SELF-CENSORSHIP AND STRATEGIC OMISSION IN MORAL COMMUNICATION

Ike Silver

A DISSERTATION

in

Marketing

For the Graduate Group in Managerial Science and Applied Economics

and

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2022

Supervisor of Dissertation:

Deborah A. Small, Laura and John J. Pomerantz Professor of Marketing, Professor of Psychology

Graduate Group Chairperson:

Nancy Zhang, Ge Li and Ning Zhao Professor, Professor of Statistics

Russell Epstein, Professor of Psychology

Dissertation Committee:

Barbara A Mellers, I. George Heyman Professor of Marketing, Professor of Psychology
Geoffrey P. Goodwin, Associate Professor of Psychology
Jonah Berger, Associate Professor of Marketing
Alex Shaw, Associate Professor of Psychology, University of Chicago

DEDICATION

This dissertation is dedicated to Nate Barnett,

who, to my astonishment, accepts me with no revisions

ACKNOWLEDGMENTS

Nothing worth doing can be accomplished alone. In writing this dissertation, and in surviving the tumultuous years during which it was conceived, I was fortunate to receive help and support from many people. I am indebted especially to the following:

To my advisor Deb Small, whose commitment to producing thoughtful and engaged doctoral students is matched only by her talent for producing thoughtful and engaging research. Passing through your orbit has changed the trajectory of my life.

To my dissertation committee and to my collaborators and mentors – Barb Mellers, Geoff Goodwin, Alex Shaw, Jonah Berger, Brent Strickland, Joshua Knobe, Phil Tetlock, Jonathan Berman, David Reibstein, Sage Baron, Jordie Gerson, and Jackie Silverman. Your perspectives and wisdom are marked indelibly in this work, and your many kindnesses were indispensable to its completion.

To my peers and friends in the academy – Erika Kirgios, Corey Cusimano, Jessie Sun, Sam Skrowonek, Joowon Klusowski, Yi Liu, Mingyung Kim, Michael Kurish, Nofar Duani, and Suneal Bedi. It is sometimes said that compassion means "to suffer together." You can count on my compassion and friendship now and always.

To my friends outside the academy – Matthew Ginsberg, Jess Wirtshafter, Mike Pontera, Brittany Hundzynski, Eden Axelrad, Jovanna Bubar, Itai Axelrad, Jamie Levinson, Nick Schwasman, Jordan Schellinkhout, Chuck Mueller, Chelsea Thomas, Riley Fressie, Christine Shaw, Teddy Terezis, Caitlin Dennis, Ahon Sarkar, Kari Wei, Nathaniel Dolquist, Mchaney Carter, Eliza Dryer, Becca Edelman, Rahim Hashim,

Calista Small, Marina Campana, Angelica Chaghouri, and many others. I am as proud of my relationships with you as I am of anything else in my life.

To my family, both biological and chosen – Kimberley Dashiell, Mark Silver, Cody Dashiell-Earp, Augy Silver, Kellen Silver, Alex Schwed, Nate Merchant, Stu Silver, Judi Sittler, Connie Silver, the Barnetts, the Ginsbergs, Lisa Axelrad, Amy and Adam Moreland, and Marla Freeman-Jackson. I am grateful for your patience, your strength, and for always offering me a soft place to land.

To my therapist, Jay Moses, the patron saint of hopeless cases.

To the Wharton Behavioral Lab, the Wharton Social Impact Initiative, the Mack Institute for Innovation Management, the Wharton Decision Processes Lab, the Penn Moral Psychology Reading Group, and the Wharton Marketing Department family for providing financial and logistical support for this work.

And, finally, to my beloved Nate and Chico. Without you, there's no joy. Without joy, there's no dissertation.

ABSTRACT

SELF-CENSORSHIP AND STRATEGIC OMISSION IN MORAL COMMUNICATION

Ike Silver

Deborah A. Small

In their words and actions, people are motivated to appear morally good, socially conscious, and politically correct. Yet while *having* a reputation for strong moral character can be extremely valuable, *cultivating* such a reputation can be a tricky business. Indeed, individuals who advertise their commitment to moral causes or who weigh in on hot-button social issues are often met with skepticism or scorn, and sometimes castigated as disingenuous braggarts or tactless virtue-signalers. To date, work in this area has focused primarily on the ways in which people try to broadcast their morality (e.g., by donating time and money, purchasing cause marketing products, voicing their political opinions online, or boycotting transgressive individuals or brands) and on how such strategies are received by others. By contrast, much less is known about the strategies people use to *avoid* moral and political communication. I present three essays which examine processes of self-censorship and strategic omission, and which highlight their unexpected costs. The first essay investigates people's reticence to talk about their charitable giving and develops an intervention to remedy it, with an eye towards boosting word-of-mouth fundraising for nonprofit organizations. The second essay explores people's discomfort sharing attitude-incongruent facts about charged social issues, even facts they simultaneously admit are true and relevant, conceptualizing

this behavior as a potential source of misinformation and distortion. The third essay investigates moral neutrality, documenting people's discomfort with sharing controversial moral opinions, but also their distrust of others who try to avoid moral issues. Drawing on experimental methods from psychology and behavioral economics, these three essays show not only that people use strategic omission strategies to protect their moral reputation and avoid morally-charged conversations, but also that such strategies frequently come with ironic costs for social discourse, for their reputations, or for the greater good.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

PREFACE

Being thought of as morally good can be a powerful asset. We preferentially associate with, like, and trust people who we judge to have strong moral character or who we believe hold the "correct" views over hot-button social issues. Similarly, we often prefer to support brands that we see as authentically committed to social responsibility or to the greater good. Indeed, in forming impressions, morality is considered a central axis along which other social judgments are organized and subsumed. In this way, and in many others, judgments of moral goodness fundamentally shape the way we engage with our social surroundings.

It is no surprise then that people feel compelled to manage their moral reputations. If being thought of as morally good confers status benefits, then advertising our moral character – moral signaling – becomes a powerful social imperative. In recent years, scholarly research on this idea has rapidly proliferated. Drawing on perspectives from psychology, economics, marketing, anthropology, and philosophy, social scientists have come to understand a wide variety of human behaviors as stemming from a desire to be seen as morally good. To name just a few examples, moral signaling has been implicated in how people choose to donate their time and money, select and consume products, behave in public forums, and treat others both within and outside their social groups. Its hallmarks can be seen in expressions of moral outrage on social media; in the posturing of politicians and public figures over issues like racial justice or public health; and in people's everyday choices about where to travel, what to eat, or what to wear.

The trouble with moral signaling is that it can go horribly wrong. Wanting to *look* good is not necessarily the same thing as wanting to *be* good, and observers are vigilant cynics about sussing out the difference. In fact, observers are prone to ascribe selfishness, inauthenticity, ulterior motives, and more generally to respond skeptically to the ostensibly moral actions of others. Moreover, taking moral stands can open one up to attributions of hypocrisy or charges that one's moral opinions are misguided, ill-informed, or wrongheaded. In this way, attempts to signal morality can end up backfiring, painting actors who try to do good in a more a negative light overall. Indeed, individuals who talk about their donations to charity or their commitment to vegetarianism, as examples, are often judged as disingenuous braggarts or holier-than-thou hypocrites. Brands that try to position themselves as socially responsible around issues like abortion or LGTBQ+ advocacy can end up looking like self-interested bandwagoners or sellouts, especially among those who disagree with their positions.

A central assumption of this dissertation is that people are aware of the fraught position they find themselves in when trying to communicate about morally tinged subjects. That is, people recognize that moral territory is reputationally risky. As a result, people often try, in various ways, to *avoid* moral communication, hoping to sidestep the potential costs associated with saying the wrong thing. The three essays in this dissertation explore manifestations of this psychology. In the first essay, I investigate people's reticence to talk about their charitable giving and develop an intervention to remedy it, with an eye towards boosting word-of-mouth fundraising for nonprofit organizations. In the second, I explore people's discomfort sharing attitude-incongruent

facts about charged social issues, even facts they simultaneously admit are true and relevant, conceptualizing this behavior as a potential source of misinformation and distortion. In the third essay, I investigate moral neutrality, documenting people's discomfort with sharing controversial moral opinions, but also their distrust of others who try to avoid moral issues.

Such strategies of calculated omission are harder to detect than more active forms of moral signaling. By definition, they involve communication that *doesn't* occur. Still, they come with their own costs and benefits, both in terms of outcomes for the actor's reputation and in terms of outcomes for the greater good. Across the three essays presented here, I suggest that despite their intuitive appeal, omissions can be costly in ways people do not readily anticipate. My hope is that the work contained in this dissertation will help to broaden our understanding of how people navigate moral communication and suggest avenues for advancing the greater good by encouraging more open and trusting moral conversations

CHAPTER 1: PUT YOUR MOUTH WHERE YOUR MONEY IS: A FIELD EXPERIMENT ENCOURAGING DONORS TO SHARE ABOUT CHARITY

Ike Silver & Deborah A. Small

Abstract.

Sharing about charity online or in personal conversations can help raise awareness and bolster fundraising for good causes. However, we show that when deciding whether to share about a charitable cause after donating to it, donors often overlook the social impact of sharing and instead focus on possible risks to their reputation (e.g., of seeming braggy, inauthentic). In a large preregistered field experiment, we tested a brief post-donation intervention designed to encourage word-of-mouth by re-orienting donors to the idea that sharing about charity means doing more good. 77,485 donors to an education non-profit received either a control or treatment message asking them to share a link to the cause via social media, text, or email. Compared to the organization's standard solicitation ('Please share your donation…'), our intervention emphasized the consequences of sharing *for the cause* ('Your donation can start a chain reaction…'). This brief message increased click-through by 5.1%, likelihood of recruiting at least one later donation by 12.4%, and funds raised via referral by 16.6% relative to control. Exploratory follow-up analyses suggest that these messaging effects are most pronounced among more generous donors: The more donors gave, the more responsive they were to

the social impact message asking them to share. While many field experiments aim to increase giving directly, we show that thoughtful marketing can also exogenously influence word-of-mouth, and we discuss approaches for encouraging sharing in the domain of charity and beyond.

Introduction.

Donations to charity could go farther if donors were more willing to talk about them with others. Indeed, posting about the charities one supports on social media or mentioning them in conversation can raise awareness about worthy causes and fundraising campaigns; it can serve as social proof that people are giving to them; and it can reinforce norms of generosity and altruism more broadly (Agerström et al. 2016; Kraft-Todd et al. 2015). Recognizing that consumer sharing drives revenue, many firms, both for-profit and non-profit, ask customers to 'refer their friends' or 'spread the word' in order to reap the benefits of word-of-mouth (WOM). In the domain of charity in particular, such benefits accrue both for the organization and for society more broadly.

Yet despite the good that can come from sharing about charity, people often treat their giving as a private matter. Many cultures prescribe modesty with respect to charity, and good deeds done anonymously are often considered especially praiseworthy (De Freitas et al. 2019). By contrast, when donors publicize their commitment to moral causes or talk openly about donations of time and money (e.g., on social media), others sometimes see them as doing good for the wrong reasons – to *look* good rather than to *be*

good – and judge them negatively as a result (Berman et al. 2015). Thus, although sharing about charity can help to advance the cause, it also entails reputational risk. The present paper explores donors' hesitancy to share about charity and tests a simple messaging intervention designed to combat it in the field.

Our investigation focuses on the choice of whether to pass along information about a charitable organization or cause after donating to it – whether to generate word-of-mouth. In line with past WOM literature (e.g., Berger, 2015), we define what constitutes sharing broadly. When sharing about charity, donors may post a link on social media, highlight their personal experiences or feelings about the organization, pass along an advertisement or news story related to the cause, or even ask others to give directly. They may talk about how much they gave or merely remark on the charity in a way that implies they donate to it without saying so explicitly[1]. Across all these cases, and inherent to sharing about charity more broadly, is a tension between reputational and altruistic goals.

We posit that when deciding whether to share about charity at all, people pay more attention to the consequences for their reputation (*What will others think of me?*) than to the consequences for the cause (*Can I influence others to donate?*). In turn, worrying about what others will think and overlooking their potential to help the cause may lead people to avoid the topic altogether. Drawing on this account, we ask whether it

[1] Note that in our field dataset, we will observe donors' click-through rates on messages asking them to share about a charitable cause after donating to it, and whether they subsequently recruit donations from others, but we will not observe what they say to others directly (e.g., in personal communications or on social media). We return to this issue in the General Discussion.

is possible to encourage WOM about charity with messaging that reorients donors' attention to the social impact case for sharing: that talking about charity means doing more good. As compared to simply asking them to share, we test a brief intervention cuing donors to consider that their generosity can 'start a chain reaction,' but only if they share about the cause they support with others.

Whereas many field experiments in marketing and economics test interventions to increase donation rates and encourage prosocial purchases (e.g., Dubé, Luo, and Feng 2017; Munz, Jung, and Alter 2020; Sudhir, Roy, and Cherian 2016; Yang and Hsee 2021), our experiments probe a less well understood route to increasing charity revenue: Encouraging donors to spread the word. More broadly, while marketing scholars tout the importance of harnessing WOM (Godes and Mayzlin 2004; Godes et al. 2005; Berger 2014), relatively few field experiments provide direct evidence for exogenous effects of marketing on consumer sharing decisions.[2] Our research unpacks the psychology of sharing about charity and investigates an intervention to encourage WOM in the field.

We report two preregistered experiments: a laboratory experiment showing that people often fail to consider social impact when deciding whether to post online about charitable causes, and a large-scale field experiment (N = 77,485) testing a proposed remedy. To preview, our treatment increases click-through rates on a solicitation to share about the cause by 5.1% relative to control, and it boosts downstream donations recruited

[2]Where such experiments exist, they typically use incentives to spur WOM – e.g., via promotions (Berger and Schwartz 2011) or referral bonuses (Wolters, Schulze, and Gedenk 2020).

via referral (the average WOM value generated per participant) by 16.6%. Before reporting our experiments, we briefly review relevant literature and derive predictions.

The Social Impact of Sharing about Charity

Decades of research in marketing have revealed that consumer decision-making is sensitive to social influence: People take cues about how to spend their money what to consume from the words and actions of others (e.g., Elster 1989; Goldstein, Cialdini, and Griskevicius 2008; Huang et al. 2020; Iyengar, Van den Bulte, and Valente 2011). Direct word-of-mouth recommendations are believed to be an especially potent mechanism (Berger 2014). When asked, people say that WOM referrals are more trustworthy than other forms of advertising (Nielsen 2012), and they spend an estimated 7-10 trillion dollars annually in response to social conversations with family and friends (Engagement Labs 2017). Recognizing that word-of-mouth drives revenue, many firms ask customers to 'refer their friends' or to 'spread the word' in order to attract more business. So important is WOM that many organizations allocate valuable marketing dollars toward incentivizing referrals (Gershon, John, and Cryder 2020) as well as identifying influencers (Nair, Manchanda, and Bhatia 2010) to broaden the reach of their products and services.

When it comes to charity in particular, WOM can have a number of desirable effects. First, sharing serves a key informational purpose: It can raise awareness about worthy causes and help potential donors learn about organizations that advance them (see also, Godes and Mayzlin 2009). People cannot contribute money, donate goods, sign

petitions, or volunteer time for charities they do not know exist. Second, telling others about charities one supports often puts them on the spot to give, and those directly solicited often feel uncomfortable saying no, contributing more as a result (DellaVigna, List, and Malmendier 2012; Flynn and Lake 2008). Third, sharing can influence others to give by reinforcing norms of generosity and providing social proof that people abide by them (Kraft-Todd et al. 2015). Outside of specific religious groups that practice tithing, the norms about giving to charity are often opaque. Indeed, the fact that people often give privately prevents such norms from developing. If, instead, donors were more open, others may feel more compelled to give, too. We refer to these benefits collectively – increased awareness and funds raised for charitable causes – as the 'social impact' of sharing about charity.

Many assert that impact should be the primary guiding force in decisions about doing good. For example, the effective altruism movement, a growing coalition of socially conscious consumers, scholars, and philanthropists, argues that donors should strive to contribute to charitable causes in whatever way *maximizes the good they can do per dollar* (MacAskill 2015). And, at least in some cases, donors are motivated by social impact information. For instance, donors select more effective charities when impact information is easily evaluable, (Berman et al. 2018) and they give more when they are told their money will be matched (Karlan and List 2007). Building on these ideas, it can be argued that if people truly care about the causes they donate to, they should be willing to talk about them with others in their social network, harnessing the marketing benefits of WOM to do more good (see, Small et al. 2018; Zaki and Cikara 2020). Thus, from the

perspective of optimizing social impact, sharing about charity seems preferable to keeping it quiet.

However, as we will find, social impact may not be the primary consequence that comes to mind when deciding whether to share. Donors also worry about how sharing about charity will make them look to others.

The Reputational Consequences of Sharing about Charity

Impression management is a central motive in consumer behavior. People deftly curate their presence on social media (Schlosser 2020), watch what they say in public (Toubia and Stephen 2013), and consume conspicuously in order to signal their status, values, and preferences to others (Bagwell and Bernheim 1996). In fact, across diverse contexts from fashion choices to athletic performance, most people *overestimate* the extent to which their peers notice and evaluative their behavior (Gilovich, Medvec, and Savitksy 2000). Such tendencies reveal a powerful drive to cultivate favorable impressions and protect one's reputation in the eyes of others. Because perceptions of *moral* character are central in social judgment (Goodwin, Piazza, and Rozin 2014), reputational concerns might be particularly important in deciding whether to share about charity.

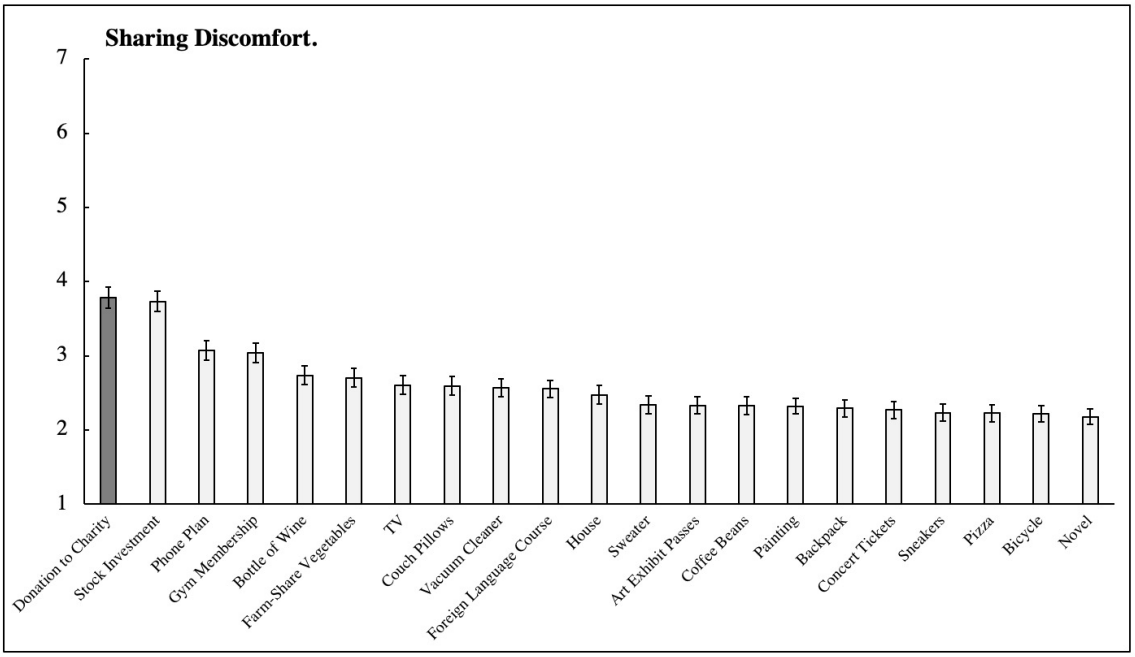
It might seem that impression management should push people to share about their giving. After all, donations are acts of selflessness and generosity, which could reflect positively on givers. However, reactions to those who advertise their goodness are often quite cynical (Critcher and Dunning 2011; Miller and Ratner 1998). This is because

the decision to share contains information not only about *what* a person does, but also about *why* she does it. For example, a Facebook user who posts about a charity she supports on her personal page might seem motivated, not by an altruistic impulse to help the cause, but by a desire to look generous. Such attributions of self-interest in the context of charity can provoke distain (Berman and Silver 2022; Silver, Newman, and Small 2021), undermining the signal of selflessness inherent to donating in the first place. In line with this idea, many cultures and religions encourage private donations, praising anonymous good deeds and treating public generosity with skepticism (De Freitas et al. 2019). Accordingly, sharing about charity can have ironic effects, sometimes painting ostensibly selfless donors as tactless braggarts or holier-than-thou hypocrites (Berman et al. 2015). Beyond appearing self-interested, donors who choose to share may also seem self-righteous, intrusive, or pushy.

With their moral reputation on the line, donors might be particularly apprehensive and uncomfortable at the prospect of sharing about charities they support relative to other sorts of purchases they make. To investigate this possibility, we ran a preregistered pilot study. We recruited 198 participants from Amazon's Mechanical Turk (M_{age} = 32, 49% female) and asked them to report how comfortable or uncomfortable they would feel talking about 21 different expenditures with their peers, friends, and family. Expenditures spanned a variety of ordinary categories from buying a dozen eggs, to signing up for a gym membership, to securing passes to an art exhibit, to purchasing a new TV. Among the expenditures we tested, a donation to charity was rated as the most uncomfortable to

talk about. People seem to see charitable giving, although intuitively praiseworthy, as a relatively unpleasant topic of conversation. See Figure 1.

Figure 1. (Chapter 1) Pilot Study: Anticipated discomfort ratings (1:"Totally comfortable" to 7:"Extremely uncomfortable") telling others about 21 ordinary expenditures



Note. Error bars represent standard errors.

That people would be especially uncomfortable sharing about charity suggests a potential hurdle for non-profits hoping to solicit WOM and access its social impact benefits. Our experiments sought to investigate this psychology and test a simple intervention to combat it in the field.

Present Research

If donors prioritize social impact, they should be willing to share about the causes they support. But if they worry about their image, they might be more hesitant to share. Which concern typically wins out? In line with evidence that people are preoccupied with social judgment (Gilovich, Medvec, and Savitksy 2000) and that worries about sending the right social signals can undermine altruism (Ariely, Bracha, and Meier 2009; Yang and Hsee, 2021), we expect that reputational consequences will loom large, often at the expense of social impact. Specifically, we predict that when deciding whether to share in this context, donors typically pay more attention to possible consequences for their reputation than to possible consequences for the cause.

If donors fail to conceptualize sharing as an opportunity to do more good, then a timely message to consider social impact may help. Along similar lines, prior research has found that well-timed reminders can prompt desirable behaviors from saving for retirement (Karlan et al. 2016) to getting immunized (Szilagyi et al. 2000). But whereas such studies typically remind consumers to follow through with tasks they have already committed to, our intervention seeks to reframe the decision at hand by reminding donors to take into account something they care about in principle but may forget to consider in practice. By re-orienting donors to the idea that sharing is a way to do more good, we sought to boost their willingness to share and, ultimately, to recruit others to donate to the cause.

We tested these predictions across two experiments. Experiment 1 was a laboratory study which shed light on what donors ordinarily think about when deciding whether to share about charity. Experiment 2 was a large-scale experiment with 77,485 donors testing our proposed intervention in the field. In short, we demonstrate that a brief and free reminder to consider social impact promotes sharing and in turn boosts donation revenue. Preregistrations, data, materials, code, and appendix materials are available at: https://researchbox.org/105&PEER_REVIEW_passcode=MKFQMJ.

Experiment 1. Overlooking social impact in deciding whether to share

Experiment 1 tested the hypothesis that, when deciding whether to share about charity, people attend more closely to possible consequences for their reputation than to possible consequences for the cause. Participants reported how likely they would be to post about a charity they care about after being prompted to consider different potential consequences of doing so.

Method.

377 participants (M_{age} = 21.7, SD = 6.5, 68% female) were recruited from a business school's behavioral lab and paid $10 for a one-hour session, of which our experiment took ten minutes. Our target sample was 400 participants, but recruitment was conducted by the lab and constrained by participant sign-ups. No participants were excluded from analysis.

Participants first wrote down the name of a "charity or charitable cause" they felt was personally important. They were then asked to imagine making a donation to the cause and receiving a follow-up email from the charity with a request to share about the cause on social media (e.g., on Facebook, Instagram). The key dependent variable was a willingness-to-share measure on a 7-point scale from "Not at all willing" to "Completely willing." But before deciding whether to share, participants completed a short writing task.

Participants were randomly assigned to one of three writing prompt conditions. In the first, participants wrote a short entry about the consequences of sharing *for their reputation* (i.e., how others would view them). In the second, participants wrote a short entry about the consequences of sharing *for the charity* (i.e., how sharing would impact the cause). In a third baseline condition, participants were simply asked to record "whatever comes to mind" when thinking about whether to share. After the writing task, participants indicated willingness-to-share. They then completed age and gender demographics as well an additional exploratory question about frequency of social media use.

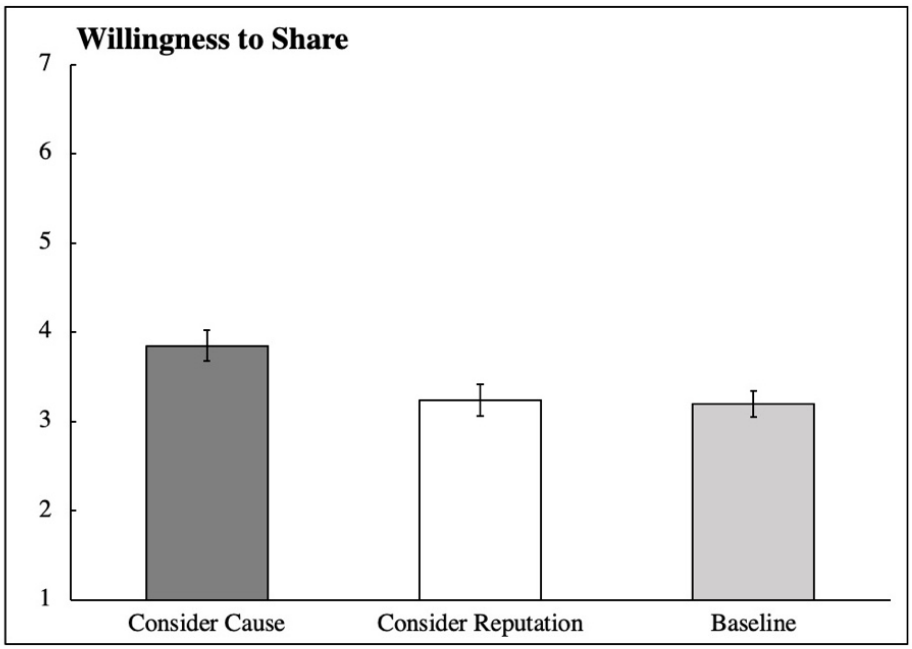
We had two predictions. First, consistent with the notion that people spontaneously consider their reputation in deciding whether to share, we predicted that sharing would not differ between those prompted to consider their reputation and those in the baseline condition. Second, we predicted that asking donors to consider consequences for the cause would increase willingness to share, relative to the other two conditions. We

also coded participants' written responses and conducted exploratory analyses, described below.

Results.

We first subjected the sharing measure to a one-way ANOVA with condition as a between-subjects factor. This procedure revealed a significant omnibus effect of condition ($F(2, 374) = 4.97, p = .007, \eta_G^2 = .026$). Planned comparison t-tests revealed results in line with our predictions. Participant indicated a higher willingness to share when prompted to consider how sharing would impact the cause ($M = 3.85$, SD = 1.92) than in either the condition that prompted reputational considerations ($M = 3.24$, SD = 1.97; $t(249) = 2.52, p = .012, d = .32$) or the baseline condition ($M = 3.20$, SD = 1.63; $t(248) = 2.91, p = .004, d = .37$). There was no difference between those prompted to consider reputation and those in the baseline condition ($t(251) = .17, p = .87, d = .02$). These results are consistent with the idea that, when considering whether to share about charity, donors ordinarily consider consequences for their reputation more readily than consequences for the cause. See Figure 2.

Figure 2. (Chapter 1) Experiment 1: Willingness to share by condition (1: "Not at all willing" to 7: "Totally willing")



Note. Error bars represent standard errors.

To investigate further, we asked a hypothesis-blind RA to code participants' responses to the writing prompts. Specifically, entries were coded according to whether they mentioned "reputation-based reasons" (how sharing would influence judgments about the sharer) or "cause-based reasons" (how sharing would influence outcomes for the charity) as two separate dummy variables. Moreover, whenever a participant mentioned a reputation- or cause-based reason, it was further coded according to whether it highlighted a positive impact or a negative impact of sharing. Each was coded as a separate dummy variable. In sum, this coding allowed us to ascertain, for each written response, whether it mentioned possible consequences of sharing about charity for the

donor's reputation and/or for the cause, and whether the consequences participants brought up were positive, negative, both, or neither.

As would be expected with the manipulation, participants in the condition which explicitly prompted them to think about reputation were quite likely to mention in their written responses how sharing about charity would impact their reputation. 91.3% of participants mentioned reputation, while only 14.2% mentioned social impact. In the condition which explicitly prompted social impact, we saw the opposite pattern. Only 13.7% mentioned reputation, while 92.7 % mentioned impact. These results provide confirmation that we successfully manipulated what participants were thinking about when reporting willingness to share. Of particular interest was whether participants in the baseline condition, where there was no prompt one way or the other, were more likely to bring up their reputation or social impact. In line with our theorizing, participants asked to write "whatever comes to mind" were more likely to mention the impact of sharing on their reputation than for the cause. 62.7% of participants in the baseline condition mentioned reputation, while only 39.7% mentioned the cause (McNemar's Chi-Square (1 df) = 12.37, $p < .001$).

Across conditions, mentions of possible consequences for the cause nearly always highlighted positive outcomes (e.g., "I want to showcase the charity so it receive[s] more money"). Among all written responses which mentioned possible outcomes for the cause, 92.9% mentioned only positive outcomes for the cause that might result from sharing, 0.5% mentioned only negative outcomes, 1.1% mentioned both positive and negative outcomes, and 5.5% were coded as neither positive nor negative. By contrast, mentions

of possible consequences of sharing for the sharer's reputation were substantially more negative (e.g., "I don't want to make it seem like I donated just to get social credit"). Among written responses which mentioned reputation, 21.6% mentioned positive outcomes only, 50.2% mentioned negative outcomes only, 19.2% mentioned both positive and negative outcomes, and 8.9% were coded as neither positive nor negative.

Finally, we estimated an OLS regression predicting willingness-to-share across conditions (N = 377) from dummy variables capturing whether participants mentioned consequences for the cause and/or for their reputation in their written responses. Results are reported in Table 1 below. In line with our theorizing, bringing up the cause was associated with greater willingness to share, while bringing up reputation was not. If anything, bringing up reputation was negatively associated with willingness to share. The fact that most people fail to mention the cause at all when asked to share at a baseline suggests a potential avenue for intervention, one we explored directly in our next experiment.

Table 1. (Chapter 1) Experiment 1: OLS regression predicting willingness-to-share from whether written responses mention consequences for reputation or for the cause

	Willingness to Share (1-7)
	N = 377
Written Response Mentions Reputation	-.26 (.22)
Written Response Mentions the Cause	.74*** (.21)
R^2	0.058

Note. Unstandardized betas and standard errors; p-values: ^ < .10, * < .05, ** < .01, *** < .001

Discussion.

The results of Experiment 1 supported our predictions and theorizing. Participants were more willing to post about a charity online when instructed to consider the consequences of doing so for the cause vs. for their reputation. Meanwhile, in the baseline condition, which simply asked them to report whatever came to mind when thinking about whether to share, participants seem more focused on their reputations. Indeed, baseline condition participants reported willingness-to-share in line with the reputation prompt condition, and their written responses were much more likely to mention reputational consequences than consequences for the cause. Moreover, across conditions, participants who highlighted consequences for the cause focused on positive outcomes of sharing for the cause, whereas those who highlighted consequences for their reputations predominantly focused on negative outcomes.

Taken together, this pattern of results suggests that donors asked to share about charity are more likely to think about their reputation than about the cause, and that doing so brings to mind reputational risks that might hinder sharing more so than reputational benefits which might encourage it. It also provides reason to believe that a message reorienting donors to the social impact case for sharing – that sharing means doing more good – might increase WOM.

Experiment 2. Encouraging donors to consider social impact in the field

Experiment 2 was a field experiment that randomly assigned donors to one of two sharing solicitation messages at check-out after completing an online donation: The firm's standard message that simply asked donors to share or a treatment message emphasizing that sharing can help the cause. We sought to test whether this brief message could re-orient donors' attention to the social impact consequences of sharing and thus encourage word-of-mouth.

Experimental Setting and Method.

Experiment 2 was conducted in partnership with the education non-profit DonorsChoose.Org. DonorsChoose.org ("DonorsChoose") is an online platform where users can learn about and donate to classroom-based fundraisers in underfunded schools across the United States (e.g., raising money for new desks, books, science equipment). After giving through DonorsChoose's online platform, donors encounter a brief pop-up message which thanks them for their donation and asks them to tell friends and family

about the cause. To better capitalize on word-of-mouth effects, a major objective for DonorsChoose's marketing team has been to increase donors' propensity, after giving, to click through and post about classroom fundraisers on social media or share them via email or text message.

During a ~5-week period from 8/13/2020 to 9/16/2020, donors who gave money via the organization's online portal served as participants in a field experiment. Every time a donor completed an online donation, they saw a pop-up message asking them to tell others about the cause. Donors were randomly-assigned to one of two versions of this pop-up solicitation to share, and donors who gave more than once during the test period saw the same version at each donation occasion. The control condition employed DonorsChoose's standard language: "Share this classroom with family and friends." The treatment condition tested an alternative version which emphasized social impact: "Your donation can start a chain reaction, but only if you tell others about the cause. Share this classroom with family and friends." We designed this treatment to draw attention to donors' capacity to influence others to get involved ("Your donation can start a chain reaction…") and to communicate that such impact was contingent on sharing ("but only if you tell others…"). In addition to the sharing solicitation, the pop-up window in both conditions displayed clickable icons (i.e., to Twitter, Facebook, and Email for desktop users, or to Twitter, Facebook Messenger, or SMS Messages for mobile users), which allowed us to observe click-through.

Data Overview and Observed Variables.

Our data include observations from 77,485 donors, who made a total of 117,090 donations during the test period[3]. 83.5% of donors gave only once and so saw the focal pop-up only once. The remaining 16.5% gave more than once during the period and so saw the pop-up at each donation occasion, always in the same condition (number of donation occasions: mean = 1.51, median = 1). Our primary outcome variables were **click-through** (whether a donor clicked on a pop-up soliciting them to share during the test period) and **recruitment** (whether a donor subsequently recruited at least one downstream donation via a unique referral link).

The click-through measure captures whether a given donor clicked at least once on any of the sharing icons on a pop-up message displayed to them during the test period. Note that for the 16.5% of donors who gave more than once, and saw a pop-up each time, we can observe whether they clicked-through at least once during the test period but cannot identify at which donation occasion they clicked or whether they clicked more than once. The remaining 83.5% donated only once during the test period, saw the pop-up one time, and thus had only one opportunity to click. Click-through was recorded during the first 30 minutes after a donor saw a given pop-up.

When a donor clicked on a sharing pop-up, they were redirected to their platform of choice and supplied with a unique referral link, allowing us to record downstream

[3] **Note on exclusions**: We removed an additional 1468 donors who evaded assignment to condition in DonorsChoose's A/B software and so saw different messages at different donation occasions. Note that because our measure of click-through is at the donor-level, and cannot be tracked to a particular donation occasion for donors who give more than once, these 1468 participants cannot be included in our analysis (or else they would be in both conditions simultaneously). We also excluded 258 donors whose individual payments were not tracked. These participants cannot be included in any analyses that consider number of donations or amount donated; however, including these 258 donors in models that predict click-through and recruitment from condition alone does not impact the results.

referrals from each participant. Because donors shared to their private social and personal media, our field partner cannot observe sharing directly (i.e., what donors' send in personal emails or post to Facebook). This is a limitation of our data. However, if a subsequent donation came in through a given donor's referral link, that participant would have necessarily clicked-through, accessed the link, shared it, and successfully influenced someone to give. In our preregistration, we planned to treat recruitment as a binary variable (assigning 1 if a participant recruited any donations, 0 otherwise) as we had no predictions about the impact of condition on donors' persuasiveness in recruiting many vs. few donations after deciding to share. Nevertheless, we also observe the number of downstream donations recruited and total amount raised by each donor via sharing.

To summarize, the dataset we obtained from our field partner included the following information for each participant: the primary independent variable (message condition: standard language control or treatment emphasizing social impact) and two key outcome variables: **click-through** (never = 0, at least once = 1) and **binary recruitment** (no donations recruited = 0, at least one donation recruited = 1). In addition, for each participant, we observe the number of donations they made (with corresponding dates and times), the total donation amount they gave during the period, the number of donations they recruited via referral, and the total donation amount they recruited during the period. We also observe the date(s) and time(s) each participant donated and whether each donation came in through a mobile, desktop, or tablet device (and present exploratory analyses of these data in our online materials). We cannot observe demographic information for individual participants, as our field partner has a policy of

not disclosing it to third parties. However, among the organization's donors at large, 78% identify as female, with age distributed as follows: 40-and-under = 27%, 40-59 =37%, 60-and-up = 36%.

To examine any longer-term effects of treatment on the donation behavior of our participants after the test period, we later obtained a binary measure of whether each donor in our field test donated again in the three-month period after our field test (between 9/16/2020 when the experiment ended and 12/15/2020).[4]

Note on Deviations from Preregistration.

We report the following necessary deviations from our preregistration. First, although we anticipated recruiting ~30,000 participants per condition, our field partner received more donations than expected during the test period, and so our data contain closer to 40,000 per condition. Second, although we planned to use mixed-effect logistic regressions, adding random intercepts by participant to the model proved unnecessary, as click-through was ultimately collected at the participant-, rather than the donation- level. Third, we removed 1726 participants prior to the reported analyses (1468 were not properly randomized and 258 were missing payment data; see footnote 3 for more detail).

[4] One possible concern with social influence experiments is network interference (Rosenbaum 2012) –that the treatment of one participant might impact outcomes for other participants. Our primary DV – click-through – should not be susceptible to interference. However, it is theoretically possible that later donations could be. Such an explanation would require (a) that participants in treatment and control clicked/shared to the same potential donors, (b) that these donors did not simply respond to the first post they encountered, and (c) that the effect of seeing later posts differs from the effect of seeing the first post. These conditions seem unlikely for more than a tiny fraction of participants, but we cannot rule them out entirely.

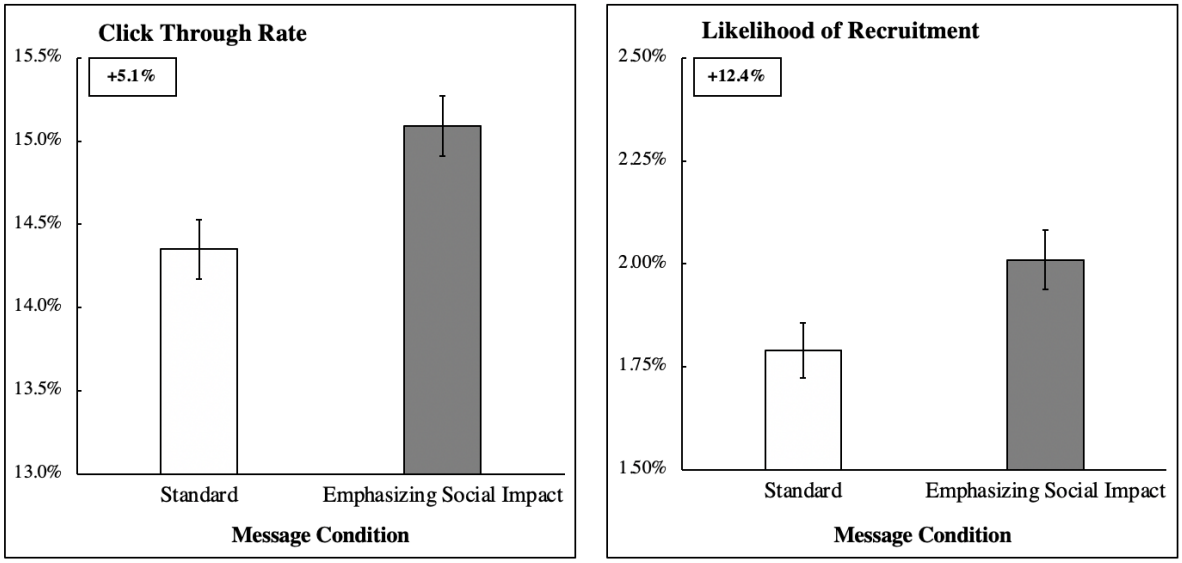
Results.

Testing Central Predictions.

We predicted that donors in the treatment condition would be more likely to click-through on the sharing solicitation message and subsequently recruit at least one downstream donation. To investigate, we began by estimating logistic regression models predicting our key DVs – click-through (*did the donor click through on any of the sharing links they saw?*; 1,0) and recruitment (*did the donor recruit at least one downstream donation*?; 1,0)? – from condition only.

The results supported our predictions. Participants were more likely to click on the sharing-solicitation pop-up in the treatment condition (15.09% click-through) vs. the control condition (14.35% click-through; $B = .059$, SE = .020, Wald $Z = 2.89$, $p = .004$). This difference corresponds to a 5.1% relative increase in participants' likelihood of clicking on the pop-up message soliciting them to tell others about their giving. Participants were also more likely to recruit at least one downstream donation in the treatment condition (2.01%) vs. control (1.79%; $B = .12$, SE = .053, Wald $Z = 2.27$, $p = .023$). This difference corresponds to a 12.4% relative increase in participants' likelihood of recruiting at least one downstream donation via their unique referral link. Note that the larger relative effect of treatment on recruitment reflects a lower baseline. See Figure 3.

Figure 3. (Chapter 1) Experiment 2: Click-through rate and likelihood of recruiting at least one donation by condition



Note. Error bars represent standard errors.

As noted above, 16.5% of participants in our field experiment gave more than once during the treatment period and so saw the focal pop-up message asking them to share on multiple occasions. The number of exposures to the pop-up, and thus the number of opportunities to click, was equal to the number of donations made during the test period. Importantly, it does not appear that our central effects on click-through or recruitment can be explained by participants seeing the pop-up more often in treatment vs. control (i.e., as a function of whether they return to donate more often after seeing the impact-focused treatment message, which they do not; see follow-up analyses below). When we control for the number of donations made during the period in our primary

models, we continue to find effects of treatment on both key outcome variables, at nearly identical effect sizes (click-through: $B = .058$, SE = .020, Wald $Z = 2.88$, $p = .004$, binary recruitment: $B = .12$, SE = .052, Wald $Z = 2.28$, $p = .023$). We also find effects of condition on click-through among the subset donors who gave only once ($B = .045$, SE = .022, Wald $Z = 2.04$, $p = .041$). Note that in all analyses that include how much or how many times a donor in the experiment gave, we winsorized these variables at their respective 95th percentiles (Number of donations 95th percentile = 3, Dollars donated 95th percentile = $275.42; Blaine 2018) to account for a small number of extreme outliers, but the results are the same regardless of whether we do so. Regression tables for key hypothesis tests are reported in Table 2 below.

Table 2. (Chapter 1) Experiment 2: Logistic regressions predicting click-through and recruitment

N = 77,445	**Click-through (1/0)**		**Recruitment (1/0)**	
Message Condition (Standard or Emphasizing Social Impact)	**.059**** **(.020)**	**.059**** **(.020)**	**.12*** **(.052)**	**.12*** **(.053)**
Donations Made - Number (# of Donation Visits During Test Period)		-.16*** (.019)		.18*** (.041)
Cox & Snell R^2	0.0001	0.0010	0.00006	0.00029

Note. Unstandardized betas and standard errors. Donation number is winsorized for this analysis (see above); p-values: ^ < .10, * < .05, ** <. 01, *** < .001.

Interestingly, these regressions also suggest that donors who give more than once during the period were *less* likely to click on average, despite seeing the pop-up message, and having the opportunity to click, every time they donated. Logically, it can't be the case that having more than one opportunity to click on a message makes it less likely that one will click at all. Rather, a more likely possibility is that donors who are more generous (i.e., give more and more often, $r = .55$) might also be more modest (i.e., less likely to tell others about their giving). In the next section, we explore this possibility, analyzing the relationship between how much a donor gave during the period, their sharing behavior, and their susceptibility to treatment.

Exploratory Analyses.

Sharing and susceptibility to treatment by donor generosity. Donors in our experiment gave a variety of amounts: Some donated only a few dollars during the test period, while others gave thousands. Importantly, how much donors gave was not impacted by condition (OLS regression: $B = -.002$, SE = .004, $t(77{,}483) = -.43$, $p > .5$), which stands to reason given that the majority of our participants (83.5%) gave only once, and so decided how much to donate before exposure to the post-donation message.[5] Nevertheless, it could be the case that donors who happened to give more might be differentially willing to *share* or differentially impacted by our treatment messaging

[5] Participants decided whether to make a second donation after exposure to the sharing pop-up (having seen it after making their first donation). Thus, in principle, how much donors gave across conditions *could* be impacted by treatment, but only if participants were more likely to return to donate again in one condition vs. the other. This is not the case (see section on likelihood of returning below). For donors who gave only once, how much to give was decided before random-assignment, and, as would be expected with balanced randomization, we see no differences in generosity among this subset ($B = -.15$, SE = .38, $t(64{,}735) = -.41$, $p > .5$).

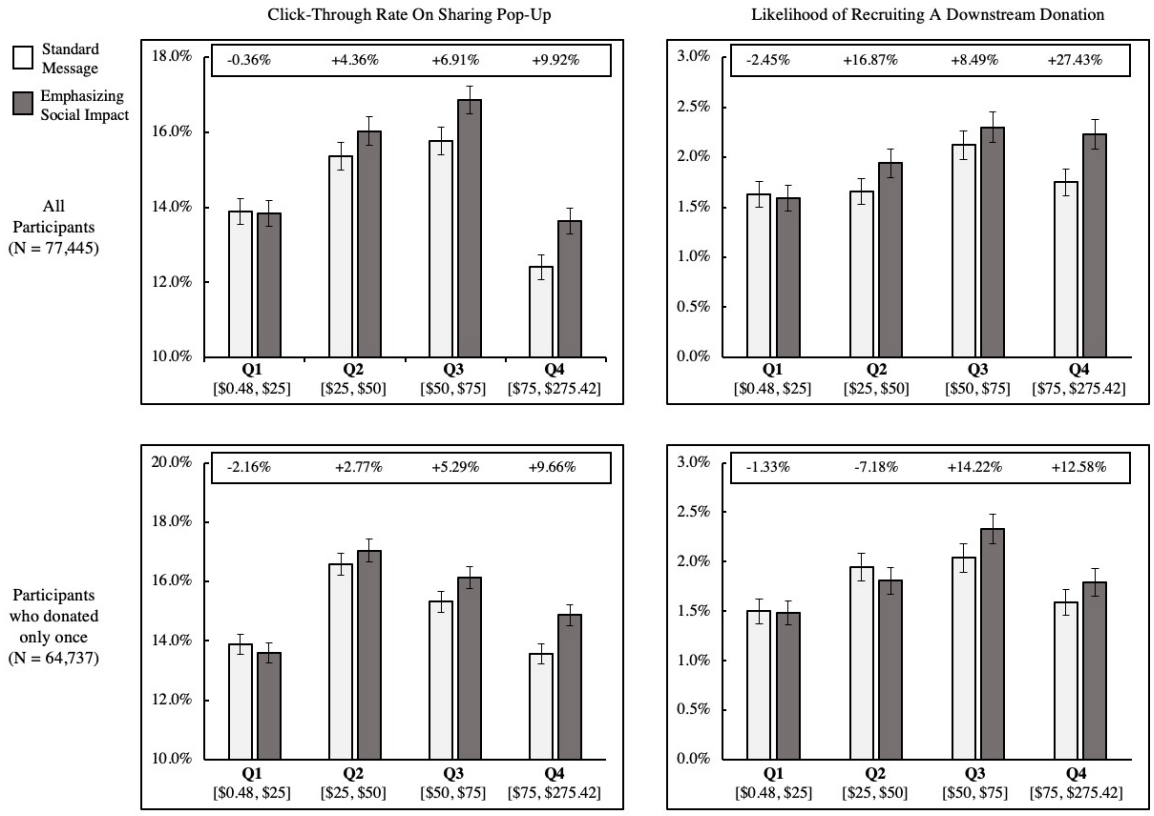
asking them to do so. The following analyses explore the relationship between generosity and sharing.

First, we analyzed the relationship between how much a donor gave during the test period and their likelihood of click-through and recruitment, collapsing across conditions. Interestingly, we found that the more a donor gave, the *less* likely they were to click-through (logistic regression: $B = .0018$, SE = .00016, Wald $Z = -10.95$, $p < .001$). In other words, donors who gave more were *less* responsive to DonorsChoose's pop-up ask to share. However, controlling for their lower rates of click-through, the more a donor gave, the *more* likely they were to recruit at least one downstream donation (logistic regression: $B = .0022$, SE = .00043, Wald $Z = 5.00$, $p < .001$). Moreover, controlling for click-through, donors who gave more during the test period recruited more donations and a greater total amount (OLS regression predicting number of donations recruited per dollar donated: $B = .000064$, SE = .000013, $t(77{,}442) = 6.42$, $p < .001$; OLS regression predicting amount recruited via referral per dollar donated: $B = .005$, SE = .00089, $t(77{,}442) = 5.91$, $p < .001$). This pattern of results suggests that more generous donors are generally more modest (i.e., less willing to tell others about their giving), but also, conditional on their willingness-to-share, more influential (i.e., more likely to bring in donations from their social networks). As more influential recruiters, these more generous donors – who may have greater means and/or a stronger commitment to the cause –represent an important target group for interventions designed to increase click-through and subsequent word-of-mouth.

Next, we examine whether our treatment effects might vary across those who gave more vs. less during the test period. To investigate, we pursued a model-free analysis, bucketing donors in our experiment into quartiles according to the amount they gave during the test period and then quantifying treatment effects on our two dependent variables: click-through and binary recruitment. This approach reveals that our treatment effects were generally stronger for those who gave more during the test period. For both dependent variables, treatments effects were null among the quartile who donated the least during the test period, but much stronger for those who donated more.

One might reasonably wonder whether these patterns can be explained by the number of donations made during the period (and thus exposures to the treated pop-up message). In other words, perhaps our effects only *appear* stronger for generous donors because those who gave more money were more likely to have donated more than once, and thus to have seen the focal pop-up more than once during the period. However, a similar approach stratifying treatment effects among those 64,737 donors who gave only once during the test period suggests this explanation is unlikely to account for the boost to our effects among more generous donors. If anything, among those who gave only once, and so saw the pop-up only once, the pattern is even more pronounced. See Figure 4 below.

Figure 4. (Chapter 1) Experiment 2: Treatment effects on click-through and recruitment by donor generosity



Note. Donor generosity is divided into quartiles according to how much donors gave during the test period. The upper-bound for the fourth quartile is winsorized at the 95th percentile of generosity. Error bars represent standard errors.

In the online materials, we report regression results which interact treatment with donation amounts and with number of donations, to provide a more formal test of these patterns of interaction observed in our by-quartile analysis. Results suggest the same positive pattern of interaction between treatment and donor generosity at varying levels of statistical significance, though these analyses impose a linear form on the interaction,

and so offer a lower resolution picture of how our effects may vary across donation amounts.

Number of donations and dollar amounts recruited. Our primary analyses find that a message focusing on social impact increases donors' likelihood of clicking on a solicitation to share about the charity and of subsequently recruiting at least one downstream donation. We can also explore whether donors who see our social impact message recruit *more* or *larger* donations, perhaps by making them more persuasive or committed recruiters. Although we cannot observe what people post on their social media accounts or say in private communications as a measure of their persuasiveness, we can test whether the number of downstream donations or the dollar amount recruited per participant varies across conditions.

Note that these variables (number and amount of donations recruited via referral) are each the product of two processes: (1) a donor's initial decision to share and (2) their effectiveness at recruiting more or larger donations if they share. Each might be separately impacted by condition, and donors who do not click through cannot recruit any donation and so appear as excess zeros in the data. To account for this data structure, we estimated zero-inflated poisson (ZIP) regression models predicting the number of donations recruited and number of dollars recruited by each participant, with condition entered as a predictor in both the logistic (predicting excess zeros) and count (predicting recruitment numbers) portions of each model. For both outcome variables, condition was a negative predictor in the zero portion of the model. Said differently, participants in treatment were *less* likely to appear as zeros in the recruitment data – likely because they

are *more* willing to click-through and share at all (Number of donations recruited model: B = -.20, SE = .070, Wald Z = -2.83, p = .005; Dollars recruited model: B = -.12, SE = .053, Wald Z = -2.27, p = .023). But the evidence is mixed as to whether donors were more effective recruiters if they did share. In the count portions of these models, treatment had a negative non-significant impact on the number of donations recruited (B = -.12, SE = .069, Wald Z = -1.69, p = .091), but a positive and significant impact on the number of dollars recruited (B = .036, SE = .0060, Wald Z = 6.09, p < .001).

A conceptually similar approach tests the average number of donations and dollars recruited across conditions among only those participants who click on the sharing pop-up to see if they subsequently become more or less effective recruiters in treatment vs. control. Conditional on click-through, participants in treatment did not bring in more or larger donations (ps > .15). Taken together, these results support the idea that our social impact message made donors more valuable recruiters by increasing their likelihood of clicking-through to share in the first place, not by making them better recruiters if they do share. Still, because they share more often, donors in treatment recruit $0.22 more on average compared to control, a relative increase of 16.6% in recruited dollars per participant.

Testing effects of treatment on donors' likelihood of returning to give again. Although we designed our treatment message to increase participants' willingness to tell others about the cause after donating, it is also instructive to explore whether our social impact message had any impact on participants' likelihood of returning to donate again themselves.

We did not have any specific predictions about such effects, and we did not find any. Looking first at participants' donation behavior during the test period, we found no differences in their likelihood of returning for a second donation in treatment vs. control (Logistic regression: $B = -.0026$, SE = .019, Wald $Z = -.14$, $p > .5$) or in the average number of donations during the test period (OLS regression: $B = -.0018$, SE = .004, $t(77{,}483) = -.43$, $p > .5$). We next examined whether participants in our experiment were any more or less likely to have given again in the four months after our test period, and we found no differences across conditions here either (Logistic Regression: $B = -.026$, SE = .018, Wald $Z = -1.44$, $p > .15$).

These results suggest that although our treatment increased donors' willingness to talk about charity, and subsequently recruit others to give to the organization, it did not have a detectable impact on their own subsequent donation behavior. They thus offer reassurance that encouraging donors to talk about charity - something they may be somewhat uncomfortable doing - does not diminish their likelihood of supporting the cause again.

All variables analyzed in the above results sections are summarized below in Table 3.

Table 3. (Chapter 1) Experiment 2: Descriptive statistics by condition

	Overall	Message Condition		Δ
		Standard	*Emphasizing Social Impact*	
N	77,485	38,621	38,864	
Key Outcome Variables				
Click-Through Rate *% clicking on sharing pop-up at least once during test period*	**14.72%**	**14.35%**	**15.09%**	**+5.1%****
Recruitment Rate *% recruiting at least one downstream donation*	**1.90%**	**1.79%**	**2.01%**	**+12.4% ***
Additional Variables - Participants' Donations				
Donations Made - Number *Mean # of donation occasions during test period*	1.24 (.58)	1.24 (.58)	1.24 (.57)	NS
Donations Made - Amount *Mean $ donated during test period*	$64.82 (67.36)	$65.16 (67.74)	$64.47 (66.98)	NS
Post-Experiment Return Rate *% of participants returning to donate again in four months after field test*	19.47%	19.67%	19.26%	NS
Additional Variables - Participants' Recruiting				
Donations Recruited - Number *Mean # of downstream donations recruited (zeros included)*	0.029 (.25)	0.028 (.25)	0.030 (.25)	NS
Donations Recruited - Amount *Mean $ in downstream donations recruited (zeros included)*	$1.45 (16.94)	$1.34 (15.96)	$1.56 (17.86)	+16.6%^

Note. Means and SDs for continuous variables; proportions for binary variables. Statistical tests are from OLS regressions for continuous variables and logistic regressions for binary variables; *p*-values: ^ < .10, * < .05, ** <. 01, *** < .001. Key

outcome variables appear in bold. Mean donations made (number and amount) are calculated after winsorization.

General Discussion.

Worthy causes lose out when donors hesitate to talk about the charities they support. Yet, our results suggest that when thinking about whether to share about charity, people lose sight of their ability to inspire others to get involved, worrying instead about what others will think of them. In a large field experiment, we find that we can encourage more donors to share about charity with messaging that makes salient the social impact case for doing so: that sharing means doing more good. Specifically, we contrasted a basic request to share with a brief message designed to reorient donors' attention to the 'chain reaction' of social impact possible if they choose to tell others about the cause. The treatment message in our field test increased click-through by 5.1% and boosted donors' likelihood of recruiting others to give by 12.4%, compared to a control condition simply asking donors to share. As a result, the average participant who saw our impact-focused solicitation to share brought in 16.6% more in recruited donations from others, relative to control.

Interestingly, exploratory analyses suggest that not all donors are equally willing to share, and that our treatment effects on WOM were more pronounced among more generous donors. Across conditions, donors who gave more during the test period were *less* likely to click on our ask-to-share pop-up, but if they did, *more* likely to recruit

others to give through their referral link. What might account for these divergent effects? One possible explanation is that generous donors are wealthier/higher-status individuals who worry more about their reputations, but who are also visible and influential within their social networks. Another possibility is that more generous donors care more authentically about the causes they support, making them more hesitant about bragging, but more impassioned and persuasive recruiters if they do decide to share (see Barasch, Berman, & Small, 2016). There may be other explanations.

No matter what drives their modesty, these reluctant influencers were more responsive to our social impact message. Comparing the highest and lowest quartiles of givers (in terms of dollars donated during the experiment) reveals heterogeneity in our treatment effects. Donors who gave the most were 9.9% more likely to click and 27.4% more likely to recruit a downstream donation in the treatment condition relative to control. By contrast, donors who gave the least were 0.4% *less* likely to click and 2.5% *less* likely to recruit in treatment relative to control. In sum, our impact-focused pop-up message boosted WOM more among more generous donors, who appear more modest about sharing but who also have greater potential for social influence if they do.

Finally, because our treatment message encouraged donors to do something that might make them feel uncomfortable, sharing about charity, we also tested whether it might have any negative effects on their likelihood of donating again. Encouragingly, our results suggest that our treatment increased WOM without reducing donors' likelihood of supporting the organization again in the future.

Implications and Future Directions

For marketers, the effects of our treatment are modest in absolute terms, but they have important economic consequences. Indeed, our treatment message yielded an increase of 16.6% in donations raised via WOM per donor treated. As a first approximation of economic significance, we can multiply the increase in WOM revenue per donor treated in our experiment ($0.22) by the number of annual donors to DonorsChoose.org (~600,000) to predict an annual revenue boost of roughly $132,000 for the organization, all from a wording change in how they ask donors to share about the cause. Note that this back-of-the-envelope calculation does not account for further network benefits of treatment – donors recruited via WOM may themselves later become recruiters – suggesting that it may be a conservative estimate.

We can also benchmark the effects we observe by comparing them to those found in previous online advertising experiments. Estimates of 'lift' (i.e., relative increase in click-through rates) from ad experiments vary (Bakshy, Eckles, Yan, and Rosenn 2012; Ghosh, Thomke, and Pourkhalkhali 2020; Lewis and Rao 2015), though recent meta-analyses suggest that the average impact of treatments in online A/B tests is around 2.3%, with modal effects typically closer to 0 (Berman et al. 2021). Comparing our effect of treatment on click-through to such benchmarks suggests an above-average effect at 5.1%. Still, it should be noted that most prior experiments measure clicks on advertisements for products and services, while we observed clicks on a solicitation to share. We know of no field experiments testing click-through rates on solicitations to share per se, either in the domain of charity or elsewhere, making comparisons to past literature difficult. We

therefore hope the effects we obtained here can serve a benchmark for related future work.

In the meantime, we note a number of potential ways that our effects might be strengthened and complemented. First, we treated an online pop-up message, which many donors might close reflexively, block automatically, or simply fail to notice. Thus, one simple way to increase its efficacy would be to embed social impact messaging into a wider set of marketing communications aimed at increasing WOM. For example, many non-profits send emails urging their donors to spread the word about the cause or host campaigns online designed to raise awareness. Our results suggest that such efforts would likely be aided by more direct and consistent messaging reminding donors that sharing about charity means doing more good. Although the WOM benefits of sharing may seem obvious on reflection, worrying about what others will think can be a distraction in the moment.

Another critical direction for increasing WOM will be to explore what information people choose to communicate when talking about charity and whether marketing messages can enhance recruitment by suggesting what they might say (e.g., by providing default messages to post after click-through). In our dataset, we cannot see what people choose to post on social media or share in their personal communications, but we can observe that fewer than 20% of those who click-through ultimately recruited a donation. Closing this gap will likely require helping donors become more persuasive recruiters, beyond just increasing their willingness to share about charity at all. However, marketers need to proceed carefully in this space, as tactics aimed at making donors more

persuasive recruiters may also make them more reluctant to share in the first place. To illustrate, consider two messages:

1. *"I just gave $50 to St. Jude's Children's Hospital. I hope a few of my friends will join in and match me. @Sally? @Andre? Get on board folks: It's a great cause!"*
2. *"Check out St. Jude's Children's Hospital. It's a great cause!"*

The features that might make the first message more direct and persuasive – (1) sharing a dollar amount donated, (2) including an explicit call to action, (3) putting particular friends on the spot to give – might also make it feel more uncomfortable to share. Similarly, tactics which make donors more willing to share may also make them less persuasive. For example, giving donors an avenue to amplify a cause anonymously may reduce worries about bragging, but it may also diminish their capacity to use social standing and relationships as a point of influence for the cause. In short, there may be important trade-offs between what donors are willing to say and what will most effectively bring in recruited donations. The most promising approaches would avoid this paradox, both increasing people's willingness to share and making them more effective recruiters when they do. Building on our framework, an intriguing extension of our intervention would be to encourage donors to cite social impact as the reason for their sharing, perhaps in a way that both makes them more comfortable sharing at all and leads observers to see their choice to share as more altruistic and compelling.

A further, complementary approach would be to explore whether marketers can increase WOM about charity by reducing apprehension about bragging. That is, while our

intervention was designed to amplify the salience and import of social impact, it could be augmented with messaging that eases donors' worry about the appearance of self-promotion. Some organizations have found success with such tactics already. For example, well-known viral campaigns like *The Ice Bucket Challenge* – in which people post videos dumping ice on their heads for ALS research – or *Movember* – in which people grow unbecoming moustaches to raise awareness for prostate cancer – explicitly introduce embarrassment or self-effacement into the sharing campaign. For another example, Facebook encourages users to post about 'donating their birthday,' forgoing selfish gifts in exchange for donations to good causes. Our point is not that any of these specific campaigns can or should be scaled across the non-profit sector. Rather, we want to highlight that non-profits designing marketing campaigns to go viral might consider getting creative with elements of silliness, embarrassment, or self-sacrifice. Although it may seem surprising that adding humorous indignities or self-effacements might boost WOM, we suspect that such strategies can succeed by helping consumers weaken or displace the signal that their sharing is aimed at self-promotion. Future researchers can look to combine messages that highlight how sharing can benefit the cause with features that diminish inferences of bragging to further spur WOM.

To what extent do our effects generalize beyond this particular setting? Our framework applies to cases in which people have acted generously in some way (i.e., contributed to a public good) and are considering sharing about it with others. Beyond charitable giving, this includes things like signing petitions, volunteering, purchasing fair-trade products, going green, etc. It might also apply to certain political activities like

voting or protesting or giving money to a political campaign. Future research should explore people's sharing decisions in the context of more divisive contributions (e.g., giving money to the NRA, attending an abortion rights rally). Such cases involve more nuanced reputational calculations: We often want to signal that we hold 'the right' moral positions among valued ingroups, but may not want to be seen as 'taking a stand' and provoke reproach from those with whom we disagree (Silver and Shaw 2022).

Messages about social impact would not logically apply in cases where consumers are making purchases for the self. Central to our theorizing and explicated in the treatment message is the idea that talking about one's generosity is a way to have more impact toward a social cause to which one has just contributed. That is, the consequence of sharing we make salient (i.e., doing more good) aligns with the goal of donating. By contrast, after making a purchase for the self – say, buying a new pair of sneakers or taking a cooking class – telling others does not necessarily further the goal of the purchase. Moreover, as evident in our pilot study, people feel particularly uncomfortable sharing about their donations to charity, more so than about many other ordinary purchases. We believe that this stems from a pervasive view that generosity is supposed to be selfless, and that publicizing it reveals that a person may have an ulterior (selfish) motive (Berman and Silver 2022). The decision to share about other purchases is therefore unlikely to involve the same level of apprehension. Still, although social impact per se may not be relevant in other purchase contexts, our broader conceptual approach may suggest a path forward for increasing WOM about other kinds of expenditures: Explore what sorts of considerations come to mind naturally for people when deciding

whether to share about a given purchase, and ensure that those which favor sharing are most salient when asking them to do so.

Although scholars argue that social impact *should* guide decisions about charity in principle (MacAskill 2015), there is ongoing debate as to how much consumers care about maximizing their impact in practice. Evidence suggests, for example, that donors are relatively scope insensitive in their charitable contributions (Jung et al. 2017), that they care more about having *some* impact than how much (Zlatev et al. 2020), and that they will prioritize personal preferences over effectiveness in selecting charities to support (Berman et al., 2018). On the other hand, donors do seem motivated by incentives to amplify their impact via donation matches, suggesting that social impact information can sometimes influence donor behavior (Karlan and List 2007). Whereas this prior work examined how considering social impact affects decisions about whether and where to donate, our contribution to the debate focuses on donors' willingness to talk about their giving. We find that when making sharing decisions, donors often fail to consider that sharing about charity can amplify their impact. However, simple reminders to construe the decision in terms of social good can increase donors' willingness to share about charity.

Conclusion.

Encouraging word-of-mouth is a central marketing objective. In the context of fundraising for charity, it also represents a critical way for generosity to spread. Our experiments document an important psychological bottleneck which stands in the way:

Donors worry about their reputation and often fail to consider the social impact of sharing as a result. Fortunately, a simple message can reorient consumers to their ability to influence others, encourage WOM, and ultimately boost fundraising for worthy causes.

References.

Agerström J, Carlsson R, Nicklasson L, Guntell L (2016) Using descriptive social norms to increase charitable giving: The power of local norms. *Journal of Economic Psychology*. 52:147-153.

Andreoni J, Rao JM (2011) The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of Public Economics*. 95(7-8):513-520.

Andreoni J, Rao JM, Trachtman H (2017) Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*. 125(3):625-653.

Ariely D, Bracha A, Meier S (2009) Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*. 99(1):544-5.

Bagwell LS, Bernheim BD (1996) Veblen effects in a theory of conspicuous consumption. *American Economic Review*. 1:349-73.

Berger J (2014) Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*. 24:586-607.

Berger J, Schwartz EM (2011) What drives immediate and ongoing word of mouth? *Journal of Marketing Research*. 48(5):869-80.

Berman B (2016) Referral marketing: Harnessing the power of your customers. *Business Horizons*. 59(1):19-28.

Berman JZ, Levine EE, Barasch A, Small DA (2015) The Braggart's dilemma: On the social rewards and penalties of advertising prosocial behavior. *Journal of Marketing Research*. 52:90-104.

Berman JZ, Barasch A, Levine EE, Small DA (2018) Impediments to effective altruism: The role of subjective preferences in charitable giving. *Psychological Science*. 29(5):834-44.

Berman R, Pekelis L, Scott A, Van den Bulte C (2021) p-hacking and false discovery in A/B testing. Working Paper.

Critcher CR, Dunning D (2011) No good deed goes unquestioned: Cynical reconstruals maintain belief in the power of self-interest. *Journal of Experimental Social Psychology*. 47:1207-1213.

Dubé JP, Luo X, Fang Z (2017) Self-signaling and prosocial behavior: A cause marketing experiment. *Marketing Science*. 36(2):161-186.

De Freitas J, DeScioli P, Thomas KA, Pinker S (2019) Maimonides' ladder: States of mutual knowledge and the perception of charitability. *Journal of Experimental Psychology: General*. 148:158.

DellaVigna S, List JA, Malmendier U (2012) Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*. 127(1):1-56.

Elster J (1989) Social norms and economic theory. *Journal of Economic Perspectives*. 3(4):99-117.

Engagment Labs (2017) New study finds that 19% of sales driven by consumer conversation. Retrieved from: https://www.engagementlabs.com/press/new-study-finds-19-percent-sales-driven-consumer-conversations-taking-place-offline-online/

Flynn FJ, Lake VK (2008) If you need help, just ask: Underestimating compliance with direct requests for help. *Journal of Personality and Social Psychology*. 95(1):128.

Nair HS, Manchanda P, Bhatia T (2010) Asymmetric social interactions in physician prescription behavior: The role of opinion leaders. *Journal of Marketing Research*. 47(5):883-95.

Gershon R, Cryder C, John LK (2020) Why Prosocial Referral Incentives Work: The Interplay of Reputational Benefits and Action Costs. *Journal of Marketing Research*. 57(1):156-172.

Gilovich T, Medvec VH, Savitsky K (2000) The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*. 78(2):211.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Science*. 23(4):545-60.

Godes D, Mayzlin D, Chen Y, Das S, ... Verlegh P (2005) The firm's management of social interactions. *Marketing Letters*. 16(3):415-428.

Godes D, Mayzlin D (2009) Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science*. *28*(4):721-739.

Goldstein NJ, Cialdini RB, Griskevicius V (2008) A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*. 35(3):472-82.

Goodwin GP, Piazza J, Rozin P (2014) Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*. 106(1):148.

Huang S, Aral S, Hu YJ, … and Brynjolfsson E (2020) Social advertising effectiveness across products: A large-scale field experiment. *Marketing Science*. *39*(6):1142-1165.

Iyengar R, Van den Bulte C, Valente TW (2011) Opinion leadership and social contagion in new product diffusion. *Marketing Science*. 30:195-212.

Jung MH, Nelson LD, Gneezy U, Gneezy A (2017) Signaling virtue: Charitable behavior under consumer elective pricing. *Marketing Science*. 36(2):187-194.

Karlan D, List JA (2007) Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*. 97(5):1774-93.

Karlan D, McConnell M, Mullainathan S, Zinman J (2016) Getting to the top of mind: How reminders increase saving. *Management Science*. 62(12):3393-3411.

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*. 130(4):1941-1973.

Kraft-Todd G, Yoeli E, Bhanot S, Rand D (2015) Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*. 3:96-101.

MacAskill W (2015) *Doing good better: Effective altruism and a radical new way to make a difference*. Guardian Faber Publishing.

Munz KP, Jung, MH, Alter AL (2020) Name similarity encourages generosity: A field experiment in email personalization. *Marketing Science*.

Nielsen (2012) Global Trust in Advertising and Brand Messages. Retrieved from: https://www.nielsen.com/us/en/insights/report/2012/global-trust-in-advertising-and-brand-messages-2/#

Rosenbaum PR (2007) Interference between units in randomized experiments. *Journal of the American Statistical Association*. 102(477):191-200.

Schlosser, AE (2020) Self-disclosure versus self-presentation on social media. *Current Opinion in Psychology*. 31:1-6.

Silver I, Newman GE, Small DA (2021) Inauthenticity Aversion: Consumer reactance to tainted actors, actions, and objects. *Consumer Psychology Review*. 4(1):70-82.

Small DA, Berman JZ, Levine EE, Barasch A (2018) Should you broadcast your charitable side? *Behavioral Scientist*. Retrieved from:https://behavioralscientist.org/should-you-broadcast-your-charitable-side

Sudhir K, Roy S, Cherian M (2016) Do sympathy biases induce charitable giving? The effects of advertising content. *Marketing Science*. 35:849-869.

Szilagyi PG, Bordley C, Vann JC, Chelminski A, … Rodewald LE (2000) Effect of patient reminder/recall interventions on immunization rates: a review. *JAMA*. 284(14):1820-1827.

Toubia O, Stephen AT (2013) Intrinsic vs. image-related utility in social media: Why do people contribute content to Twitter? *Marketing Science*. 32(3):368-92.

Wolters HM, Schulze C, Gedenk K (2020) Referral reward size and new customer profitability. *Marketing Science*. 39(6):1166-80.

Yang, AX, Hsee, CK (2022) "Obligatory Publicity IncreasesCharitable Acts." *Journal of Consumer Research* 48.5 (2022): 839-857.

Zaki J, Cikara M (2020) Don't be afraid to virtue signal. *TIME*. Retrieved from:https://time.com/5859459/in-defense-of-virtue-signaling-2/

CHAPTER 2: SELF-CENSORSHIP AND THE STRATEGIC OMISSION OF FACTS FROM COMMUNICATION

Ike Silver, Deborah A. Small, & Geoff P. Goodwin

Abstract.

People often acquire and exchange facts, statistics, and ideas about charged issues like public health, climate change, and racial justice. In turn, such exchanges inform beliefs and attitudes that guide a variety of consequential consumer decisions. Seven pre-registered experiments examine how word-of-mouth communications about charged issues are distorted by moral attitudes and social pressure. Specifically, we investigate *self-censorship*, people's tendency to omit relevant, attitude-incongruent facts from communication. We argue that although people reflectively endorse norms of unbiased evidence-sharing, a desire to avoid social sanction for surfacing "the wrong sorts of facts" leads them to avoid sharing pertinent information with valued ingroups if it cuts against their shared opinions. This tendency cannot be explained by selective attention or motivated reasoning: Across three paradigms and six issues, we find that even after affirming that attitude-incongruent facts are true and relevant, participants hesitate to share them. By contrast, and consistent with impression management, self-censorship is moderated by observability and mediated by expectations of social judgment. Finally, people see self-censorship as a perfectly acceptable distortion: When asked if omitting relevant facts because they go against one's beliefs is wrong, they disagree. Whereas

prior work on misinformation has focused on the sharing of false information (i.e., fake news), the non-sharing of true information represents a parallel threat to information ecosystems, one which is harder to detect, but which may nonetheless distort beliefs, entrench disagreements, and harm consumer welfare.

Introduction.

Whether posting an article about the latest political controversy or discussing current events over dinner with friends, people often acquire and share information about charged social issues. Such exchanges of data and evidence about issues like climate change, public health, or racial justice inform attitudes which fundamentally shape consumer decision-making. Indeed, recent scholarship illustrates that moral and political sentiments guide a wide variety of consequential judgments and choices, from which brands to support and what to buy from them (Crockett and Wallendorf 2004; Jost 2017; Jost, Langer, and Singh 2017; Kidwell, Farmer, and Hardesty 2013, Luttrell, Teeny, and Petty 2021; Shavitt 2017; Silver, Newman, and Small 2021), to whether to trust experts (Scheufele and Krause 2019), comply with medical advice (Fridman, Gershon, and Gneezy 2021; Gollwitzer et al. 2020), or even prepare for natural disasters (Long, Chen, and Rohla 2020). More broadly, the social transmission of morally charged information can spur social movements, crash markets, topple public figures, and build or shake faith in public institutions and private firms alike.

All else equal, processes of information sharing would ideally proliferate the most accurate and relevant facts. After all, increasing the body of information available to consumers, managers, and policy-makers should only increase the expected utility of their decisions (Good 1967; Mesmer-Magnus and DeChurch 2009). But recent studies reveal that people often share in ways that distort the evidence and lead to biased beliefs and choices. In particular, scholars have highlighted the problem of 'fake news,' people's pernicious tendency to propagate false or misleading information (e.g., Pennycook and Rand 2021), and recent years have seen an explosion of interest in psychological factors which give rise to 'misinformation ecoystems' (Lazer et al. 2018).

Here, we investigate a parallel threat to the quality of information that circulates in social networks: *Self-censorship*. But whereas the problem of fake news – and the lion's share of research on misinformation thus far – concerns the *sharing of falsehoods*, our work explores people's *reluctance to share facts*, particularly those that cut against morally charged ingroup attitudes. Specifically, we explore people's tendency, when communicating about charged issues with like-minded peers (i.e., those in one's moral or political ingroup), to omit attitude-incongruent facts from discussion. As we will show, people self-censor even information they affirm to be factual and relevant if it contravenes valued political attitudes or identities. This phenomenon, we will argue, emerges primarily out of a desire to manage impressions and avoid social sanction for surfacing "the wrong sorts of facts." Moreover, while people report that sharing false information is wrong (e.g., Pennycook and Rand 2021), we will find that they see self-

censorship as perfectly acceptable, suggesting that different interventions may be required to remedy these different threats to the information ecosystem.

To illustrate, imagine a staunch gun control advocate reads an article in *the Wall Street Journal* suggesting that banning assault weapons – a policy goal held dear by the gun control movement – might not impact total annual gun deaths very much, because assault weapons are responsible for only a very small proportion of them (Koper, Woods, and Roth 2004). How likely would she be to bring up this information while talking about the issue with like-minded peers or to post about it on social media? On one hand, if she is concerned with reducing gun deaths or simply getting the relevant facts on the table, this information might seem especially important to share. On the other, if she is concerned with *appearing* committed to the gun control cause, this information might seem potentially threatening, and sharing it might harm her reputation. Of course, the conflict between sharing the facts and signaling one's loyalties is not restricted to liberals or to the issue of gun control. A similar dilemma might arise for a freedom-of-choice conservative deciding whether to acknowledge the possible benefits of COVID-19 shots among anti-vaccine friends; or for a climate change activist choosing whether to inform fellow environmentalists that nuclear power plants are surprisingly clean and efficient. Across a variety of morally-charged issues, raising relevant facts may risk sending detrimental social signals. We predict and find that in such cases, impression management considerations (e.g., Berger and Heath 2007; Leary and Kowalski 1990; Savitsky, Epley, and Gilovich 2001; Zwebner and Shrift 2020) will often win out.

Importantly, we aim to investigate self-censorship effects separately from processes of motivated reasoning (e.g., Kunda 1990). Prior work in information processing has consistently documented people's persistent tendency to avoid, discredit, and forget information which contradicts their attitudes (e.g., Lord, Ross, and Lepper 1979; Nyhan and Reifler 2010; Tetlock et al. 2000). But some attitude-incongruent facts cannot be reasoned away or forgotten. Our experiments aim to show that people are hesitant to share attitude-incongruent facts *even when they believe them to be accurate and relevant.* Beyond motivated assimilation of information, we investigate a motivated sharing effect, or more accurately a motivated *withholding* effect, which leads people to share incomplete evidence and, in so doing, bias the flow of information about important issues. We contend that self-censorship – the non-sharing of true information – constitutes the 'other half' of the misinformation ecosystem, a threat which is harder to observe than fake news, and one which, we predict and show, communicators feel less compelled to correct.

We present seven pre-registered experiments which establish that people self-censor, investigate why, and examine their perceptions of this behavior. Before describing our methods and results, we review relevant literature and derive predictions.

Conceptual Background.

Motivated Information Processing.

When people encounter information that threatens their attitudes, they deploy an impressive ability to avoid, misinterpret, and forget. When searching for information,

people seek attitude-congruent information (Nickerson 1998) and avoid attitude-incongruent information (Golman, Haggman, and Loewenstein 2017; Woolley and Risen 2021). Moreover, when they encounter attitude-incongruent evidence, they deny and discredit. For example, people misinterpret scientific research if the data conflict with their perspectives on issues like gun ownership or climate change, even when the correct interpretations of the results are clearly explained to them (Washburn and Skitka 2018). People will sometimes even forget information which threatens their beliefs or identities (Dalton and Huang 2014; Rotella and Richeson 2013).

Importantly, motivated belief formation seems to be amplified in social contexts. That is, people are driven to protect valued social identities and fit in with others, and they often adopt the views of their ingroup as a result. Theories of belief-based utility (Loewenstein, 2017), identity protective cognition (Kahan 2017), socially adaptive belief (Williams 2020), and accountability pressure (Lerner and Tetlock 1999) all argue that psychological processes of belief formation are sensitive not only to objective evidence but also to the social rewards and penalties of adopting different attitudes. For example, Mackie and Cooper (1984) found that college undergraduates were more likely to change their opinions after listening to a taped discussion attributed to ingroup (vs. outgroup) members. Similarly, Tetlock and colleagues (1989) demonstrated that when expected to justify their views on controversial political issues to others, participants tended to conform their beliefs towards those of their prospective audience. Generally speaking, such social effects on belief formation are stronger for hot-button issues which touch on valued identities (Tajfel et al. 1979).

Although a great deal of work has examined how people reconcile attitude-incongruent information internally, less is known about whether attitude-(in)congruence impacts how communicators *transmit* facts that they affirm to be true and relevant. Indeed, to the extent that past research has documented effects of ideology on communication, it has often understood them to be mediated by motivated reasoning – that people *believe* the evidence for their side and so communicate it more frequently (e.g., Myers and Lamm 1976, Janis 1972). And while some more recent work has begun to examine possible direct effects of political ideology on headline-sharing (Marie, Altay, and Strickland 2021, Ekstrom and Lai 2021), the role of impression management concerns in guiding the strategic omission of facts has yet to be explored directly. Our experiments ask whether, holding perceptions of accuracy and relevance constant, social concerns might bias people against sharing relevant facts and evidence. If self-censorship exists, it represents an additional source of biased information transmission, one which might further distort and polarize.

Drivers of Sharing

According to a recent Pew Research survey ("The Political Environment on Social Media" 2016), more than half of consumers say they encounter political content when they log in to Facebook or Twitter. And although people often self-select into consuming media and following others who generally agree with their viewpoints (Mosleh et al. 2021), there remain ample opportunities for them to encounter and pass

along facts which run counter to their moral and political attitudes[6]. What factors predict whether such attitude-incongruent evidence is propagated?

There is reason to believe that sharing might track perceptions of accuracy and relevance closely. People often try to communicate information they believe to be true and useful (Lovett, Peres, and Schachar 2013), and, at least upon reflection, they endorse ideals of accurate and impartial fact-sharing (Pennycook et al. 2021). Accordingly, in a brief pilot study motivating our research (reported in full online), we found that >90% of participants, when offered a list of possible objectives for talking about social issues, indicated that 'unbiased and accurate fact-sharing' was *the most important goal* communicators should strive to follow. Such sentiments stand to reason: For individual communicators, conversational norms dictate the inclusion of relevant information (Grice 1975) and moral norms prohibit concealing the truth (John, Barasz, and Norton 2016). There is even direct field evidence that people strive to share what's true in the domain of political issues. Recent work on fake news, for example, finds that simple reminders to consider factual accuracy reduce the spread of disinformation on Twitter (Pennycook et al. 2020, Pennycook et al. 2021). Taken together, such results suggest that people value accuracy and generally aim to transmit what they believe to be true and relevant.

At the same time, studies of word-of-mouth reveal that myriad factors beyond perceived accuracy influence what people end up sharing (e.g., affective arousal, Berger and Milkman 2012; need for uniqueness, Cheema and Kaikati, 2010; controversy, Chen

[6]For instance, an opinion article about immigration policy might qualify its position with important countervailing evidence; or a news outlet might report the results of a surprising scientific study which disconfirms viewers' assumptions about climate change.

and Berger 2013). One central driver of sharing behavior is impression management and social signaling (Berger 2014; De Angelis et al., 2012; Toubia and Stephen 2013). People care about appearances: They want to be liked and trusted, and they want to be thought of as smart, capable, high-status, well-connected, and morally upstanding (Baumeister and Leary 1995). Such goals guide communication across a host of contexts. People tweet about how busy they are to signal productivity and value (Bellezza, Paharia, and Keinan 2017); they post photos to signal taste and consumption expertise (Barasch, Zauberman, and Diehl 2018); and they conspicuously consume products and support brands to signal status (Bagwell and Bernheim 1996) and identity (Berger and Heath 2007). There is also evidence that people conspicuously *withhold* self-relevant information that might paint them in a negative light. For instance, people avoid talking about personal achievements or good deeds to deflect the appearance of bragging (Berman et al. 2015; Roberts, Levine, and Sezer 2020); and they avoid sharing past mistakes (John, Barasz, and Norton 2017), embarrassments (Miller and McFarland 1987), deviant attitudes (Prentice and Miller 1993), or honest feedback (Levine and Cohen 2018) to deflect perceptions that they are incompetent, immoral, or unkind.

It is perhaps unsurprising that impression management goals would inhibit the sharing of self-relevant consumption choices or unflattering attitudes or behavior, which seem like diagnostic signals of the sharer's personal qualities and character. But it less obvious that people would avoid sharing facts. For one thing, factual evidence does not pertain as directly to the sharer's individuating characteristics as do specific choices, opinions, or behaviors; and for another, sharing factual information might readily be seen

in a positive, helpful light. For instance, a person's willingness to share attitude-incongruent, factual evidence might be seen as a costly signal of their impartiality or open-mindedness, or it might communicate their authentic commitment to the underlying *values* of a cause (e.g., a desire to save lives from gun violence vs. a desire to *look like* a good gun control advocate; Silver, Newman, and Small 2021). By contrast, we will argue that communicators treat attitude-incongruent facts as reputational threats.

Present Research

In this paper, we investigate self-censorship, its constituent mechanisms, and people's judgments about its acceptability. We propose that although people say they should share accurately in the abstract (Pennycook et al. 2021), in practice, they will often omit relevant evidence if it contravenes moral attitudes held jointly with valued ingroup audiences. People self-censor because doing so allows them to avoid being perceived negatively by others in their social groups – those typically like-minded peers with whom they interact most frequently and rely on for information. For example, a political conservative who believed that Donald Trump lost the 2020 election fairly might not want to share evidence of the election's legitimacy with predominantly conservative followers on Facebook; or a supporter of COVID-19 vaccination efforts may not want to mention that inoculations reliably cause a non-trivial number of anaphylactic reactions to avoid the judgment of pro-vaccine friends. In such cases, we contend, communicators fear that *sharing* may be construed, not as an effort to get the facts on the table, but as tacit endorsement of a morally opprobrious position.

Being mistaken for a member of an opposing political ideology is a reasonable worry. Indeed, receivers of information sometimes 'shoot the messenger,' disliking and distrusting bearers of bad news (John, Blunden, and Liu 2019) and prefer others who affirm their pre-existing preferences when making decisions (Mojzisch et al. 2015). People especially dislike political outgroup members (Finkel et al. 2020; Rogowski and Sutherland 2016), and they voice moral outrage at disagreeable news articles, public figures, and corporate policies (Crockett 2017). To the extent that bringing up an attitude-incongruent fact might be misunderstood as espousing an opinion that threatens ingroup values (Tetlock et al. 2000), sharers will be inclined to avoid sharing altogether.

If this hypothesis is right, it suggests that we need to be concerned not only with peoples' tendency to share false information – fake news – but also with their tendency *not to share* useful and true information. Perhaps more troublingly, owing to its status as a lie of omission (Levine et al. 2018), we predict and find that self-censorship is considered much less wrong than fake news sharing, suggesting that people may be considerably less motivated to correct it. Investigating self-censorship can thus help to broaden our understanding of how people take in and transmit information about issues of considerable importance. It may also advance our understanding of the behavioral processes that give rise to 'echo chambers' and point us towards useful interventions to foster healthy information ecosystems. We return to broader implications and potential debiasing strategies in the General Discussion.

Overview of Experiments

We report eight preregistered experiments which explore self-censorship across issues and paradigms. Our studies aim to establish that people are reluctant to share attitude-incongruent evidence even when they believe it to be true and relevant, and that this effect is driven by impression management concerns. We focus specifically on communication with like-minded ingroup members, whom people preferentially rely on for information and interact with most frequently in word-of-mouth contexts online and in person. Study 1 demonstrates a self-censorship effect in people's communications about COVID-19 policy. Study 2 replicates the effect with information that is hard to reason away as factually inaccurate – the race of a criminal suspect. Studies 3 and 4 further replicate the effect across new contexts, issues, and paradigms, with further controls for motivated perceptions of accuracy and relevance. Studies 5 and 6 demonstrate process, showing that the tendency to self-censor is moderated in the absence of observers and mediated by predictions about social judgment. Finally, Studies 7 and 8 ask if people see self-censorship as a bias worth correcting, exploring how people say they should communicate about attitude-incongruent evidence and whether they judge self-censorship as morally wrong.

Although not their primary focus, our experiments also provide additional indirect evidence for motivated reasoning effects. Relevant analyses are reported in the web appendix. Importantly, our account does not deny that people's prior positions might lead them to see attitude-incongruent information as less true or less relevant. Rather, we seek to show that people self-censor beyond the influence of these effects. All reported

experiments were pre-registered. We report all conditions, exclusions, and response variables. Pre-registrations, stimuli, data, and the web appendix are posted at: https://researchbox.org/311&PEER_REVIEW_passcode=BXQOFM

Study 1

Study 1 provides an initial demonstration of self-censorship in people's communications about public health policy at the height of the COVID-19 pandemic. Participants read an article about the effects of lax shutdown policies in a foreign country and subsequently summarized what they had read for a peer. We predicted that participants would be less likely to communicate public health outcomes (which we manipulated between-subjects) when these contradicted (vs. aligned with) their incoming perspectives about how actively governments should respond to the pandemic.

Method.

Three hundred and fifteen participants were recruited from Amazon's Mechanical Turk (mean age = 35.7, 35% female). Participants first reported their attitudes about the best course of action for the US in reopening its economy after the first wave of the COVID-19 pandemic, indicating either that opening up 'faster and more optimistically' (26.7%) or 'slower and more cautiously' (73.3%) would be best. They then read a short, real news article taken from *Vox.com,* which we lightly altered according to condition. The article described how Sweden had made the decision, unique among its neighbor countries, to avoid shutting down altogether. Participants learned, for example, that

Sweden had kept open movie theaters, cafes, and ski resorts during the pandemic, and that Sweden's approach sought to balance protecting Sweden's economic interests against managing COVID-19's spread.

Participants were randomly assigned to one of two between-subjects conditions – an attitude-congruent or an attitude-incongruent condition – which varied how the *outcomes* of Sweden's unique policy were described. At the time of the experiment, Sweden had seen *more* COVID-19 deaths than its stricter Northern neighbors (Finland, Denmark, and Norway), but *fewer* COVID-19 deaths than its stricter Southern neighbors (France, Spain, and Italy). We embedded into the otherwise-identical article one of these two facts according to condition. For example, in the attitude-congruent condition, participants who believed in slower, more cautious reopening learned that Sweden had a relatively high death rate compared to its Northern neighbors, while participants who believed in faster, more optimistic reopening learned that Sweden had a relatively low death rate outcomes compared to its Southern neighbors. In the attitude-incongruent condition, participants who believed in slower, more cautious reopening learned about Sweden's relatively low death rates, while participants who believed in faster, more optimistic reopening learned about Sweden's relatively high death rates. While both death rate facts were true at the time of the experiment, each comparison (more deaths than Northern neighbors vs. fewer deaths than Southern neighbors) supported one possible attitude about lockdowns/reopening while opposing the other.

Participants were then asked to consider having a conversation with a work colleague who generally shared their views about COVID-19 lockdowns and to write a

short entry (~25 words) describing what they would tell this person about the situation in Sweden. Our key dependent variable was a binary measure, coded by a hypothesis-blind RA, of whether participants included the death rate fact they had learned when choosing what to communicate. We predicted that participants who encountered an attitude-incongruent fact about death outcomes from Sweden's policy – one which if shared might suggest that they do not agree with their colleague's beliefs – would be more likely to omit this fact from communication.

We used an additional check at the end of the study and preregistered exclusion rules to control for the possible influences of motivated memory and belief (i.e., that participants in the attitude-incongruent condition simply forgot or disbelieved the statistic they read). Specifically, we planned to include only participants who answered a final question about Sweden's actual death rates correctly (i.e., in line with the real facts presented). Participants who read about death rates in Sweden compared to its Northern neighbors (Norway, Denmark, and Finland) were included only if they later reported in a multiple-choice question that Sweden had indeed seen *fewer* deaths; Participants who read about death rates in Sweden compared to its Southern neighbors (Italy, Spain, and France) were included only if they later reported that Sweden had indeed seen *more* deaths. Following our preregistration, we excluded an additional 46 participants who failed these checks (22 and 24 from the congruent and incongruent condition, respectively), but the same effects are observed if all participants are instead included. Finally, participants were asked to report their age, gender, and general political attitudes on a 7-point scale from -3 "Strongly Liberal" to +3 "Strongly Conservative" (mean = -

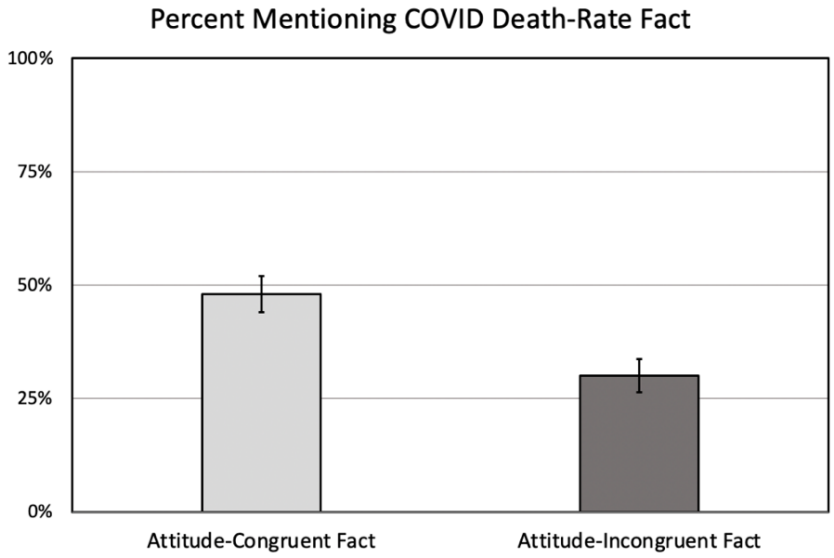
.45, SD = 1.87). Political identification was correlated with attitudes towards reopening, with liberals relatively more likely to believe in a slow and cautious approach (r = .29).

Results.

A logistic regression predicting participants' likelihood of including the focal death-rate fact in their written communication found that attitude-(in)congruence reduced sharing. Participants were less likely to mention death rate outcomes in Sweden if the comparison fact they had learned contravened their views about reopening (30% mentioning) vs. supported them (48% mentioning, B = - .79, SE = .24, Wald Z = -3.3, p < .001)[7]. See Figure 5.

[7] This effect was primarily driven by the 73% participants who favored a slower and more cautious reopening (p < .001). Among the other 27% participants, who favored a faster reopening, we did not find an effect (p > .5), though we were underpowered to detect one in this latter group. Following our pre-registered plan, we collapsed across these groups for our primary analysis. In the remainder of our studies, we find consistent evidence of self-censorship across a variety of ideologies.

Figure 5. (Chapter 2) Rates of mentioning (in a summary for others) a fact about COVID death rates, based on whether the fact was congruent or incongruent with the sharer's attitude



Note. Error bars represent standard errors.

Discussion.

The results of Study 1 support the self-censorship hypothesis. When addressing a like-minded peer, participants were less likely to communicate a fact which contradicted (vs. supported) their shared attitudes about a hot-button issue. This pattern arose among participants who, when subsequently asked, recalled and reported the focal fact correctly, suggesting that they didn't simply misremember or dismiss it when considering what to share. Our next study sought to replicate these effects in a similar paradigm, but in a new context.

Study 2

Study 2 investigated whether people would self-censor in a case where the relevant information, the suspect's race in a crime, cannot not be easily reasoned away as factually inaccurate. In other words, we sought to replicate our effect with a different sort of control for motivated reasoning. To do this, we presented participants with a news report about a real mass shooting event, manipulated suspect race, and examined whether participants would include or omit this information when recounting the event for others. In line with self-censorship, we predicted that differences in political attitudes (between liberals and conservatives) would predict whether sharers would include or omit race information when recounting a crime. As in Study 1, we further controlled for motivated memory and belief by focusing on participants who correctly reported the race information they had read at study's end.

Method.

One thousand and forty-one participants were recruited from Prolific.com (mean age 36.9, 49% female). To achieve a balance of political attitudes, we posted two work tasks: one visible only to American citizens on Prolific who identified as politically conservative, and the other visible only to American citizens on Prolific who identified as politically liberal. Both posts led to the same survey.

All participants first answered a question about their political attitudes, indicating that they were either more 'liberal' (50.3%) or more 'conservative' (49.7%) in a forced

choice. Participants then read a short excerpt from a news article on NPR.com which described a real mass shooting event, but which manipulated suspect race. The excerpt provided basic details of the crime, in which two suspects entered a Kosher deli in New Jersey and opened fire, killing five people. The five-paragraph stimulus was identical for all participants, save for one important detail manipulated between-conditions. Half of participants learned that the suspects in the shooting were "Jason N. Williams and Sue Miller," a "white couple" (White suspects condition), or "Javier N. Gonzalez and Rosa Cortes," a "Latino couple" (Latino suspects condition).[8]

Participants then wrote a short entry describing the key facts from the news report as they would communicate them to a friend or peer who shared their political beliefs (i.e., liberals addressed another liberal, and conservatives addressed another conservative). Two hypothesis-blind research assistants independently coded these written entries for whether they contained information about the suspects' race, and met to resolve disagreements. The resulting binary variable of mentioning suspect race while recounting the crime served as our key dependent variable. Prior work has found that people are sometimes uncomfortable mentioning race information in conversation (Norton et al. 2006). Here, we predicted an interaction on this variable between participants' political attitudes and condition. Specifically, we expected that liberals, who are generally concerned about portraying racial minorities negatively (Janus 2010), would be less likely to mention race in the Latino suspect condition (vs. the White suspect condition). By contrast, we suspected that conservatives, who are less concerned

[8] In reality, the crime on which our study was based was committed by a Black couple, so both conditions differed from reality. Participants were debriefed about this aspect of our design at check-out.

with racial sensitivity (and increasingly concerned with the portrayal and status of White people; Mutz 2016), might show no – or even the opposite – effect.

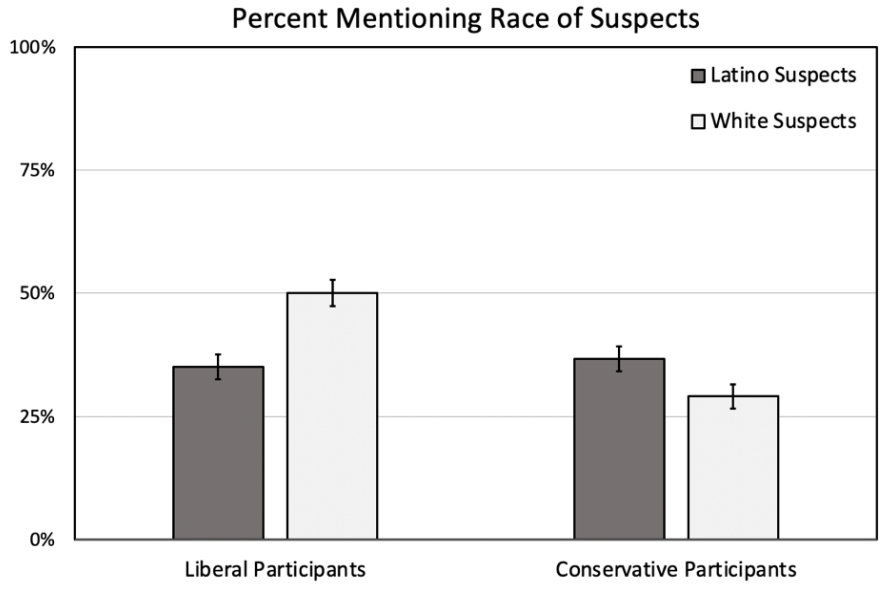
As in Study 1, to avoid any influence of motivated memory, we planned to include only participants who accurately recalled the focal factual information from the case – the manipulated names and races of the suspects – from three multiple-choice options presented after writing their entries. Following our pre-registration, 22 participants were excluded for failing this check, and an additional 29 participants were excluded for providing different political affiliations on Prolific vs. in our actual experiment. Participants were also asked whether they had heard of this crime (92% had not) and to self-report whether they had mentioned suspect race in their written summary (as an exploratory measure not formally analyzed). Finally, participants reported age, gender, and race/ethnicity demographics.

Results.

We first subjected the race-mention variable to a logistic regression using political ideology (measured: liberal vs. conservative), condition (randomly-assigned: Latino vs. White suspects), and their interaction as predictors. This model detected no main effect of condition (B = .07, SE = .065, Wald Z = 1.14, p = .25). However, in line with our predictions, the model detected a crossover interaction (B = - .24, SE = .065, Wald Z = -3.80, p < .001). Liberal participants were less likely to mention the suspects' race in the Latino suspects condition (35.2% mentioning) vs. the White suspects condition (50.1%

mentioning; χ^2 (1 df) = 12.97, $p < .001$). However, this pattern reversed for conservative participants (36.7% in the Latino suspects condition vs. 29.1% in the White suspects condition; χ^2 (1 df) = 3.39, $p = .066$). We also detected an unexpected main effect of participant ideology, such that, at baseline, conservatives were less likely to mention the suspects' race in their summary than liberals (B = -.21, SE = .065, Wald Z = 3.29, $p <$ .001). Note that because participants' political ideology was measured rather than randomly-assigned, and liberal and conservative participants might differ in many ways, it is difficult to interpret absolute levels across ideologies. Rather, our focus is the patterns of mentioning race (Latino suspects vs. White suspects, manipulated between-subjects) within each ideological group. See Figure 6.

Figure 6. (Chapter 2) Proportion including suspect race in a crime summary to be shared with others, broken down by suspect race condition (between-subjects) and political ideology (measured)



Note. Because participant ideology was measured (and not randomly-assigned), our primary interest was the differing patterns within each ideology, rather than in comparing absolute levels of mentioning race across them. Error bars represent standard errors.

Discussion.

In Study 2, attitude-incongruence reduced sharing in a case where the information in question, the race of criminal suspects, could not be easily dismissed as factually inaccurate. More specifically, participants self-censored suspect race information when it contradicted (vs. aligned with) their social group's political ideology and goals: Liberals were less likely to mention the race of Latino suspects (compared to White suspects), while conservatives displayed the opposite pattern. Importantly, our results held while focusing on participants who accurately recalled the key facts from our stimuli, suggesting that the tendency to omit these facts was not driven by thinking them inaccurate or forgetting them altogether.

Studies 1 and 2 provide initial demonstrations of self-censorship using a behavioral task: Summarizing a news story for others. One advantage of this approach is that it provides a test of self-censorship in a free response context, where participants could easily deploy other sorts of impression management strategies they might want to use instead. For example, in Study 1, there was nothing stopping participants from sharing the incongruent death rate statistic while restating their commitment to their side of the argument or while caveating the fact's importance for the broader debate about

reopening. However, even with these substitute strategies available, many chose to self-censor, omitting attitude-incongruent facts altogether.

Studies 3 and 4 took a complementary approach, this time exploring participants' willingness to share a variety of attitude-congruent and attitude-incongruent facts. Here, instead of using recall tasks to account for motivated reasoning, we directly measured perceptions of factual accuracy, and, as an additional assurance that our effects are not the product of motivated beliefs, perceptions of relevance. We sought to show that even when people judge attitude-incongruent evidence to be factual and directly relevant, they may still exhibit reluctance to share.

Studies 3 and 4

Studies 3 and 4 sought to replicate the self-censorship effect in new paradigms which took a different, complementary approach to controlling for motivated reasoning. In these studies, we presented participants with evidence pertinent to hot-button moral issues and measured willingness to share as a function of attitude-(in)congruence. But whereas Studies 1 and 2 used recall tasks to control for motivated reasoning, Studies 3 and 4 measured perceptions of accuracy and relevance directly. Study 3 tested participants' willingness to share facts that either supported or else undermined Donald Trump's performance as president while statistically controlling for and conditioning on perceived accuracy and relevance. Study 4 tested whether participants would reshare a social media post about human gene-editing that either highlighted or omitted an attitude-

incongruent fact about the issue, immediately after agreeing that the fact was true and relevant.

Study 3 Method.

Three hundred participants were recruited from MTurk (mean age = 34.3, 48% female). Participants reported their attitudes about Donald Trump's presidency by completing the statement "Generally, I think Donald Trump has done a _____ job as President of the United States" on a 6-point scale from 'truly terrible' to 'truly great.' 19.3% gave an answer greater than 3, indicating a favorable view of the president, while 80.3% gave an answer less than 4, indicating an unfavorable view of the president. Participants then read an excerpt from an article published in *The New Yorker* discussing reasons Donald Trump might or might not deserve credit for positive US economic performance during his administration's tenure.[9] The excerpt provided evidence for both sides of the issue and was identical for all participants.

After reading, participants encountered six facts taken directly from the article and made a series of judgments about them. Three of these facts suggested that Trump deserved credit for economic growth, while three suggested that Trump did not deserve credit. The exact items are listed below:

<u>Anti-Trump:</u>

1. *The current upward trajectory of economic growth began well before Donald Trump was inaugurated, during Barack Obama's presidency in 2009.*

[9] At the time of the experiment (Fall 2019; prior to the COVID-19 pandemic), The US was experiencing record-breaking stock market growth, but the public was polarized as to whether economic prosperity should be credited to the president and his policies.

2. *Job growth has gradually slowed since Donald Trump's inauguration.*
3. *The US economy created nearly 30,000 more jobs per month during the final 29 months of Obama's presidency than during the first 29 months of Trump's presidency.*

Pro-Trump:

4. *The current upward trajectory of the economy has, during President Trump's administration, grown to be the longest expansion – 10 years and 1 month – since the government began keeping records.*
5. *Early last week, the stock market hit a new all-time high, something that has happened multiple times under President Trump.*
6. *Job growth tripled between May and June of 2019, the third year of Trump's presidency.*

For each participant, we coded whether a given item aligned with or contradicted their incoming attitude toward Trump's performance as president. This binary coding served as a within-subjects manipulation of attitude-(in)congruence[10]. Each participant saw three facts congruent with and three facts incongruent with their attitudes about Trump. After reading, participants made a series of judgments, rating each item on three dimensions: perceived accuracy, perceived relevance, and willingness to share. Judgments (accuracy, relevance, willingness to share) were presented in randomly ordered blocks. Within each block, items were presented in random order on consecutive pages. The accuracy question asked participants to rate agreement on a 7-point Likert scale with the statement: "This piece of information is factual: I believe it is true as

[10] We initially planned in Studies 3 and 5 to use a *degree* of attitude-(in)congruence variable (relative to participants incoming political attitudes on a continuous scale) to predict sharing; however, for simplicity, we opted to report a binary congruence variable, as both approaches yield consistent and strongly significant results (see Web Appendix for robustness checks).

stated." The relevance question asked participants to rate agreement with the statement: "This piece of information is relevant to understanding whether Trump has done a good job as president." The sharing question – our key dependent variable – asked participants to consider having a conversation with a group of acquaintances who shared their views of the president and to rate how likely they would be to bring up each fact on a 7-point scale from "Extremely unlikely" to "Extremely likely." Note that although hypothetical, self-reported willingness to share measures have been shown to reliably correlate with actual sharing behavior in the context of political issues specifically (Mosleh, Pennycook, and Rand 2020).

Participants then answered a simple true/false attention check indicating whether they had read an article containing evidence on both sides of the issue of Trump's economic performance. Following our pre-registered procedure, we excluded 24 additional participants who answered this question incorrectly. No results depend on this exclusion. Finally, participants answered the same demographic questions. Unsurprisingly, participants' overarching political identification (mean = -.90, SD = 1.67) was strongly correlated with their attitudes about Trump's performance (r = .74).

Study 3 Results.

Our key prediction was that participants would be less willing to share attitude-incongruent (vs. attitude-congruent) facts, controlling for their perceptions of accuracy and relevance. To test this hypothesis, we estimated a linear mixed-effects regression predicting willingness to share each fact as a function of attitude-incongruence, perceived

accuracy, and perceived relevance. The model also included random effects to control for repeated measures by participant, and dummy variables for each specific item as well as for participants' initial attitude about the president (favorable or unfavorable). In line with predictions, we found that controlling for differential effects of ideology on perceptions of accuracy and relevance, participants were substantially less willing to share attitude-incongruent vs. attitude-congruent facts (regression estimated means: M = 3.41, SE = .07 vs. M = 4.79, SE = .07; B = -1.38, SE = .084, t(1557.74) = -16.54, p < .001).

As a further robustness check, we refit the same model, but this time we also *conditioned on* perceptions of accuracy and relevance in addition to controlling for them. That is, we included only observations in which participants agreed (answered above the midpoint) that a given item under consideration to share was both true and relevant. This approach yielded a virtually identical effect: Participants were less willing to share attitude-incongruent vs. attitude-congruent facts (regression estimated means: M = 3.88, SE = .12 vs. attitude-congruent facts: M = 5.26, SE = .09; B = -1.38, SE = .13, t(713.85) = -10.30, p < .001).

Study 4 Method.

Our designs thus far have documented relative differences in sharing between attitude-congruent and attitude-incongruent evidence. However, they have not yet cleanly distinguished a motivation to *omit* attitude-incongruent facts from a motivation to *include* attitude-congruent ones. While both dynamics likely play a role in real-world sharing

decisions, Study 4 sought to disentangle them. Specifically, Study 4 aimed to demonstrate that adding a single incongruent fact to an otherwise identical communication reduces willingness to share.

For this study, we examined an emerging social issue: the use of the gene-editing tools to alter human DNA. Recent advances in genetics have opened the door for scientists to target and tailor specific gene sequences in viable human embryos. Although the development of this technology may ultimately help scientists cure a variety of diseases, many people find the prospect of 'genetically modified humans' disturbing, and the research remains highly controversial. In Study 4, we examined whether people who oppose gene-editing would be comfortable communicating to their social network that experimenting on human DNA may confer important scientific benefits.

Eight hundred and fifty-one MTurk participants (mean age = 40.5, 50% female) read a short blurb about gene-editing technology. The blurb explained that a new gene-editing tool called CRISPR now allows scientists to "tailor the DNA of humans before they are born" and that the technology poses "difficult choices about desirable/undesirable aspects in the human gene pool and whether scientists can/should try to fix them." Participants then indicated whether they were generally IN FAVOR or AGAINST experiments which genetically modify human embryos before birth. As we expected, a majority of participants (63.7%) opposed gene-editing experiments. Following our preregistration, these 542 anti-gene-editing participants comprised our target sample. Results from the remaining participants were treated as exploratory.

Participants then read an excerpted article from the *MIT Technology Review,* titled "Chinese Scientists are Creating CRISPR Babies." The article, which we condensed for brevity, discussed the rapid advance of CRISPR technology in China, and the surprising development that Chinese doctors were planning to birth "the first genetically tailored humans." The article primarily raised questions and concerns about CRISPR technology, noting that the technology was controversial, that international scientists had asked the Chinese team not to pursue its aims, and that one scientist had gone to jail for violating ethics rules. However, it also included indications that gene-editing experiments might yield important benefits, noting that such experiments would likely help to cure diseases like HIV, Smallpox, and Cholera. This focal fact was attitude-incongruent for our target sample of participants, who opposed the idea of editing human DNA.

After reading, we presented participants with three facts taken directly from the article. One of these facts was our target attitude-incongruent item: "*Experiments altering human DNA could help scientists learn to eradicate certain diseases like smallpox, HIV, and cholera."* The other two items were meant to serve as fillers ("*Genetically modifying human embryos before they are born is considered highly controversial"*; "*Experiments on human DNA may alter the human gene pool permanently and in unexpected ways"*). For each of these three facts, participants indicated in a binary choice whether they "mostly agree[d]" or "mostly disagree[d]" that it was (a) factually accurate and (b) relevant to understanding the risks and benefits of gene-editing technology. We also asked participants to indicate whether each fact was included in the passage – all three were – as an additional filler exercise. Our primary concern was whether participants who

explicitly agreed that the focal attitude-incongruent was accurate and relevant would be willing to share it with others.

In the final phase of the experiment, participants were asked to imagine encountering a post on social media sharing the article they had read. In the control condition, the post read:

"Whoa. I had no idea this was happening in China already. Gene-editing and eugenics is going to be a huge issue going forward. If we aren't careful, this could introduce mutations which change the human gene pool permanently...
Many potential risks of this technology to consider..."

The treatment condition was identical, except that the post also highlighted the attitude-incongruent fact (manipulation bolded here only):

"Whoa. I had no idea this was happening in China already. Gene-editing and eugenics is going to be a huge issue going forward. If we aren't careful, this could introduce mutations which change the human gene pool permanently...

Still, experiments altering human DNA could help scientists learn about diseases like HIV and Smallpox...
Many potential risks ***and benefits*** *of this technology to consider..."*

Participants rated how willing they would be to reshare the post, on a 7-point scale from "Completely Unwilling" to "Completely Willing." This designed allowed us to isolate the impact of highlighting a single attitude-congruent fact on willingness to share and to focus on responses from participants who nevertheless saw the fact as true and relevant.

Finally, participants completed age, gender, and political identification demographics. Political identification was correlated with attitudes towards CRISPR technology, with conservatives relatively more likely to be against gene-editing (r = .26).

Study 4 Results.

Following our preregistered plan, we focused our analysis on participants who agreed that the focal attitude-incongruent fact differentiating the two conditions was both factually accurate and relevant to understanding the risks and benefits of gene-editing technology. For these 452 participants the only difference between conditions was the addition of a single factual, relevant attitude-incongruent fact to the message they were asked to consider sharing. In line with the self-censorship hypothesis, however, participants were less willing to reshare a post which highlighted such a fact (M = 2.68, SD = 1.89) compared to one that didn't (M = 3.09, SD = 2.04, t(450) = 2.20, p = .028, d = .21).

As an exploratory follow-up, we broadened our analysis to include the additional 280 participants who also believed the focal fact about CRISPR's benefits to be true and relevant, but who actually expressed *support* for gene-editing at the outset of our study. Including these additional pro-gene-editing participants wiped out the overall effect of condition on sharing (t(730) = .18, p = .86, d = .013), because, consistent with our account, participants with the *opposite* attitudes from our focal sample actually showed the *opposite* pattern of sharing preferences ($M_{control}$ = 2.47, SD = 1.78 vs. $M_{treatment}$ = 3.05, SD = 1.99, t(278) = 2.58, p = .01, d = .31). That is, participants who were in favor of gene

editing technology were more inclined to share the article with a post highlighting a potential benefit of gene-editing than one which omitted this fact.

Studies 3 and 4 Discussion.

In Study 3, attitude-incongruence reduced sharing across an array of pertinent facts, even while directly controlling for and conditioning on perceptions of accuracy and relevance. In Study 4, seconds after agreeing that a specific attitude-incongruent fact was true and relevant, participants said they would be less willing to share a message that highlighted it (vs. one which omitted it altogether). Study 4 provides perhaps the cleanest demonstration so far of self-censorship: There, the inclusion of a single attitude-incongruent fact in an otherwise attitude-congruent post reduced sharing.

Taken together, Studies 1-4 provide convergent evidence of self-censorship across paradigms, issues, and controls for motivated reasoning. In Studies 5 and 6, we sought to show that self-censorship effects are driven by impression management concerns.

Study 5 and 6: The Social Underpinnings of Self-Censorship

Thus far, we have demonstrated that people will self-censor attitude-incongruent evidence that they simultaneously know to be accurate and relevant. Studies 5 and 6 aimed to show, via mediation and moderation respectively, that self-censorship effects are driven by concerns about impression management and social signaling. People avoid sharing information that might undermine the appearance of commitment to a cause

valued by their ingroup. The design of Study 5 mirrored that of Study 3: Participants read an article, judged the accuracy and relevance of a variety of attitude-congruent and attitude-incongruent facts, and then indicated willingness to share. But this time, to demonstrate that self-censorship effects depend on the presence of observers, we manipulated whether participants were sharing with a group of peers or with their future self. The design of Study 6 mirrored that of Study 4: Participants read an article, judged the accuracy and relevance of a single attitude-incongruent fact, and then indicated willingness to share a post that either highlighted or omitted that specific fact. But this time, we also measured predictions about social judgment, to show that participants expect sharing attitude-incongruent facts to provoke social reproach and that this expectation mediates the self-censorship effect. Studies 5 and 6 also generalized self-censorship to a new moral issue: gun control.

Study 5 Method.

We recruited 519 MTurkers (mean age = 36.7, 37% female). Participants first indicated their beliefs about the issue of gun control by rating their agreement on a 6-point scale (strongly disagree to strongly agree) with the statement: "*Generally, I think the United States should allow ordinary citizens to access and purchase guns with minimal restriction*." We used this item to classify participants as either against (35.6%) or in favor of (64.4%) restricting gun ownership. Participants then read an excerpt from a real article from ThePerspective.org describing factual pros and cons of gun control policies, which we condensed for brevity. After reading, participants considered six facts from the article and made a series of judgments about them. Three of these facts aligned

with arguments for gun restrictions, while the other three aligned with arguments against gun restrictions. The exact facts were:

In favor of gun restrictions:

1. *Universal background checks can help prevent criminals and the mentally ill from obtaining guns: Some estimate that background checks on gun ownership could cut gun deaths by up to 90%.*
2. *High-capacity magazines, which many gun control measures would heavily restrict, are used in over 50% of mass shootings.*
3. *Banning sales to civilians of assault weapons such as the AK-47 and the AR-15, which fire many more rounds per minute than other kinds of guns, would reduce carnage in mass shootings.*

Against gun restrictions:

4. *Chicago, a city with some of the toughest gun restrictions in the country (including on assault and high-capacity weapons), suffers from rising gun and gang violence.*
5. *57% of people feel that carrying a gun would prevent them from being victimized.*
6. *40% of convicted felons say that they would be deterred from carrying out their crimes if they thought that their potential victim was armed.*

For each participant, we coded whether a given item aligned with or contradicted their incoming attitude about gun restrictions. This binary coding served as a within-subjects manipulation of attitude (in)congruence, and the design thus included three attitude-congruent and three attitude-incongruent facts for each participant. We also manipulated between-subjects whether participants considered sharing in a private or a public context. Half of participants were told to imagine preparing a summary of key information on gun control for a group of neighbors who held similar views to theirs and would soon discuss the issue together (public condition), while the other half imagined

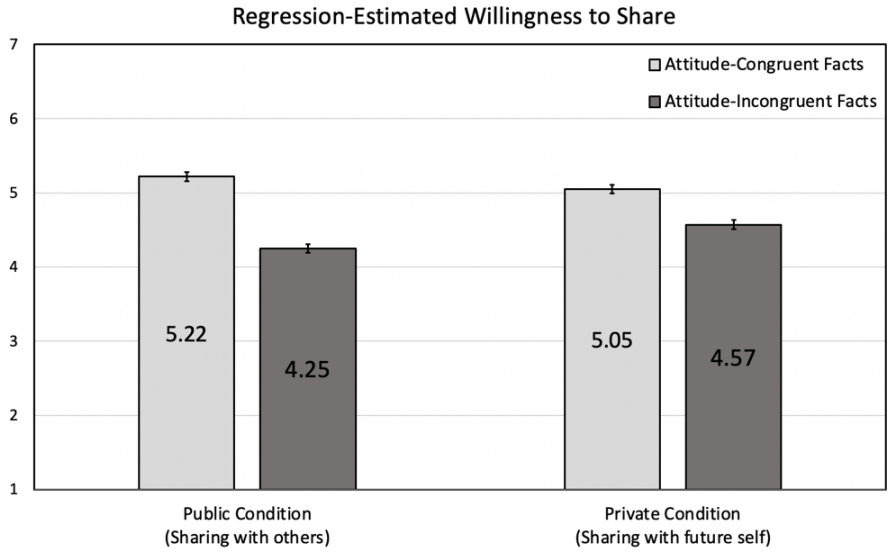
preparing a summary of key information on gun control for their own private future use (private condition). Participants rated each item on three dimensions: perceived accuracy, perceived relevance, and willingness to share. These tasks were presented in random blocks with item order randomized within blocks. The accuracy and relevance questions were identical to those used in Study 3. To measure willingness to share, participants were asked to indicate, for each of the six facts, how likely they would be to include it in their public/private summary (1-7 scale, "Very unlikely" to "Very likely"). Our key prediction was that people would be less willing to include attitude-incongruent (vs. attitude-congruent) facts in a summary to be shared with peers publicly, but that this tendency to self-censor would be reduced when considering what to include in a summary for their future self. We theorized that in the absence of observers, participants might be feel less pressure to avoid attitude-incongruent facts.

Next, participants answered a true/false attention check, indicating whether the article they had read contained both pro- and anti-gun control arguments. Following our pre-registered plan, an additional 34 participants were excluded for failing this simple check. Finally, participants answered our usual demographics, and answered a comprehension/manipulation check, reporting whether the summary from the study would be public or private. Most answered correctly (86.7% overall: 92% in the public condition and 81% in the private condition). Political identification was correlated with attitudes about gun control, with conservatives more likely to oppose restrictions on ownership (r = .45).

Study 5 Results.

To test our predictions, we estimated a linear mixed-effects regression predicting willingness to share each fact as a function of attitude-incongruence (congruent vs. incongruent), condition (public vs. private) and their interaction, while controlling for perceptions of accuracy and relevance. The model also included random effects to control for repeated measures by participant, and dummy variables for each specific item as well as for participants' initial attitude about gun restrictions (for or against). Regression-estimated means by attitude-(in)congruence and condition are displayed in Figure 7.

Figure 7. (Chapter 2) Willingness to share (1-7) broken down by condition (between-subjects) and attitude-(in)congruence (within-subjects) from Study 5



Note. Means displayed are regression-estimated: They were generated from our primary model which controls for perceptions of accuracy and relevance. Error bars represent standard errors.

In line with our hypotheses and our prior results, the model detected a main effect of attitude-incongruence (B = -.36, SE = .027, t(2738.13) = -13.31, p < .001). People were less willing to share attitude-incongruent (vs. attitude-congruent) facts. Importantly, the predicted attitude-incongruence by condition interaction was also significant (B = -.12, SE = .024, t(2590.70) = -5.05, p < .001), such that the effect of attitude-incongruence on participants' willingness to include facts in a summary was larger when preparing a public summary for others than when preparing a private summary for their own future use.

We followed up on this interaction by examining the effect of attitude-incongruence within the public and private conditions separately. Attitude-incongruence had a significant negative effect in public (B = -.48, SE = .04, t(1387.66) = -12.11, p < .001), and it also had a significant, albeit smaller, negative effect in private (B = -.24, SE = .04, t(1341.70) = - 6.48, p < .001). Reflecting our focal interaction, the private condition saw the effect of attitude-incongruence reduced by roughly half.

Finally, focusing on attitude-incongruent facts only, we found that participants were less likely to share in a public summary vs. a private one (B = -.16, SE = .05, t(509.21) = - 3.10, p = .002).

Study 6 Method.

Study 6 sought to demonstrate (a) that people anticipate more negative observer reactions when sharing attitude-incongruent content and (b) that this expectation mediates the negative effect of attitude-incongruence on willingness to share facts. To do this, we

presented participants with an article which included a target attitude-incongruent fact, confirmed that they believed this fact to be true and relevant, and then asked them to indicate how willing they would be reshare a post on social media that either highlighted or omitted it (mirroring the design of Study 4). This time, we also measured anticipated social judgment to establish the role of impression management concerns.

In Study 6, we again tested the issue of gun control, but here we planned to focus specifically on participants who supported a national assault weapons ban (the majority position in the US; Newport 2019). In particular, we sought to explore whether such participants would be willing to share the fact that banning assault weapons might have a relatively small impact on overall gun deaths (because assault weapons are responsible for only a tiny fraction of annual gun fatalities).

One thousand nine hundred and ninety-four participants were recruited from Prolific.com (mean age = 37.3, 48% female). At the outset of the study, participants read a short introductory blurb about recent mass shootings in the United States and indicated whether they generally SUPPORT or OPPOSE a national ban on assault weapons and high-capacity magazines. 77.3% of our participants indicated support for a ban, and these 1541 participants comprised our target sample. Results from the remaining participants were treated as exploratory.

Participants then read a short article which contained facts about the movement for a national assault weapons ban in the United States. For this study, the article was constructed by the experimenters to resemble a real news excerpt. Included in the article

were a variety of attitude-congruent, factually accurate arguments, such as the fact that the US previously had a national assault weapons and that reinstating such a ban is supported by a majority of Americans. The article also mentioned the fact that although assault weapons are commonly used in mass shootings specifically, they are responsible for only about 3% of total annual gun deaths, suggesting that an assault weapons ban might have a relatively small overall impact. We expected that this fact, which raises important questions about the effectiveness of an assault weapons ban for reducing gun fatalities, would be attitude-incongruent for our sample of pro-ban participants.

After reading, participants were presented with three facts from the article. One of these facts was our target attitude-incongruent item: "*Because mass shootings are relatively rare, assault weapons are responsible for only about 3% of gun deaths. Banning them may not impact total annual gun deaths all that much."* The other two facts served as fillers ("*According to Gallup, a majority of Americans support a ban on assault weapons and high capacity magazines"*; "*The US used to have a federal assault weapons ban, but it lapsed in 2004. Studies suggest that it reduced the frequency of mass shootings during that time"*). For each of these three facts, participants indicated in a binary choice whether they "mostly agree[d]" or "mostly disagree[d]" that it was (a) factually accurate and (b) relevant to understanding the possible effects of assault weapons restrictions. We also asked participants to indicate whether each fact was included in the passage – all three were – as an additional filler exercise. Our primary concern was whether participants who explicitly agreed that the focal attitude-

incongruent fact was accurate and relevant would be willing to share it with others and whether any hesitancy to share might be related to impression management concerns.

In the final phase of the experiment, participants were asked to imagine encountering a post on social media sharing the article they had read. In the control condition, the post read:

"This issue continues to be timely and relevant, so we need to keep talking about it. We suffer from too many mass shootings, and banning assault weapons would help to reduce the carnage. The majority of Americans support a ban."

The treatment condition was identical, except that the post also highlighted the attitude-incongruent fact (manipulation bolded here only):

"This issue continues to be timely and relevant, so we need to keep talking about it.

We suffer from too many mass shootings, and banning assault weapons would help to reduce the carnage. The majority of Americans support a ban.

Still, assault weapons are responsible for only about 3% of gun deaths. Banning them may not impact total annual gun deaths all that much."

Participants rated how willing they would be to reshare the friends' post, on a 7-point scale from "Completely Unwilling" to "Completely Willing." As a measure of anticipated social judgment, participants also completed the following statement, "If others in my social network saw that I had shared this, they would get a ____ impression of me and my views," answering on a 1-7 scale from "Strongly negative" to "Strongly positive."

Finally, participants completed our usual demographics. They also reported their frequency of posting articles on social media on a 1-7 scale from "never" to "all the time" as an additional exploratory variable, which was not formally analyzed (M = 2.66, SD = 1.63). Political identification was correlated with attitudes towards assault weapons bans, with liberals relatively more likely to be in favor (r = .46).

Study 6 Results.

In line with our preregistration, we focused on participants who agreed that the focal attitude-incongruent fact differentiating the two conditions was both true and relevant. For these 1272 participants, the only difference between conditions was the addition of a single factual, relevant attitude-incongruent fact to the posted they were asked to consider resharing. Replicating our previous results, participants were less willing to reshare a message which highlighted an attitude incongruent fact (M = 3.38, SD = 1.92) than one that did not (M = 4.15, SD = 2.09; t(1270) = 6.84, p = < .001, d = .38). We also found that participants anticipated more negative reactions from their social network when the message included (M = 4.28, SD = 1.40) vs. did not include (M = 4.93, SD = 1.23) the attitude-incongruent fact (t(1270) = 8.79, p = <.001, d = .49). Moreover, we found evidence of significant mediation, such that the effect of highlighting an attitude-incongruent fact on willingness to share was mediated by anticipated social judgment: The more negatively participants expected their social network to react, the less willing they were to share (Indirect Effect: B = -.49, SE = .06, 95% CI = [-.38, -.61]).

Next, we turned to the 413 additional participants who also saw the focal fact as true and relevant, but who actually *opposed* an assault weapons ban. Consistent with our account, and our results from Study 3, participants with the opposite attitudes from our focal sample showed the opposite pattern of results on both willingness to share ($M_{control}$ = 1.67, SD = 1.12 vs. $M_{treatment}$ = 2.22, SD = 1.67, $t(411) = 3.94$, $p < .001$, $d = .39$) and anticipated social judgment ($M_{control}$ = 3.45, SD = 1.42 vs. $M_{treatment}$ = 3.60, SD = 1.29, $t(411) = 1.14$, $p = .25$, $d = .11$). For these subjects, highlighting an attitude-congruent fact increased willingness-to-share but did not necessarily lead to more favorable anticipated reactions from others.

Studies 5 and 6 Discussion.

In line with our theorizing, Studies 5 and 6 provide evidence that self-censorship effects are driven in part by concerns about impression management. In Study 5, the self-censorship was effect was significantly reduced when considering sharing with peers vs. with the future self. Interestingly, in that study, we still detected a (weaker) effect of attitude-incongruence on sharing facts with the future self, which may reflect related processes of self-signaling. In Study 6, we found that adding an incongruent fact to a social media post increased concern that sharing would result in negative judgment, and that this mediated the negative effect of attitude-incongruence on willingness to share.

Having established that people will self-censor for impression management purposes, we turned in Studies 7 and 8 to investigating people's reflective judgments about self-censorship.

Studies 7 and 8: Do people recognize self-censorship as a problem?

Work in the burgeoning misinformation literature suggests that people see sharing factually dubious information as wrong and that they will try to avoid doing so if their own accuracy goals are made salient to them (Pennycook and Rand 2020). Similarly, our own pilot data suggest that people endorse accurate and unbiased fact-sharing as the most important goal people should have in mind when talking about hot-button issues. However, in our experiments, people readily self-censor even after explicitly affirming the accuracy and relevance of the information they subsequently omit, suggesting that the activation of accuracy considerations may do less to constrain the omission of true information as compared to the transmission of false information. To explore this tension between people's professed commitment to accuracy and their tendency to self-censor, we sought to investigate people's reflective judgments about the phenomenon: Do they see self-censorship as biased? Are they motivated to correct it?

We investigated this question in two ways. First, in Study 7 we manipulated whether we asked people to report what they *would* share or else what they *should* share after reading a factual, two-sided article about a contentious topic, which allowed us to test whether people self-censor more than they think they ought to. Second, in Study 8, we asked people to make judgments about another person who either chooses to transmit

false information that supports her side of an issue (i.e., to share fake news), or else to withhold true information that goes against her side (i.e., to self-censor), which allowed us to compare people's judgments of fake news sharing and self-censorship directly. Broadly, we suspected that self-censorship would be judged less harshly. This could be because self-censorship is seen as a lie of omission (leaving things out), while fake news sharing is more like a lie of commission (actively deceiving others), and we know that lies of omission are typically judged more leniently (Levine et al. 2018).

Study 7 Method.

We recruited 530 participants (mean age = 34.8, 52% female) from MTurk. For this study, we focused on the issue of protests against police violence. At the time of the experiment (summer 2020), cities across the US were experiencing protests and social unrest in response to the killing by police of unarmed Black men and women. The protest movement was a lightning-rod issue, with one side believing that the protests were a necessary and productive response to police violence, and the other believing that protests were destructive and unhelpful. We designed Study 7 to capitalize on this divergence.

Participants first reported their beliefs about the protest movement, indicating in forced choice either that they thought the protests had so far done "more good than harm" (60.4%) or "more harm than good" (39.6%). Participants then read a two-sided article, again constructed by the experimenters, which outlined both benefits and costs of the protest movement. Participants then made judgments of accuracy and relevance for six

facts (three from each side) taken from the article (see below), and they also considered sharing. As previously, these judgments were blocked and randomized, and items were coded as attitude-congruent or incongruent relative to participants' incoming attitudes.

Pros of protests:

1. *In response to the protest movement, Americans from across the political spectrum have come together to donate hundreds of millions of dollars to non-profits committed to leveling the playing field for all citizens.*
2. *Protests leaders have successfully convinced many municipalities to ban chokeholds, return unnecessary and expensive military equipment, and invest in de-escalation trainings.*
3. *Disruptive public demonstrations are more effective at getting attention and forcing conversation and reform than more peaceful forms of protests (such as those employed by NFL players) which haven't resulted in much concrete change.*

Cons of protests:

1. *Public protesting has caused hundreds of millions of dollars of property damage, much of which has afflicted Black communities and Black owned businesses in urban areas.*
2. *Violent crime has risen sharply in urban communities where protests have taken place, in part because protests provide cover for bad actors to get away with criminal behavior.*
3. *Public protests, where people are typically congregated closely together, are responsible for some of the uptick in recent transmissions of COVID-19.*

This time, the sharing task was randomly manipulated between-subjects. All participants were asked to consider having a conversation with a group of acquaintances who shared their view of the protests (either that they had done more good than harm or more harm than good). Half of participants were then asked to rate their agreement on a 7-point

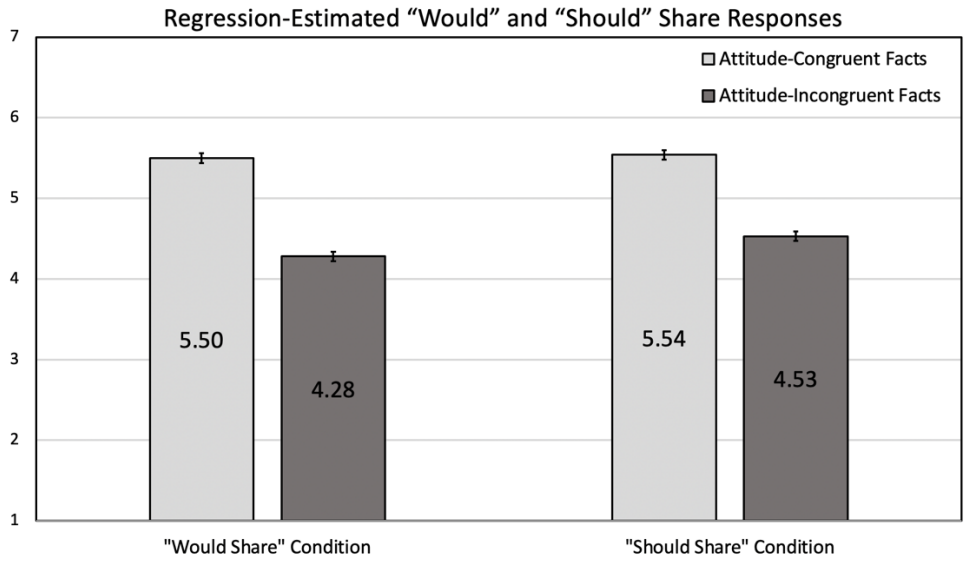
Likert scale with the statement "I **should** probably bring this fact up for discussion" while the other half rated agreement with the statement "I **would** probably bring this fact up for discussion" (bolding added here for emphasis). Using this simple manipulation, we sought to compare participants' judgments of what they should bring up in conversation about a hot-button issue to their judgments of what they actually would bring up. In line with the idea that people endorse accurate, unbiased sharing upon reflection, we thought that asking people what they *should* share might find weaker effects of attitude-incongruence than asking them what they *would* share, and so predicted a reduction of the self-censorship effect between attitude-incongruence and should vs. would share judgments. Still, owing to their status as omissions that may not obviously resemble violations of accuracy, we were unsure if cuing normative sharing behavior would totally eliminate the tendency to censor.

Participants were then asked, as a simple true/false attention check, whether the article had contained arguments on both sides of the issue. An additional 18 participants failed and were excluded per our preregistered plan. Participants also answered a simple multiple-choice comprehension/manipulation check, indicating whether they had reported what they 'probably would' or 'probably should' share. Most answered correctly (78.3% overall: would: 93.5%, should: 63.01%). Finally, participants reported age, gender, and political attitudes as previously. For this study, conservative participants were more likely to believe the protests had done more harm than good (r = .61).

Study 7 Results.

As previously, we modeled effects of interest using linear mixed-effects regression. The model predicted sharing intentions (whether participants would or should bring up each fact) as a function of attitude-incongruence (congruent vs. incongruent), condition (would vs. should share) and their interaction, while controlling for perceptions of accuracy and relevance. The model also included random effects to control for repeated measures by participant, and dummy variables for each specific item as well as for participants' initial attitude about the protests (for or against). Regression-estimated means by attitude-(in)congruence and condition are displayed in Figure 8.

Figure 8. (Chapter 2) Sharing (1-7 scale), broken down by would/should condition (between-subjects) and attitude-(in)congruence (within-subjects) in Study 7



Note. Means displayed are regression-estimated: They were generated from our primary model which controls for perceptions of accuracy and relevance. Error bars represent standard errors.

The model detected no main effect of condition (would vs. share), but a sizable main effect of attitude-incongruence, such that participants indicated that they either would or should share attitude-incongruent facts less than attitude-congruent ones (B = -.56, SE = .026, t(2788.67) = - 21.17, p < .001). As in our prior experiments, this effect held when participants were explicitly considering whether they *would* share congruent vs. incongruent facts (B = - .57, SE = .039, t(1401.65) = -15.15, p < .001). Here, participants also indicated that they *should* share attitude-incongruent facts less than attitude-congruent ones (B = - .54, SE = .037, t(1377.77) = - 14.76, p < .001). Interestingly, we also detected an interaction, such that the effect of attitude-congruence was smaller when participants considered what they *should* share vs. what they actually would share (B = -.054, SE = .024, t(2643.64) = - 2.28, p = .023).

We followed up on this interaction by testing the effect of condition on attitude-congruent and attitude-incongruent facts separately. Results revealed that while there was no significant difference between should and would share judgments for attitude-congruent facts (B = -.020, SE = .039, t(525.07) = - .59, p = .56), participants reported that they would share attitude-incongruent facts significantly less than they said they should do so (B = - .14, SE = .06, t(526.01) = - 2.31, p = .021).

Study 8 Method.

Five hundred and ninety-seven MTurkers were recruited for this study (mean age = 41.4, 44.9% female). All participants read a short vignette about a character named Daisy. Daisy is described as someone who "really cares about the issue of gun rights/gun

control," who "has strong opinions" about the topic, and who "sometimes posts interesting information about gun control to her social media accounts." In all conditions, Daisy is considering whether or not to post online an article that she had read which contains surprising and potentially relevant information about the issue. In the fake news condition, participants read that Daisy knows the information in the article to be false, but that she decides to share it anyway because "even though it's not true, it aligns with her position" and "reinforces her personal opinion." In the self-censorship condition, participants read that Daisy knows the information in the article to be true, but that she decides not to share it, because "even though it's true, it does not support her position" and "goes against her personal opinion." We also manipulated whether Daisy's specific ideology on the issue aligned with the participants' beliefs (ideological ally vs. ideological opponent). To do this, Daisy was described as either "strongly supporting" or else "strongly opposing" new gun regulations, and the condition variable was coded according to whether Daisy's randomized beliefs matched (ally condition) or did not match (opponent condition) participants' professed beliefs. Participants' own ideology was captured in a binary choice between supporting and opposing stronger regulations (67.3% supported, 32.7% opposed) which was embedded in the demographic portion of the survey. In summary, participants were randomly-assigned to one of four conditions in a 2 (fake news sharing vs. self-censorship) x 2 (ally vs. opponent) design, which varied both how Daisy behaved and whether she was ideologically aligned with the participant judging her.

To assess their judgments of Daisy's behavior, participants completed two items: "Daisy's decision about what information to post on social media is …" (*biasedness*: 1 "Not at all biased" to 7 "Deeply biased"); and "What Daisy did in this situation is…" (*moral wrongness*: 1 "Not wrong at all" to 7 "Deeply wrong"). We predicted that participants would see self-censorship as less wrong and less biased than sharing fake news, and further suspected that allies would be judged less harshly than opponents, but we did not have any predictions about how these effects might interact.

Finally, participants indicated their beliefs about gun control (as noted above), completed age and gender demographics, and indicated their broader political affiliation in a forced choice between "lean liberal" and "lean conservative" (60.1% and 39.9% respectively). Political attitudes were correlated with beliefs about gun control (r = .58).

Study 8 Results.

Attributions of Bias. We began by subjecting participants' attributions of bias to a two-way ANOVA, with Daisy's behavior (fake-news sharing vs. self-censorship) and her ideological alignment to the participant (ally vs. opponent) as manipulated factors. This procedure revealed a significant main effect of ideological alignment, such that participants judged Daisy's sharing behavior as less biased when she was on their side of the issue (an ally, M = 5.93, SD = 1.45) vs. on the opposing side (an opponent, M = 6.40, SD = 1.03; $F(1, 593) = 16.95$, $p < .001$, $\eta^2_G = .028$). In line with our predictions, it also detected a main effect whereby self-censorship was seen as less biased (M = 5.85, SD = 1.40) as compared to fake news sharing (M = 6.50, SD = 1.03, $F(1, 593) = 38.35$, $p <$

.001, η^2_G = .061). There was also an interaction such that the difference in attributions of bias between self-censorship and fake news was smaller when judging an ideological opponent than an ideological ally (F(1, 593) = 5.43, p = .023, η^2_G = .009), which may reflect that attributions of bias to ideological opponents who share fake news neared the ceiling of the scale.

Follow-up t-tests revealed that self-censorship was seen as less biased than fake news sharing both when judging an ideological ally (M_{sc} = 5.55, SD_{sc} = 1.56 vs. M_{fn} = 6.40, SD_{fn} = 1.15; t(286) = -5.15, p < .001, d = -.61) and when judging an ideological opponent (M_{sc} = 6.19, SD_{sc} = 1.12 vs. M_{fn} = 6.57, SD_{fn} = .92, t(307) = -3.32, p = .001, d = -.38). Interestingly, we note that the means all four cells fell significantly above the midpoint of the scale (all ps < .001), suggesting that our participants saw self-censorship as less biased than fake news sharing, but still tended to agree that it represents a biased way to share information.

Moral Wrongness Judgments. We next subjected participants moral wrongness judgments to a similar 2x2 ANOVA model, which revealed a main effect of ideological alignment (F(1, 593) = 14.32, p < .001, η^2_G = .024): Participants saw the sharing behavior of an ideological ally (M = 4.19, SD = 2.14) as less wrong than the sharing behavior of an ideological opponent (M = 4.91, SD = 1.88). Of interest, we also found a main effect whereby self-censorship (M = 3.37, SD = 1.89) was seen as substantially less wrong than fake-news sharing (M = 5.76, SD = 1.39; F(1, 593)= 301.50, p < .001, η^2_G = .34). This effect held separately when judging an ideological ally (M_{sc} = 3.06, SD_{sc} = 1.90 vs. M_{fn} = 5.56, SD_{fn} = 1.53, t(286) = -12.12, p < .001, d = - 1.43) as well as an ideological opponent

(M_{sc} = 3.73, SD_{sc} = 1.81 vs. M_{fn} = 5.91, SD_{fn} = 1.26, $t(307) = -12.43$, $p < .001$, $d = -1.42$); and we detected no interaction between ideological alignment and sharing behavior ($F(1, 593) = 1.36$, $p = .24$, $\eta^2_G = .0023$). Note that despite participants seeing self-censorship as biased on average, they tended to *disagree* that self-censorship is wrong, regardless of whether done by an ideological ally or an ideological opponent. Moreover, despite their tendency to judge allies more leniently than opponents, participants tended to regarded an ideological *opponent* who self-censored more favorably than an ideological *ally* who shared fake news ($t(271) = -8.98$, $p < .001$, $d = -1.09$).

Studies 7 and 8 Discussion.

The results of Studies 7-8 shed light on people's reflective judgments about the tendency to self-censor and they suggest an important point of divergence between the psychology of fake news and that of self-censorship: People don't necessarily view their tendency to self-censor as a problem. In Study 7, participants indicated that they *should* share attitude-incongruent facts less than attitude-congruent ones, even while controlling for differential perceptions of factual accuracy and relevance. In Study 8, while participants strongly agreed that sharing fake news was morally wrong, they disagreed that omitting true news was morally wrong. Still, people's judgments of self-censorship are not totally rosy either: People report they *should* share attitude-incongruent facts a bit more than they would (Study 7) and they readily admit that self-censorship is a form of biased communication (Study 8).

These results suggest an interesting tension between people's judgments of self-censorship and their commitment to norms of accuracy. Although people will often say that being accurate is important, they are sometimes convinced that other goals, especially moral ones, are more important in practice. For example, Cusimano and Lombrozo (2021) recently demonstrated that people sometimes explicitly endorse biased reasoning, particularly if they think it will lead to morally preferable conclusions. Similarly, our results suggest that people endorse, to some extent, their tendency to self-censor. Why might this be the case? One possibility is that leaving facts out is not obviously recognized as a violation of accuracy norms, or, to the extent that it is seen as a violation of accuracy, it just isn't coded as a particularly egregious one by virtue of its status as an omission. Another possibility is that although people value accuracy and see self-censorship as biased, they may worry that sharing attitude-incongruent information might have undesirable consequences for the broader debate: Sharing attitude-incongruent evidence might feel more like arming the opposition than merely getting the relevant facts on the table, and it may seem like the lesser of evils as a result.

Whatever the reason, these results raise some worry that debiasing self-censorship may prove challenging, likely requiring a broader conceptualization of people's accuracy goals and how they play out in charged exchanges of information. If people fail to see self-censorship as wrong in the same way they judge fake news as wrong, or if think they *should* omit attitude-incongruent facts they also judge to be true and relevant, then interventions which teach them the correct facts or which alert them to their tendency to share in a biased manner may prove insufficient. In other words, while sharing false info

and omitting true info may represent parallel threats to the information ecosystem, the sorts of interventions required to remedy them may be different.

General Discussion.

Across eight pre-registered experiments investigating six different moral and political issues, we document and explore people's tendency to self-censor. When faced with the opportunity to share facts which contravene valued positions on hot-button issues, people often demur, omitting from communication information they simultaneously agree to be true and relevant. Importantly, we show that self-censorship effects persist while controlling for possible effects of motivated memory and belief. That is, while people's prior position on moral issues does influence their interpretation of the evidence (see Web Appendix), this alone cannot explain why they self-censor. Regardless of whether we a) focused analysis on participants who reported the correct facts *ex post*, b) used information that was hard to reason away as untrue, or c) directly measured and controlled for perceived accuracy and relevance, we found that attitude-incongruence bred omissions and reduced willingness to share. What can explain this effect? Consistent with our prediction that self-censorship is driven by impression management concerns, we demonstrate that the tendency to self-censor is moderated by the presence of observers and mediated by expectations about social judgment. Finally, comparing judgments of whether they would vs. should share incongruent facts, we

found that participants endorse their tendency to self-censor to some extent, but also that their inclination to self-censor outstrips what they think is appropriate upon reflection.

Implications of self-censorship

In recent years, scholars and practitioners alike have raised serious concerns about the quality of evidence that consumers share within social networks. In particular, much has been made of the problem of "fake news" – the circulation of false or misleading information via word-of-mouth about consequential social issues like election integrity or public health (Lazer et al. 2018). Indeed, the proliferation of fake news distorts the accuracy of beliefs (Pennycook and Rand 2021), decreases trust in science and media (Allcott and Gentzkow, 2017; Scheufel and Krause 2019), fuels conspiracy theories (Albarracín 2020), and contributes to political animosity and extremism (Finkel et al. 2020). Such dynamics have critical implications for marketing and consumer behavior, both at the individual-level, where moral and political beliefs can impact consumption decisions (Lutrell, Teeny, and Petty 2021) – e.g., about what to eat (Fernbach et al. 2019) or watch (Stroud 2010) – and at the societal-level, where broader dynamics of political cynicism and mistrust can lead to market instability.

Our experiments suggest that the sharing of fake news may not be the only ailment plaguing the information ecosystem: Self-censorship is a parallel threat. But whereas the danger of fake news is an abundance of false information, the danger of self-censorship is a dearth of useful, true information. From our vantage point, improving the quality of information shared across social networks requires not only addressing why

people share falsehoods, but also investigating why they might avoid sharing relevant facts. Importantly, while partisan motivations may underlie both problems, the sorts of interventions required to remedy them may be different. Consider recent evidence that prompting sharers to consider accuracy decreases their tendency to share disinformation online (Pennycook et al. 2021). The logic here is that people (a) think it is wrong to share falsehoods in principle, but (b) often forget to check for factual accuracy in practice, such that a simple reminder to do so can serve as a potential antidote. By contrast, many of our experiments find that even after explicitly affirming a given fact's accuracy and relevance, participants nonetheless self-censor. Why might this be the case? We suspect that part of the answer relates to the psychology of omissions and commissions. Prior research on the 'omission bias' finds that people object more strenuously to harmful *actions* than to equally harmful *inactions* (Spranca, Minsk, and Baron 1991). In our contexts, while sharing fake news may resemble a harmful commission, self-censoring may be perceived as a benign omission. That is, sharers may less readily recognize self-censorship as a violation of accuracy, and even if they do, they may see it as a less objectionable one than actively sharing false information, as borne out in our Study 8 (see also, Levine et al. 2018). Further complicating matters is the likelihood that self-censorship may be even harder to detect and correct from the receiver side. While observers can be taught to recognize the hallmarks of disinformation (Bago, Rand, and Pennycook 2020), they cannot be taught to detect facts that go unshared in the first place. In these respects, self-censorship might be even more pernicious than the fake news problem.

Beyond their immediate practical implications for information quality, processes of self-censorship can also help to explain a variety of group-level phenomena previously documented. For one example, consider the case of 'echo chambers.' With the help of network-level analyses, scholars have recently documented that individuals on opposite sides of moral issues exist in increasingly separate, non-overlapping communities both online and in person (Cinelli et al. 2021; Brown and Enos 2021). And while partisan segregation is thought to breed biased beliefs and moral disagreement, less is known about the specific individual-level processes that might give rise to these troubling dynamics. We theorize that in addition to seeking attitude-congruent news (Stroud 2010), and affiliating with like-minded others (Mosleh et al. 2021), people also *actively withhold* attitude-incongruent facts from communication, filtering information that might favor more moderate or nuanced positions. Relatedly, self-censorship could provide an additional explanation for group polarization, the classic finding that groups of like-minded individuals become more extreme in their opinions after discussing issues together. Typical accounts of group polarization hold that discussions with ingroups (a) expose participants to new attitude-confirming arguments and (b) nudge them to align their beliefs with group consensus positions (Myers and Lamm 1976). But self-censorship likely plays an important and complementary role in such contexts. Discussants who possess important attitude-disconfirming facts may be unlikely to bring them up for fear of social reproach, and their failure to do so may further amplify group polarization effects.

In summary, self-censorship effects have implications for both theory and practice. Theoretically, self-censorship represents a potential explanation for a number of social phenomena that give rise to biased beliefs and extreme attitudes. Practically, self-censorship represents a possible threat to information ecosystems, one which may be harder to detect than explicit disinformation and whose ultimate implications are not yet well-understood.

Limitations

We note a few important limitations. First, it should be noted that our data come from lab and online experiments, in which participants summarized news articles or rated their willingness to share specific facts. Such measures are common in the marketing literature (e.g., summarization tasks; Melumad, Meyer, and Kim 2021; sharing intentions; Chen and Berger 2016) and their use in controlled paradigms allows us to cleanly separate self-censorship from motivated reasoning. In more naturalistic contexts, we suspect that impression management motives and motivated reasoning processes might combine to produce stronger patterns of biased information transmission than those found here. It is also possible that, in more spontaneous and realistic circumstances, the imperative to manage impressions may be even more potent than that felt by our participants (Van Boven et al. 2012), further amplifying the pressure to self-censor.

Another important limitation of this research concerns our focus on communication within *ingroups* specifically. We focused on ingroups for two reasons. First, ingroup communications more often resemble reality: People preferentially

associate with others who share their political attitudes (Brown and Enos 2021), follow them on social media (Mosleh et al. 2021), and rely on them for advice and information (Johnson, Rodrigues, and Tuckett 2021). Second, people likely care more what other ingroup (vs. outgroup) members think of them, suggesting that our mechanism of interest, impression management concerns, are likely stronger in ingroup contexts. Still, a natural further question concerns how sharing might be pressured or distorted in communications with moral *outgroups*. Although our studies cannot speak to this issue directly, it is possible that sharers might feel similarly motivated to omit attitude-incongruent evidence when interacting with outgroups, but for a different reason. Whereas they may worry about appearing disloyal to ingroup members, people may not want to provide outgroup members with ammunition for their (opposing) arguments.

Future Directions

Our paradigms open a number of avenues for future research. As a starting point, we hope future work will seek to replicate our effects and examine generality and boundary conditions across different sorts of sharers, social relationships, and conversations. On this front, we note a few possibilities. First, it would be helpful to know whether particular individuals experience stronger or weaker pressure to self-censor. A greater willingness to bring up inconvenient facts might be associated with traits like actively open-minded thinking (AOT; Baron 2019) or need for cognition (NFC; Cacioppo and Petty 1982), while a greater tendency to self-censor might be associated with tendencies towards self-monitoring (SM; Snyder 1974), fear of negative evaluation (FNE; Leary 1983) or neuroticism (N; Widiger 2009). Second, future work can examine

the prevalence of self-censorship across different relationships and audiences. Drawing on previous studies of 'broadcasting and narrowcasting' (Barasch and Berger 2014), we suspect that people may self-censor more in communication with larger or less familiar audiences, but they may do so less among smaller groups of trusted others. If true, this tendency may prove perverse, as it would mean that communications reaching larger audiences are more biased. Third, it will be helpful to establish self-censorship in other conversational settings beyond those we tested. For instance, there are bound to be cases in which bringing up an incongruent fact entails not just volunteering useful information, but actually *correcting* factual inaccuracies communicated by others. Although future work is needed to generalize our effects to such settings, our account would predict that people are even more inclined to self-censor when clarifying the facts might seem combative or antagonistic.

A second set of interesting questions concerns whether other social forces – beyond the need to manage impressions – might shape strategic omissions of attitude-incongruent evidence. In Study 5, for instance, participants were less likely to share attitude-incongruent evidence with others than with their future selves. Although our other results suggest that impression management concerns likely play an important role, this pattern may also reflect a desire to shield others from information that might weaken *their* loyalty to valued causes – a kind of *persuasion* management. In other words, when exchanging information, people may be concerned not only with managing their reputations, but also with managing *others'* beliefs and attitudes. By blocking access to attitude-incongruent evidence, people may seek to ensure that others remain on their side

of an issue. Although not the primary focus of this work, persuasion management is an intriguing hypothesis for future work in this area. Importantly, we note that it cannot explain what we found in Study 6: That people's tendency to self-censor correlates strongly with their predictions about social judgment.

Finally, future work can investigate strategies for debiasing self-censorship. Here we note two possible approaches. The first involves increasing the value people place on accuracy and helping them to recognize the costs of censorship[11]. As noted above, we are pessimistic about the efficacy of light-touch accuracy nudges for reducing self-censorship, as people do not readily see strategic omissions as violations in need of correction. On the other hand, as we learn more about the costs of self-censorship for discourse, it may be possible to teach people about the perils of withholding useful information in ways that might shift behavior. Anecdotally, prior authors have published persuasive essays decrying the perils of censorship and idea suppression (e.g., Loury 1994, Weiss 2021), and it remains to be seen whether exposure to such arguments might make people more critical of self-censorship or more forthright in sharing unpopular facts.

A second approach to debiasing might try to address people's concerns about social sanction. It may be that sharers' worries about the costs of sharing are overblown, particularly those associated with the mere inclusion of an incongruent, but clearly relevant, fact in an otherwise attitude-confirming conversation. In line with this idea,

[11] It should be noted that merely teaching people the correct facts will likely prove insufficient, as self-censorship occurs even when people attest directly to the accuracy and relevance of the evidence in question. Rather, the critical task is to convince people to share accurately, above and beyond their beliefs about what is true.

recent research suggests that while they dislike and distrust outright political opponents, people sometimes appreciate political allies who seek, rather than avoid, attitude-incongruent perspectives (Heltzel and Laurin 2021). Such results raise questions for future work. Are sharer's expectations accurate? And, if not, can helping them become better calibrated reduce the pressure to self-censor? More broadly, we suspect that interventions which increase trust between communicators or decrease reputational concerns are promising avenues for future investigation. Such approaches might find it useful to focus on the receiver side of communication: Receivers can always encourage sharers to be more forthright with the facts, regardless of which side of the argument they support, and doing so should help put sharer's at ease to be honest and accurate.

Finally, we note that while self-censorship erodes the accuracy and completeness of important conversations, there may be other competing reasons that sometimes render it defensible. From the sharer's perspective, it may be rational to avoid risking their moral reputation. And from a broader social good perspective, it could be argued that in certain cases, sharing relevant facts might do more to upset, confuse, or offend than to inform and enlighten[12]. Not every fact needs to be shared in every communication. Still, while accuracy and completeness may not be the only factors worth maximizing in communication, a greater willingness to open-mindedly surface facts about complex issues should, all else equal, improve discourse and decision-making.

Conclusion

[12] Imagine, for example, bringing up the base rates of survival to a friend whose loved one was recently diagnosed with a deadly disease (e.g., Cusimano and Lombrozo 2021; Levine 2021).

Processes of information consumption and sharing form the basis of people's moral and political beliefs. These, in turn, serve as important inputs to a wide variety of consequential decisions, from choosing whether to get vaccinated against deadly diseases, to considering whether to support or boycott major brands, to deciding whether to go green or consume sustainably. In an ideal world, such processes would surface and transmit the most relevant evidence – regardless of which side of the argument it supports. Against this normative benchmark, the present research identifies a departure. People treat evidence-sharing as a signal of their moral commitments, and they will self-censor as a result, selectively omitting attitude-incongruent facts from discussion to protect their reputation and affirming this behavior as acceptable.

References.

Allcott, Hunt, and Matthew Gentzkow, (2017), "Social media and fake news in the 2016 election." *Journal of Economic Perspectives*, 31 (2), 211-36.

Albarracín, Dolores (2020), "Conspiracy Beliefs: Knowledge, ego defense, and social integration in the processing of fake news," *The Psychology of Fake News*, Routledge, 196-219.

Bagwell, Laurie Simon, and B. Douglas Bernheim (1996), "Veblen effects in a theory of conspicuous consumption." *The American Economic Review*, 349-373.

Barasch, Alixandra, and Jonah Berger (2014), "Broadcasting and Narrowcasting: How Audience Size Affects What People Share," *Journal of Marketing Research*, 51 (3), 286–99.

Barasch, Alixandra, Gal Zauberman, and Kristin Diehl (2018), "How the intention to share can undermine enjoyment: Photo-taking goals and evaluation of experiences," *Journal of Consumer Research*, 44 (6), 1220-1237.

Baron, Jonathan (2019), "Actively open-minded thinking in politics," *Cognition*, 188, 8-18.

Bago, Bence, David G. Rand, and Gordon Pennycook (2020), "Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines," *Journal of Experimental Psychology: General*.

Baumeister, Roy F. and Mark R. Leary (1995), "The need to belong: desire for interpersonal attachments as a fundamental human motivation," *Psychological Bulletin*, 117 (3), 497-529.

Bellezza, Silvia, Neeru Paharia, and Anat Keinan (2017), "Conspicuous consumption of time: When busyness and lack of leisure time become a status symbol," *Journal of Consumer Research*, 44 (1), 118-138.

Berger, Jonah (2014), "Word of mouth and interpersonal communication: A review and directions for future research," *Journal of Consumer Psychology*, 24 (4), 586-607.

Berger, Jonah, and Chip Heath (2007), "Where consumers diverge from others: Identity signaling and product domains," *Journal of Consumer Research*, 34 (2), 121-134.

Berger, Jonah, and Katherine L. Milkman (2012), "What makes online content viral?" *Journal of Marketing Research*, 49 (2), 192-205.

Berman, Jonathan Z., Emma E. Levine, Alixandra Barasch, and Deborah A. Small (2015), "The Braggart's dilemma: On the social rewards and penalties of advertising prosocial behavior," *Journal of Marketing Research*, 52 (1), 90-104.

Brown, Jacob R., and Ryan D. Enos (2021), "The measurement of partisan sorting for 180 million voters," *Nature Human Behaviour*, 1-11.

Cacioppo, John T., and Richard E. Petty (1982), "The need for cognition," *Journal of Personality and Social Psychology*, 42 (1), 116.

Cheema, Amar, and Andrew M. Kaikati (2010), "The effect of need for uniqueness on word of mouth," *Journal of Marketing Research,* 47 (3), 553-563.

Chen, Zoey, and Jonah Berger (2013), "When, why, and how controversy causes conversation," *Journal of Consumer Research,* 40 (3), 580-593.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini (2021), "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, 118 (9).

Chen, Zoey, and Jonah Berger (2016), "How content acquisition method affects word of mouth," *Journal of Consumer Research*, 43 (1), 86-102.

Crockett, Molly J. (2017), "Moral outrage in the digital age," *Nature Human Behavior*, 1 (11), 769-771.

Crockett, David, and Melanie Wallendorf (2004), "The role of normative political ideology in consumer behavior," *Journal of Consumer Research*, 31 (3), 511-528.

Cusimano, Corey, and Tania Lombrozo (2021), "Morality justifies motivated reasoning in the folk ethics of belief," *Cognition*.

Dalton, Amy N., and Li Huang (2014), "Motivated forgetting in response to social identity threat." *Journal of Consumer Research*, 40 (6), 1017-1038.

De Angelis, Matteo, Bonezzi, Andrea, Peluso, Alessandro, Rucker, Derek, Costabile, Michele (2012), "On Braggarts and Gossips: A Self-Enhancement Account of Word-of-Mouth Generation and Transmission," *Journal of Marketing Research*, 49 (4), 551–63.

Ekstrom, Pierce D., and Calvin K. Lai (2021), "The Selective Communication of Political Information," *Social Psychological and Personality Science*.

Fernbach, Philip M., Nicholas Light, Sydney E. Scott, Yoel Inbar, and Paul Rozin (2019), "Extreme opponents of genetically modified foods know the least but think they know the most," *Nature Human Behavior,* 3 (3), 251-256.

Finkel, Eli J., et al. (2020), "Political sectarianism in America," *Science*, 370 (6516), 533-536.

Fridman, Ariel, Rachel Gershon, and Ayelet Gneezy (2021), "COVID-19 and vaccine hesitancy: A longitudinal study," *PloS one*, 16 (4), e0250123.

Gollwitzer, Anton, Cameron Martel, William J. Brady, Philip Pärnamets, Isaac G. Freedman, Eric D. Knowles, and Jay J. Van Bavel (2020), "Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic," *Nature Human Behavior*, 4 (11), 1186-1197.

Golman, Russell, David Hagmann, and George Loewenstein (2017), "Information avoidance," *Journal of Economic Literature*, 55 (1), 96-135.

Good, Irving John (1967), "On the principle of total evidence," *The British Journal for the Philosophy of Science*, 17 (4), 319-321.

Grice, Herbert P. (1975), "Logic and conversation," *Speech Acts*, Brill, 41-58.

Heltzel, Gordon, and Kristin Laurin (2021), "Seek and ye shall be fine: Attitudes toward political perspective-seekers," pre-print.

Han, Young Jee, Joseph C. Nunes, and Xavier Drèze. "Signaling status with luxury goods: The role of brand prominence," *Journal of Marketing*, 74 (4), 15-30.

Janis, Irving L. (1972), "Victims of Groupthink: A psychological study of foreign-policy decisions and fiascoes."

Janus, Alexander L. (2010), "The influence of social desirability pressures on expressed immigration attitudes," *Social Science Quarterly*, 91 (4), 928-946.

John, Leslie K., Kate Barasz, and Michael I. Norton (2016), "Hiding personal information reveals the worst," *Proceedings of the National Academy of Sciences,* 113 (4), 954-959.

John, Leslie K., Hayley Blunden, and Heidi Liu (2019), "Shooting the messenger," *Journal of Experimental Psychology: General*, 148 (4), 644.

Johnson, Samuel GB, Max Rodrigues, and David Tuckett (2021), "Moral tribalism and its discontents: How intuitive theories of ethics shape consumers' deference to experts," *Journal of Behavioral Decision Making*, 34 (1), 47-65.

Jost, John T (2017), "The marketplace of ideology: Elective affinities in political psychology and their implications for consumer behavior," *Journal of Consumer Psychology*, 27(4), 502-520.

Jost, John T., Melanie Langer, and Vishal Singh (2017), "The politics of buying, boycotting, complaining, and disputing: An extension of the research program by Jung, Garbarino, Briley, and Wynhausen," *Journal of Consumer Research*, 44 (3), 503-510.

Kahan, Dan M. (2017), "Misconceptions, misinformation, and the logic of identity-protective cognition."

Kidwell, Blair, Adam Farmer, and David M. Hardesty (2013), "Getting liberals and conservatives to go green: Political ideology and congruent appeals," *Journal of Consumer Research*, 40 (2), 350-367.

Koper, Christopher S., Daniel J. Woods, and Jeffrey A. Roth, (2004), "An updated assessment of the federal assault weapons ban: impacts on gun markets and gun violence, 1994-2003," *National Institute of Justice, US Department of Justice*.

Kunda, Ziva (1990), "The case for motivated reasoning," *Psychological Bulletin,* 108 (3), 480.

Lazer, David M. J. et al. (2018), "The science of fake news," *Science,* 359 (6380), 1094-1096.

Long, Elisa F., M. Keith Chen, and Ryne Rohla (2020), "Political storms: Emergent partisan skepticism of hurricane risks," *Science Advances*, 6 (37), 7906.

Leary, Mark R. (1983), "A brief version of the Fear of Negative Evaluation Scale," *Personality and Social Psychology Bulletin*, 9 (3), 371-375.

Leary, Mark R. and Robin M. Kowalski (1990), "Impression management: A literature review and two-component model," *Psychological Bulletin*, 107 (1), 34.

Lerner, Jennifer S., and Philip E. Tetlock (1999), "Accounting for the effects of accountability," *Psychological Bulletin*, 125 (2), 255.

Levine, Emma E., and Taya R. Cohen (2018), "You can handle the truth: Mispredicting the consequences of honest communication," *Journal of Experimental Psychology: General*, 147 (9), 1400.

Levine, Emma E., et al. (2018), "The surprising costs of silence: Asymmetric preferences for prosocial lies of commission and omission," *Journal of Personality and Social Psychology*, 114 (1), 29.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Loury, Glenn C. (1994), "Self-censorship in public discourse: A theory of "political correctness" and related phenomena," *Rationality and Society*, 6 (4), 428-461.

Lovett, Mitchell J., Renana Peres, and Ron Shachar (2013), "On brands and word of mouth," *Journal of Marketing Research*, 50 (4), 427-444.

Luttrell, Andrew, Jacob D. Teeny, and Richard E. Petty (2021), "Morality Matters in the Marketplace: The Role of Moral Metacognition in Consumer Purchasing," *Social Cognition*, 39 (3), 328-351.

Mackie, Diane, and Joel Cooper (1984), "Attitude polarization: Effects of group membership," *Journal of Personality and Social Psychology*, 46 (3), 575.

Marie, Antoine, Sacha Altay, and Brent Strickland (2021), "Moral Conviction Predicts Robust Sharing Preference for Politically Congruent Headlines," preprint.

Melumad, Shiri, Meyer, Robert, & Kim, Yoon D. (2021), "The Dynamics of Distortion: How Successive Summarization Alters the Retelling of News," *Journal of Marketing Research*.

Mesmer-Magnus, Jessica R. and Leslie A. DeChurch (2009), "Information sharing and team performance: A meta-analysis," *Journal of Applied Psychology*, 94 (2), 535.

Miller, Dale T., and Cathy McFarland (1987), "Pluralistic ignorance: When similarity is interpreted as dissimilarity," *Journal of Personality and Social Psychology*, 53 (2), 298.

Mojzisch, Andreas, et al. (2014), "The consistency principle in interpersonal communication: Consequences of preference confirmation and disconfirmation in

collective decision making," *Journal of Personality and Social Psychology*, 106 (6), 961.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G. Ran (2021), "Shared partisanship dramatically increases social tie formation in a Twitter field experiment," *Proceedings of the National Academy of Sciences*, 118 (7).

Mosleh, Mohsen, Gordon Pennycook, and David G. Rand (2020), "Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter," *Plos One*, 15 (2), e0228882.

Myers, David G., and Helmut Lamm (1976), "The group polarization phenomenon," *Psychological Bulletin*, 83 (4), 602.

Newport, Frank (2019), "Analyzing Surveys on Banning Assault Weapons," *Polling Matters,* Gallup.

Nickerson, Raymond S. (1998), "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of General Psychology*, 2 (2), 175-220.

Norton, Michael I., et al. (2006), "Color blindness and interracial interaction: Playing the political correctness game," *Psychological Science*, 17 (11), 949-953.

Nyhan, Brendan, and Jason Reifler, (2010), "When corrections fail: The persistence of political misperceptions," *Political Behavior*, 32 (2), 303-330.

Pennycook, Gordon et al. (2021), "Shifting attention to accuracy can reduce misinformation online," *Nature*, 592 (7855), 590-595.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand (2020), "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge interventionm" *Psychological Science*, 31 (7), 770-780.

Pew Research (2016), "The Political Environment On Social Media," PewResearch.org.

Pennycook, Gordon, and David G. Rand (2021), "The psychology of fake news," *Trends in Cognitive Sciences.*

Prentice, Deborah A., and Dale T. Miller (1993), "Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm," *Journal of Personality and Social Psychology*, 64 (2), 243.

Roberts, Annabelle R., Emma E. Levine, and Ovul Sezer (2020), "Hiding success." *Journal of Personality and Social Psychology*.

Rogowski, Jon C. and Joseph L. Sutherland (2016), "How ideology fuels affective polarization," *Political Behavior*, 38 (2), 485-508.

Rotella, Katie N., and Jennifer A. Richeson (2013), "Motivated to “forget” the effects of in-group wrongdoing on memory and collective guilt," *Social Psychological and Personality Science*, 4 (6), 730-737.

Savitsky, Kenneth, Nicholas Epley, and Thomas Gilovich (2001), "Do others judge us as harshly as we think? Overestimating the impact of our failures, shortcomings, and mishaps," *Journal of Personality and Social Psychology*, 81 (1), 44.

Scheufele, Dietram A., and Nicole M. Krause (2019), "Science audiences, misinformation, and fake news," *Proceedings of the National Academy of Sciences*, 116 (16), 7662-7669.

Shavitt, Sharon (2017), "Political ideology drives consumer psychology: Introduction to research dialogue," *Journal of Consumer Psychology*, 27 (4), 500-501.

Silver, Ike, George Newman, and Deborah A. Small (2021), "Inauthenticity aversion: Moral reactance toward tainted actors, actions, and objects," *Consumer Psychology Review*, 4 (1), 70-82.

Snyder, Mark (1974), "Self-monitoring of expressive behavior," *Journal of Personality and Social Psychology* 30 (4) 526.

Spranca, Mark, Elisa Minsk, and Jonathan Baron (1991), "Omission and commission in judgment and choice," *Journal of Experimental Social Psychology*, 27 (1), 76-105.

Stroud, Natalie J. (2010), "Polarization and partisan selective exposure," *Journal of Communication*, 60 (3), 556-576.

Tajfel, Henri, John C. Turner, William G. Austin, and Stephen Worchel (1979), "An integrative theory of intergroup conflict," *Organizational Identity: A reader*, 56 (65).

Tappin, Ben M., Gordon Pennycook, and David G. Rand (2020), "Bayesian or biased? Analytic thinking and political belief updating," *Cognition,* 204, 104375.

Tetlock, Philip E., Orie V. Kristel, S. Beth Elson, Melanie C. Green, and Jennifer S. Lerner, (2000), "The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals," *Journal of Personality and Social Psychology*, 78 (5), 853.

Tetlock, Philip E., Linda Skitka, and Richard Boettger (1989), "Social and cognitive strategies for coping with accountability: conformity, complexity, and bolstering," *Journal of Personality and Social Psychology*, 57 (4), 632.

Toubia, Olivier and Andrew T. Stephen (2013), “Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter?” *Marketing Science*, 32 (3), 368–92.

Van Boven, Leaf, George Loewenstein, Edward Welch, and David Dunning (2012), "The illusion of courage in self-predictions: Mispredicting one's own behavior in embarrassing situations," *Journal of Behavioral Decision Making*, 25 (1), 1-12.

Washburn, Anthony N., and Linda J. Skitka (2018), "Science denial across the political divide," *Social Psychological and Personality Science*, 9 (8), 972-980.

Weiss, Bari (2021), "The Self-Silencing Majority," *DeseretNews.*

Widiger, Thomas A. (2009), "Neuroticism," In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior*, 129–146. The Guilford Press.

Williams, Daniel (2020), "Socially adaptive belief," *Mind & Language*.

Woolley, Kaitlin, and Jane L. Rise (2021), "Hiding from the Truth: When and How Cover Enables Information Avoidance," *Journal of Consumer Research*, 47 (5) 675-697.

Zwebner, Yonat, and Rom Y. Schrift (2020), "On my own: the aversion to being observed during the preference-construction stage," *Journal of Consumer Research*, 47 (4), 475-499.

CHAPTER 3: WHEN AND WHY "STAYING OUT OF IT" BACKFIRES IN MORAL AND POLITICAL DISAGREEMENTS

Ike Silver & Alex Shaw

Abstract.

People care where others around them stand on contentious moral and political issues. Yet when faced with the prospect of taking sides and the possibility of alienating observers with whom they might disagree, actors often try to "stay out of it" – communicating that they would rather not to take a side at all. We demonstrate that despite its intuitive appeal for reducing conflict, opting not to take sides over moralized issues can harm trust, even relative to siding against an observer's viewpoint outright. Across eleven experiments (N = 4,383) using controlled scenarios, real press video clips, and incentivized economic games, we find that attempts to stay out of the fray are often interpreted as deceptive and untrustworthy. When actors choose not to take sides, observers often ascribe concealed opposition, an attribution of strategic deception which provokes distrust and undermines real-stakes cooperation and partner choice. We further demonstrate that this effect arises only when staying out of it seems strategic: Actors who seem to hold genuine middle-ground beliefs or who lack incentives for impression management are not distrusted for avoiding conflict. People are often asked to take sides in moral and political disagreement. Our findings outline a reputational risk awaiting those who opt not to do so.

Introduction.

People dislike and distrust those who oppose them over contentious political issues like immigration policy, gun control, COVID-19 safety, or abortion rights (Finkel et al., 2020; Rogowski & Sutherland, 2016). Indeed, both liberals and conservatives see their own moral and political beliefs as objectively superior to alternative perspectives, avoid exposure to opposing viewpoints, and often construe disagreement over hot-button issues as a threat to their core values (Toner, Leary, Asher, & Jongman-Sereno, 2013; Frimer, Skikta, & Motyl, 2017; Chambers & Melnyk, 2006).[13] More broadly, across a variety of interpersonal settings, those who find themselves on "the wrong side" of moralized issues often encounter outrage, intolerance, and prejudice (Brandt, et al., 2014; Chambers, Schlenker, & Collisson, 2013; Ditto et al., 2019). It is no surprise then, given the costs of moral disagreement, that many individuals feel discomfort at the prospect of sharing their views about hot-button issues in personal and professional settings (Pew Research Center, 2019).

One intuitive way to avoid hostility when asked to weigh in on such issues is simply not to take sides. Is expressing, for example, that one would prefer to "stay out of it" an effective interpersonal strategy? Although much is known about how observers respond to those who side with or against them in moral and political conflict, we know little about how people evaluate those who indicate that they would rather not take sides

[13] Such dynamics have powerful behavioral consequences: To name just a few, they impact where people work (Gantz & Murray, 1980) and what they buy, (Jost, 2017), as well as whom they choose to collaborate with (Marks et al., 2019), talk to (Chen & Rohla, 2018), or vote for (Parker & Isbell, 2010).

at all. In this paper, we propose and find that opting not to weigh in on contentious issues can backfire, often drawing stronger reproach than opposing an observer's ideological position outright.

At first blush, opting not to take sides seems sensible enough, particularly when staking out a clear position on a contentious issue might anger audiences who hold the opposing view. In such settings, a desire not to weigh in might reasonably seem intellectually humble or interpersonally courteous.[14] At a minimum, choosing not to take sides (vs. disagreeing with one's audience outright) may seem less likely to provoke moral outrage or alienate one from relevant social groups (Byrne, 1971; Crockett, 2017). As noted, outright opposition can be costly. In everyday settings, disputes over morality and politics can end friendships, alienate coworkers, and strain family ties. And for prominent individuals in the public eye, contradicting observers' moral beliefs can provoke reproach in the form of protests, boycotts, or public ridicule. To sidestep these costs, it may seem sensible to stay out of the fray altogether, avoiding a position offensive to either side by saying things like "I'd rather not get into that right now"; or "I'm neutral on this issue"; or "Let's not talk politics over dinner."

By contrast, we submit that, across a variety of situations, opting not to take sides can backfire. Why? In the context of charged discussions, observers may readily interpret such stances through the lens of social incentives. In general, people are attuned to the risks and rewards others face in social contexts and will account for such incentives when

[14] Even on Gricean conversational frameworks, which often prescribe answering questions with adequate information, demurring when one's answer might cause discomfort or interpersonal conflict is seen as normative, polite, and considerate (Grice, 1975; Yoon et al., 2020).

interpreting behavior (Schlenker, 1980; Jones & Pittman, 1982; Sperber et al., 2010). For example, observers tend to discount bragging and self-promotion because they recognize these as deceptive impression management tactics intended to cultivate a positive reputation (Wosinka, Dabul, Whetstone-Dion, & Cialdini, 1996; Berman, Levine, Barasch, & Small, 2015). Similarly, an actor who opts not to take sides when asked to respond to a hot-button issue may seem to be doing something strategic. More specifically, we propose that observers will draw on contextual information about an actor's intended audience to interpret relevant incentives, infer underlying beliefs, evaluate trustworthiness, and make decisions about cooperation and support.

Our contexts of interest are any in which an actor is asked to weigh in on a contentious moral or political issue in front of an audience with a known position. Such situations arise frequently. Consider, for example, a workplace conversation in which a new member of an office is asked for her opinion about gun control. Or imagine a prominent businesswoman asked in a TV interview for her perspective on COVID-19 policy; or a politician called upon to indicate whether or not a recent election was fraudulently decided. Across these cases and many others like them, actors are put on the spot to provide an opinion: They can take a position aligned with or opposed to their audience's, or they can choose not to take a position at all. Importantly, when actors address divisive issues, observers likely have intuitions, if not common knowledge, about what the actor's audience believes (e.g., the ideological lean of an office, a friend group, or a news station's viewers), especially given that people often sort their personal and professional networks along partisan lines (e.g., Alford et al., 2011; Bonica, Rosenthal,

Blackwood, & Rothman, 2020). Here, we explore how observers use audience information in such contexts to interpret and evaluate those who try to stay above the fray.

First, we predict that choosing not to take sides will be interpreted as tacit opposition when addressing an audience with a known position. Take one of our examples from above. Imagine a new employee who says she "doesn't want express an opinion" about gun control in an office where most are politically liberal and support stricter gun laws. On our account, her co-workers will likely infer that their new colleague may be concealing underlying conservative, pro-gun convictions. After all, if she *agreed* with the opinion held by most of her coworkers, saying so would be both truthful and aligned with social incentives. We posit that "not wanting to express an opinion" may instead signal that an actor's underlying beliefs actually *oppose* those of her audience. Thus, when addressing a predominantly liberal audience, such an assertion will signal more conservative underlying beliefs: When addressing a conservative audience, saying the same thing might signal more liberal underlying beliefs.

Second, we predict that if an actor seems to harbor private convictions that differ from what they say publicly (e.g., 'I really can't get involved'), trying to stay out of it may seem like a strategic concealment tactic, and therefore elicit distrust. Converging evidence suggests that when actors send "false-signals" about their underlying beliefs or qualities, observers respond negatively (Jordan, Sommers, Bloom, & Rand, 2017; Silver, Newman, & Small, 2020), often with distrust in particular. For example, individuals who choose not to disclose unsavory past behavior (like illegal drug use) on personal

questionnaires are less attractive dates and employees because intentionally hiding past behavior seems untrustworthy (John, Barasz, & Norton, 2016). Similarly, when audience information is available as context, opting not to weigh in may seem like strategically concealing one's oppositional views. Indeed, even for powerfully divisive issues, opting not to get involved may elicit stronger distrust than opposing an observer's viewpoint outright.

Unlike more duplicitous ways of avoiding tough questions, such as 'dodging' (answering unasked but related questions; Rogers & Norton, 2011), 'paltering' (answering with truthful but misleading statements; Rogers et al., 2017), or 'deflecting' (answering with further questions; Bitterly & Schweitzer, 2019), the strategy we investigate entails making a clear choice not to get involved. In this way, trying to stay out of it may be understood – at least by observers – as a form of active non-disclosure (Baum & Critcher, 2019), and distinguished both from covert attempts to change the subject and from efforts to avoid situations in which contentious topics might be discussed in the first place. Thus, unlike previous work, we do not focus on whether observers can 'detect' an actor's desire not to take sides, but on how they will attribute underlying beliefs and intentions to actors who openly choose to "stay out of it" and on how these belief inferences impact downstream judgments and decisions.

Importantly, our account holds that people interpret attempts to stay out of it through the lens of social incentives, such that opting not to take sides should be judged more negatively to the extent that it seems more strategically motivated. If, for example, an observer believes that an actor's choice not to take sides reflects genuinely moderate

underlying beliefs (i.e., no underlying commitment one way or the other) or can attribute it to non-strategic motives, we do not expect it to backfire. In this way, we do not predict a broad aversion to moderate beliefs or to non-responsiveness per se, but a more specific cynicism about the strategic motives of those who try to skirt contentious topics. This feature distinguishes our investigation from other accounts which might predict more general negative feelings toward moral apathy or uncooperativeness (e.g., Zlatev, 2019, Grice, 1975). Our effects depend on specific patterns of belief inference and motive attribution drawn from common and predictable social contexts.

Present Experiments

Eleven experiments conducted with online and student samples examined how observers respond to actors who try to stay out of moral issues. These employed several complementary methods, including scenario studies (Experiments 1a-b, 3, 5, 6a-b), judgments of real press video clips (Experiments 2a-b), and incentive compatible economic games (Experiments 4a-b). Across our stimuli, we presented participants with a variety of examples of real and hypothetical actors (e.g., other study participants, friends, family members, professors, celebrities, politicians) opting not to take sides over a variety of divisive political issues (e.g., abortion, immigration, gun rights, COVID-19 safety). We then gauged their reactions in terms of belief inferences, attitudinal trust, intended voting support, partner-choice, and real-stakes cooperation. Importantly, we also conducted a follow-up study which confirms that actors put on the spot to take sides (a) see staying out of it as a strongly appealing option and (b) spontaneously generate expressions much like those we test in our main experiments, suggesting that speakers

may not fully grasp the costs that come with choosing not to take sides. In summary, we find robust backfire effects in response to common tactics people use to avoid taking sides, and we also document theoretically informed boundaries. In the General Discussion, we consider more effective ways to stay out of moral and political disagreements.

Before reporting these experiments, we highlight two final points. First, it should be noted that we investigate participants' responses to what people say, which is readily observed, rather than to what they believe, which must be inferred. That is, although an actor's choice to stay out of it may be consistent with a variety of underlying goals and dispositions (e.g., a concealed opinion, a moderate or mixed opinion, no opinion), observers typically rely on the charged social contexts in which such communication is embedded when making inferences. We contend that, whatever one's true motives, opting not to weigh in often *resembles* strategic concealment. Second, our experiments pay special attention to the conservative comparison between staying out of it and opposing an observer's view outright, but we note that avoiding a position also entails an additional opportunity cost: that of failing to side *with* either side. Thus, an actor who avoids taking sides may fail both to placate those who would oppose their underlying position and to woo those who would support it. We document this latter (and somewhat more obvious) cost in Experiment 3.

Experiments 1b and 4a-6b, and our follow-up study were all pre-registered at AsPredicted.Org. We report all manipulations and measures. All sample sizes were

determined in advance, and any reported exclusions were pre-registered[15]. Study procedures were approved by the IRB at the University of Chicago. All preregistrations, materials, and data can be accessed at: https://researchbox.org/118&PEER_REVIEW_passcode=OJNSOY

Experiments 1a-1b. Attributing beliefs to those who choose not to take sides

Experiments 1a and 1b tested the prediction that opting not to take sides in front of a primarily conservative audience would signal more liberal beliefs while doing so in front of a primarily liberal audience would signal more conservative beliefs. Experiment 1a used scenarios about judges, politicians, and businessowners asked for their views in front of large public audiences. Experiment 1b used scenarios about a college professor speaking at a faculty meeting and a cousin interacting with family at a barbeque.

1a Method.

Three hundred and one MTurk participants (mean age = 36.8, SD = 11.8, 49.8% female) read one of three short stimulus-sampled scenarios. Each described a case in which an actor is asked for their position on a contentious issue in a public forum. The first scenario described a county judge who is asked in an internet forum whether abortion should be legal or illegal. The second described a prominent businessman who is

[15] We initially sought to avoid exclusions altogether; however, we pre-registered exclusions based on attention in studies conducted during the revision process (Experiments 1b, 4b, and 5) in response to worries about data quality online in the wake of the COVID-19 pandemic. All reported results are robust to exclusions.

asked at a press conference whether NFL players should be required to stand or allowed to kneel during the national anthem. The third described an elected representative who is asked during a town hall whether she thinks confederate statues in her district should be taken down or left standing. In each case, the public figure opts not to take sides, saying that the issue is "very complex" and that they "cannot take a side at this time." Note that across our studies, we vary how actors verbalize their choice to stay out of it (see Table 4). We confirm that our stimuli resemble what people are actually inclined to say in our follow-up study (which appears after Experiment 6b).

We manipulated, in each scenario, the prevailing viewpoint of the actor's audience. For example, in the confederate statue vignette, the representative's constituents predominantly believe either (a) that the statues should be taken down or (b) that the statues should remain up. We refer to conditions in which the neutral actor's audience holds a stereotypically conservative position (i.e., that statues should remain up, that abortion should be illegal, or that NFL players should stand during the national anthem) as 'conservative audience' conditions (versus 'liberal audience' conditions when the audience holds the opposite views)[16]. We therefore randomly assigned participants to one of six cells in a 3 (vignette) x 2 (audience view) between-subjects design.

Participants then made an inference about the actor's underlying personal beliefs using a 7-point scale from -3 to +3, with -3 referring to a strongly held liberal view and +3 referring to a strongly held conservative view. For example, in the confederate statue

[16] Condition names were chosen for ease of exposition, rather than to test anything about liberal or conservative ideology.

scenario, -3 indicated that the representative "Strongly believes that the statues should be taken down" whereas +3 indicated "Strongly believes that the statues should be left up." In all cases, the midpoint of the scale indicated "is neutral on this issue." We expected to observe a medium effect of audience in each vignette ($d \sim .5$) and so set our target sample size to 50 per cell for this first experiment.

For this and all subsequent studies, age and gender demographics were collected at the end. Short multiple-choice attention checks (e.g., "what issue did you read about?") were also collected at the end of all studies to audit comprehension of basic details in our stimuli.

1a Results.

A two-way ANOVA detected a significant main effect of audience condition ($F(1,295) = 241.05, p < .001, \eta^2_G = .45$, 90% CI [.38, .50]). Participants interpreted an actor's choice not to take sides as signaling more conservative beliefs when addressing a liberal audience ($M = +1.11$, SD = 1.09) and more liberal beliefs when addressing a conservative audience ($M = -.89$, SD = 1.19). This effect of audience condition on belief inference held separately in all three vignettes (abortion: $t(95) = 11.96, p < .001, d = 2.4$, 95% CI [1.90, 2.96]; anthem-kneeling: $t(100) = 8.40, p < .001, d = 1.66$, 95% CI [1.21, 2.12]; confederate statues: $t(100) = 7.1, p < .001, d = 1.41$, 95% CI [.97, 1.84]; see Figure 9). Moreover, all six scenario conditions differed from the neutral midpoint (all *p*s < .001). Specifically, participants always attributed to neutral public figures beliefs which *opposed* the prevailing views of their audiences (e.g., when in front of a liberal audience,

staying out of a discussion about abortion rights signaled pro-life beliefs, and vice versa). We also found a main effect of vignette ($F(2,295) = 3.14$, $p = .045$, $\eta^2_G = .021$, 90% CI [.00, .050]) and a marginally significant vignette x condition interaction ($F(2,295) = 2.77$, $p = .065$, $\eta^2_G = .018$, 90% CI [.00, .047]), which each reflect variance in the predicted effect across qualitatively different scenarios. These results do not qualify our interpretation of the main effect of audience.

1b Method.

Experiment 1b was a pre-registered replication of Experiment 1a with a larger sample in two new contexts: a department meeting of college professors and a socially-distant barbeque among extended family. Each represented an interesting potential extension of our 1a result. The college professor scenario was designed to test a case in which the public figure's profession (teaching English) does not directly intersect with the issue in question (protests against the police). The extended family scenario was designed to generalize our effect beyond public figures.

Four hundred and thirty MTurkers were recruited (mean age = 38.9, SD = 13.3, 46.3% female) and followed the same procedure from Experiment 1a. Following our preregistration, nine additional participants were excluded for failing a simple multiple-choice attention check. We aimed to collect 100 subjects per cell in all scenario studies conducted after 1a. Participants read one of two vignettes. One described a college English professor who is asked at a department meeting what he thinks of protests against the police. The professor responds, "Gosh there are so many perspectives on this issue.

Truthfully, I'd rather not get into it right now." The other described a cousin at a family barbeque who is asked for her views on mask-wearing requirements at the local supermarket. She responds, "You know, this whole COVID mask thing is so complicated, and I really don't like talking politics with family." Audience beliefs were once again randomly assigned. In conservative audience conditions, the audience described in the scenarios (the professor's department colleagues, other family members at the barbeque) were depicted as holding stereotypically conservative beliefs (supporting police over protesters, opposing mask mandates). In liberal audience conditions, these observers held the opposite views. Participants again made inferences about the actor's personal beliefs using similar 7-point scales from -3 to +3, with -3 referring to a strongly held liberal view and +3 referring to a strongly held conservative view. 0 again indicated neutrality.

In addition to age and gender demographics, participants in Experiment 1b (and all remaining studies) also indicated their overarching political attitude on a 7-point scale from -3 "Strongly Liberal" to +3 "Strongly Conservative." 0 indicated "In the middle" (mean = -.5, SD = 1.8).[17]

1b. Results.

A two-way ANOVA detected a main effect of audience condition ($F(1, 426) = 289.51$, $p < .001$, $\eta^2_G = .40$, 90% CI [.35, .45]). Participants again interpreted an actor's choice not to take sides as signaling more conservative beliefs when addressing a liberal

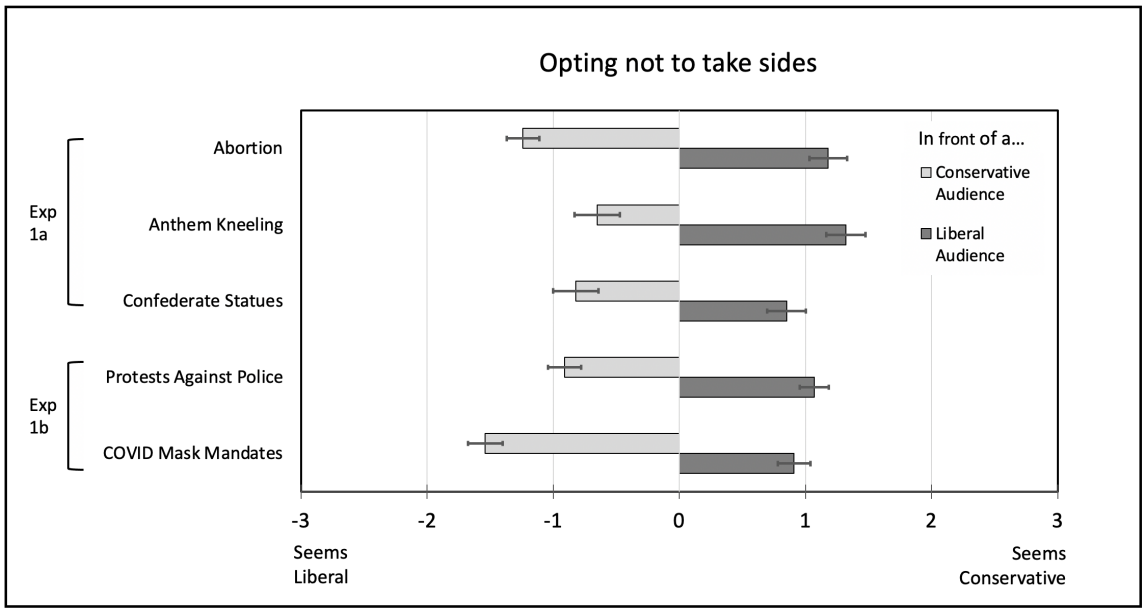
[17] Across studies, we did not detect any consistent patterns of interaction between either participants' general political attitudes or their views on specific issues and their inferences about the beliefs of actors who opted not to take sides.

audience (M = .99, SD = 1.28) and more liberal beliefs when addressing a conservative audience (M = -1.22, SD = 1.43). Effects of audience held separately in each vignette (protests: $t(213) = 11.29$, $p < .001$, $d = 1.54$, 95% CI [1.23, 1.85]; COVID masks: $t(213) = 12.74$, $p < .001$, $d = 1.74$, 95% CI [1.42, 2.05]; see Figure 9). Here, too, all four scenario conditions differed from the scale's midpoint (all ps < .001), such that participants always inferred opposition to the actor's audience from their choice not to take sides. We also found a weaker main effect of vignette ($F(1,426) = 9.17$, p =.003, η^2_G= .021, 90% CI [.004, .048]) and a marginal interaction ($F(1,426) = 2.99$, p =.085, η^2_G= .007, 90% CI [.00, .026]) which do not qualify our interpretation of audience effects on belief inference.

Discussion.

In Experiments 1a-b, opting to stay out of it telegraphed diametrically opposed underlying views when communicated to different audiences. Actors who opt not to take sides seem liberal in front of a conservative audience, but conservative in front of a liberal audience. This pattern held across actors from elected officials to academics to family members and across five different political issues.

Figure 9. (Chapter 3) Experiments 1a and 1b: Belief inferences by vignette and audience condition



Note. Error bars represent standard errors.

Table 4. (Chapter 3) How actors opted to "stay out of it" in our stimuli

Exp	Actor	Issue	Staying out of it stimulus
1a	Businessman	Kneeling for national anthem	"Well, I think the issue is quite complex, and I really can't take a side on it at this time."
	State Representative	Confederate Statues	
	Judge	Abortion	
1b	English Professor	Protesting the police	"Gosh there are so many perspectives on this issue. Truthfully, I'd rather not get into it right now."
	Cousin	COVID-19 mask mandates	"You know, this whole COVID mask thing is so complicated, and I really don't like talking politics with family."
2a	NFL Owner	Kneeling for national anthem	"We aren't doing anything on that … there's really nothing to talk about."
2b	Former Celebrity	Presidential politics	"We're neutral. I'm not taking either side. It's just uncomfortable."
3	State Representative	Kneeling for national anthem	"Well, I think the issue is quite complex, and I can't really take a side at this time."
4a	Anonymous Prolific worker	Gun control	(Computer message): "Your partner declined to take a side on this issue"
4b	Anonymous Mturk worker	Gun control	(Computer message): "Your partner declined to take a side on this issue"
5	New friend	Gun control	"You know, I'd really rather not take sides on that issue right now."
			"You know, I'd really rather not take sides on that issue right now. I just don't know enough about guns or gun policy to have an opinion."
			"You know, I'd really rather not take sides on that issue right now. I think there are good arguments on both sides of the gun debate."
			"You know, I'd really rather not take sides on that issue right now. I typically try not to talk about political issues with friends."
6a	County Judge	Gun control	"This is a very important and complex issue, and I'm not sure what I think. Consider me neutral."
6b	State Representative	Immigration/ Border Security	"You know, I think this is a very important issue, and personally, I just can't take a side on it at this time."

Note. Scenarios: 1a-b, 3, 5, 6a-b; Video clips: 2a-b; Incentivized economic games: 4a-b.

Experiments 2a and 2b. Replications with real video stimuli

Experiments 2a-b generalized Experiment 1a-b's belief inference results to more naturalistic stimuli: Press video clips of actors opting not to take sides in response to questions from reporters about their moral and political beliefs. In other words, participants observed actors staying out of it in the same fashion they might in the real world, via short news clips of public figures declining to respond to contentious topics. This time, participants were also asked to imagine that instead of staying out of it, the

actor had sided against their viewpoint explicitly and to predict whether they would, in that case, find them more or less trustworthy.

2a Method.

One hundred and eighty-seven participants (mean age 24.7, SD = 9.5, 73.2% female) from a university behavioral lab (i.e., one session of sign-ups) watched a ~30-second clip of the owner of an NFL team (the Kansas City Chiefs) responding to political controversy about whether players should be allowed to kneel during the national anthem (https://www.youtube.com/watch?v=XHe1w-7aRno). In the clip, a reporter at a press event asks: "Can you tell us where you are with the NFL's and the Chiefs' stance on the national anthem?" The owner avoids taking a side, saying "We aren't doing anything on that," and, "there's really nothing to talk about." Between conditions, we manipulated what participants were told about the press event's audience: that the fans of the team and viewers of the news station were either 'mostly conservative' (conservative audience condition) or 'mostly liberal' (liberal audience condition).

Participants completed the 7-point belief inference measure used previously (-3 'the owner strongly believes players should be allowed to kneel' to +3 'the owner strongly believes players should be required to stand'). On a separate page, participants were then asked for their own personal view on the issue in a binary forced choice: "I mostly believe players should be allowed to kneel" or "I mostly believe players should be required to stand" (86% and 14% picked each option, respectively). This binary variable

allowed us to assign each participant to an appropriate counterfactual for the following question.

We next asked participants to imagine that instead of opting to stay out of it the owner had made a statement that opposed *their* viewpoint directly (either that he believes 'players should be allowed to kneel' or 'players should be required to stand,' based on their answer to the previous question). For example, a participant who thought players should be allowed to kneel considered how they would feel if the owner had said, instead of avoiding a side, that he thought players should be required to stand. Participants predicted how much more or less sincere, trustworthy, and honest (random order, α = .77) they would find the owner on a scale from -3: "Much less" to +3: "Much more." Our two key dependent variables were belief inferences and this composite predicted trust measure. The specific trust items – trustworthiness, sincerity, honesty – were chosen to tap broader perceptions of trust as well as morally-relevant subcomponents, benevolence (sincerity of intention) and integrity (honesty/ethicality; Levine & Schweitzer, 2015). Participants also answered the same overarching political attitudes question from Experiment 1b (M = -1.19, SD = 1.40).

2a Results.

Belief Inferences. Replicating Experiment 1, participants who witnessed the football team owner choose not to take sides while speaking to a liberal audience attributed to him more conservative underlying beliefs (M = +.24, SD = 1.06) than those who learned he was speaking to a conservative audience (M = -.04, SD = .97; t(185) =

1.91, p = .058, d = .28, 95% CI [-.01, .57]). One-sample t-tests comparing mean belief inferences in each condition to the midpoint of the scale detected a significant effect only in the liberal audience condition ($t(95)$ = 2.21, p = .030, d = .23, 95% CI [.022, .43]), but not in the conservative audience condition ($t(90)$ = .43, p = .67, d = -.05, 95% CI [-.25, .16]), which may reflect participants' prior assumption that wealthy NFL owners lean conservative at baseline. In any case, these results suggest that the actor's implied beliefs seemed relatively more conservative when addressing a liberal audience and relatively more liberal when addressing a conservative audience.

Trust. We next tested whether participants might distrust someone who *implicitly* opposed them by opting not to take sides more strongly than someone who *explicitly* opposed them. To do this, we examined the subset of participants whose personal position on the kneeling issue happened to align with that of the actor's (randomly-assigned) audience (n = 97). For these participants, we expected that staying out of it would seem like implicit opposition. A one-sample t-test on the composite trust measure for this subset revealed a significant effect: Participants predicted that they would trust the actor more if he had espoused a view on the NFL kneeling controversy that they themselves eschewed (M = +0.33, SD = 1.23; $t(96)$ = 2.67, p = .009, d = .27, 95% CI [.069, .48]).

For participants who happened to *disagree* with the actor's audience (n = 90), remaining neutral signaled tacit *support* for the participants' viewpoint. The predicted trust question therefore invoked a different comparison; namely, between *implicitly agreeing with* the participants' position and *explicitly opposing it*. For example, would a

liberal participant respond more favorably to someone who tries to stay out of it but seems, in reality, to agree with them or to someone who explicitly opposes their position? We did not have any predictions about this comparison, but interestingly, we found that even in this case, people did not trust the actor who avoided a side any more than someone who opposed them outright (M = +.04, SD = 1.20; $t(89) = .29$, $p = .77$, $d = .031$, 95% CI [-.18, .24]).

2b Method.

Experiment 2b replicated Experiment 2a with a larger sample viewing a different video clip. In 2a, our sample size was restricted to one sitting of participants in a behavioral lab session. In 2b, we recruited 300 MTurkers (mean age = 37.1, SD = 11.9, 39.7% female), increasing our target to 150 per cell in light of smaller effects observed in 2a. Participants this time viewed a clip of a celebrity musician (Backstreet Boy AJ Mclean; https://www.youtube.com/watch?v=Qukk3xQM1Ac). In the clip, Mclean is approached in an airport baggage carousel by a reporter and asked for his personal political views. He responds, "We are neutral. I'm not taking either side. It's just uncomfortable." We again manipulated whether the reporter was said to represent a conservative channel with a primarily conservative audience or a liberal channel with a primarily liberal audience. Participants then inferred the celebrity's personal political beliefs on a 7-point scale from -3 'strongly liberal' to +3 'strongly conservative.' 0 indicated 'politically neutral.' Participants next indicated their own general political attitudes on a 6-point scale. In this experiment only, we eliminated the midpoint from this general political attitude measure, so as to separate participants by their predominant

personal viewpoint for the predicted trust question. (64% leaned liberal, 36% leaned conservative). As in Experiment 2a, liberals were asked to imagine that instead of saying that he would not take either side, the celebrity had said he leans conservative. Conservatives were asked to imagine that instead of saying that he would not take either side, the celebrity had said he leans liberal. Participants predicted how much more or less sincere, trustworthy, and honest they would find the actor in this imagined alternative scenario ($\alpha = .90$).

2b Results.

Belief Inferences. Participants who believed that the celebrity was speaking to a liberal news station attributed more conservative beliefs ($M = +.38$, SD = .96), and the opposite was true for those who believed he was speaking to a conservative news station ($M = -.34$, SD = 1.09; $t(298) = 6.15$, $p < .001$, $d = .71$, 95% CI [.48, .94]). One-sample t-tests comparing mean belief inferences in each condition to the neutral midpoint of the scale detected significant effects in both the conservative audience ($t(147) = 3.86$, $p < .001$, $d = -.32$, 95% CI [-.48, -.15]) and liberal audience conditions ($t(151) = 4.92$, $p < .001$, $d = .40$, 95% CI [.23, .57]).

Trust. We again examined first the subset of participants who shared the political view of the celebrity's audience (n = 147). A one-sample t-test on the predicted trust measure for this subset revealed a marginally significant effect, indicating that participants would trust the celebrity more on average if he had espoused a general

political attitude that they themselves opposed instead of saying he would not take either side (M = + .20, SD = 1.26; $t(146) = 1.92$, $p = .057$, $d = .16$, 95% CI [-.005, .32]).

Participants who did not agree with the audience they read about (n = 153) again judged an actor whose implied position aligned with their own as no more trustworthy than one who sided against them explicitly ($M = -.03$, SD = 1.11; $t(152) = .32$, $p = .75$, $d = -.02$, 95% CI [-.18, .13]).

Discussion.

In Experiments 2a-b, using more realistic stimuli, audience information again dictated participants' attributions of belief to actors who opt not to take sides. Specifically, in three of our four conditions, staying out of it was interpreted as tacit disagreement. In addition, among these actors' intended audience, staying out of it was generally expected to be less trustworthy than direct opposition. Meanwhile, when participants inferred that trying not to take sides meant tacit *agreement* with *their* view, they did not respond any more positively to it than to siding against their view outright.

Experiment 3. Comparing staying out of it to outright opposition

In Experiments 2a-b, participants predicted that they would trust an actor who opted not to take sides less than one opposed their viewpoint outright. However, these studies asked participants to compare staying out of it to an imagined alternative and predict trust. To eliminate potential demand characteristics and more directly examine

whether staying out of it backfires relative to stating an opposing position, Experiment 3 randomized participants to read about either an actor who opts not to take sides or one who expresses outright opposition. Participants then evaluated trustworthiness and reported voting intentions.

Method.

Four hundred and one MTurkers (mean age = 37.8, SD = 12.9, 41.1% female) read a short scenario adapted from Experiment 1a. In it, a businessman considering a run for office holds a press conference to get to know his constituents. He is asked by a reporter whether he thinks NFL players should be allowed to kneel or required to stand during the national anthem. We experimentally manipulated both what the businessman's audience mostly believed about the kneeling issue (conservative vs. liberal audience conditions) and also whether the businessman stays out of it or sides against his audience explicitly (not-taking-sides vs. opposition conditions). In both the not-taking-sides and opposition conditions, the businessman notes that "the issue is quite complex." In the not-taking-sides conditions, he continues, "and I really can't take a side at this time." In the opposition conditions, he continues, "but I believe that players should be (allowed to kneel/required to stand)," espousing whichever viewpoint *opposed* his audience's prevailing position. Participants were thus sorted into one cell in a 2 audience (conservative vs. liberal) x 2 response (not-taking-sides vs. opposition) between-subjects experiment.

Next, participants completed the belief inference measure from Experiment 1 and 2a. They then rated trust in the businessman on 7-point agreement scales ("Strongly disagree" to "Strongly agree") with three items: "The businessman is sincere," "The businessman is trustworthy," "The businessman is honest," ($\alpha = .96$). We also included an exploratory measure meant to assess participants' willingness to vote for the businessman for political office: "If I were a voter in this district, I would consider voting for the businessman."

Finally, participants provided their personal view on the kneeling issue on a 6-point scale from "believe strongly that players should be required to stand" to "believe strongly that players should be allowed to kneel." We binary coded this variable to capture each participant's prevailing view on the issue (64% believed NFL players should be allowed to kneel; 36% believed the opposite). Participants then completed age and gender demographics and answered the same broader political attitude measure used previously. Participants leaned liberal in their overarching political attitudes ($M = -.54$, $SD = 1.84$).

Results.

Belief inferences. A two-way ANOVA with audience condition (liberal vs. conservative) and response condition (not-taking-sides vs. opposition) as factors detected a main effect of audience ($F(1,397) = 468.81$, $p < .001$, $\eta^2_G = .54$, 90% CI [.49, .58]). This effect held separately for planned comparisons within each response condition. Unsurprisingly, when the businessman opposed a conservative audience outright by

saying outright that he thought NFL players should be allowed to kneel during the national anthem, he was believed to hold more liberal views on the issue (M = -1.62, SD = 1.38) than when he opposed a liberal audience by saying the opposite (M = +1.95, SD = 1.08; $t(197) = 20.36$, $p < .001$, $d = 2.88$, 95% CI [2.49, 3.28]). Replicating our previous results, saying that he "couldn't take sides" telegraphed more liberal beliefs when addressing a conservative audience (M = -.68, SD = 1.24) and more conservative beliefs when addressing a liberal audience (M = +1.08, SD = 1.21; $t(200) = 10.18$, $p < .001$, d = 1.43, 95% CI [1.12, 1.74]). Although there was no main effect of actor response ($F(1,397) = .076.$, $p = .78$, $\eta^2_G = .00$), we detected an interaction ($F(1,397) = 54.45$, $p <$ .001, $\eta^2_G = .12$, 90% CI [.074, .17]), indicating that choosing not to take sides led to less extreme belief attributions than outright opposition in both audience conditions (both ps < .001). In other words, as would be expected, explicit statements of disagreement signaled stronger opposition than the tacit opposition implied by staying out of it.

Trust. In this design, we again expected that saying that one "couldn't take sides" would be received differently based on whether the participant broadly agreed or disagreed with the actor's audience. For instance, participants who think that NFL players should be free to kneel during the national anthem should prefer someone who expresses support for that view in the face of an audience who thinks the opposite, as compared to someone who takes no side. In this case, taking a side clearly seems more praiseworthy than staying out of it. A more interesting and strict test concerns whether participants who agree with the actor's audience would respond to outright *opposition* more favorably than to staying out of it. For this reason, we first examined the results for trust and voting

intentions aggregated over all participants (n = 401), and then focused on the subset of participants (n = 181) who happened to share the viewpoint of the audience in their randomly assigned condition.

We did not detect an interaction between audience condition and actor response ($F(1,397) = .0059$, $p = .93$, $\eta^2_G = 0.00$), and so collapse across audience conditions for ease of exposition. Among the full sample, participants saw the public figure as substantially less trustworthy when he said he could not take sides ($M = 3.48$, SD = 1.56) than when he opposed his audience outright ($M = 5.39$, SD = 1.37; $t(399) = 13.05$, $p < .001$, $d = 1.30$, 95% CI [1.09, 1.52]). Among the target subsample (i.e., those participants who agreed with the group addressed in the scenario), not taking sides was also seen as much less trustworthy than outright opposition ($M = 3.25$, SD = 1.57 vs. $M = 5.01$, SD = 1.29; $t(179) = 8.03$, $p < .001$, $d = 1.20$, 95% CI [.88, 1.52]).

Voting Intentions. We again detected no interaction ($F(1,397) = 3.32$ $p = .069$, $\eta^2_G = 0.01$), and so again collapsed across audience conditions. Among the full sample, the businessman received more intended voting support for opposing his audience ($M = 4.62$, SD = 1.92) than for staying out of it ($M = 3.28$, SD = 1.67; $t(399) = 7.44$, $p < .001$, $d = .74$, 95% CI [.54, .95]). Among participants who agreed with the audience they had read about, the same trend emerged, albeit non-significantly ($M = 3.33$, SD = 1.80 vs. $M = 2.91$, SD = 1.60; $t(179) = 1.66$, $p = .10$, $d = .25$, 95% CI = [-.047, .54]). People were no more likely to vote for a businessman who took no side than for one who opposed their view outright: If anything, they were less likely to do so.

Discussion.

In Experiment 3, observers again used audience information to attribute underlying political commitments to actors who choose not to sides. Unsurprisingly, taking no side signals weaker opposition than outright disagreement: When the businessman directly sided against his audience, he was believed to oppose them more strongly than when he declined to take sides. Nevertheless, participants were more likely to trust and no less likely to consider voting for the businessman when he sided against their viewpoint outright than when he took neither side. Indeed, even among those whom such a strategy seems tailored to placate, staying out of it may backfire. Given that doing so also incurs the opportunity cost of failing to side *with* the supporters of either outright position, the aggregate costs of staying out of it can be steep.

Experiments 4a-4b. Behavioral effects on cooperation and partner choice

We have argued that when staying out of it resembles strategically concealed opposition, it can provoke distrust and backfire. Experiments 4a - 4b tested for effects of distrust on incentivized behavioral measures (4a: cooperation and 4b: partner choice) in an economic game. In Experiment 4a, participants signaled their beliefs on a contentious political issue to an anonymous partner and then learned that their partner had decided either to signal agreement, signal opposition, or decline to take sides. In Experiment 4b, participants signaled their beliefs and then picked between two potential partners, one who had responded with opposition and one who had declined to take sides. If staying out

of it truly harms trust, it should harm cooperation and partner-choice when real money is on the line.

Notably, these experiments further generalize our findings. Our studies thus far have explored observer responses to politicians, businesspeople, celebrities, college professors, and family members who opt to stay out of it. Experiments 4a-b sought to show that our effects can arise in any interpersonal setting where moral and political issues are up for discussion and there exists some incentive to conceal a controversial perspective.

4a Method.

Six hundred American citizens (mean age = 35.7, SD = 12.4, 49.7% female) were recruited from Prolific.com to participate in a study about political beliefs and cooperation. All were informed that they would be partnered with another worker, signal beliefs about an important issue, and then play a game with their partner for real bonus money.

Participants first learned the rules of a Prisoner's Dilemma game (Axelrod, 1980) and answered two multiple-choice comprehension check questions to ensure that they understood them. Participants were given two chances to answer these questions correctly before being removed for inattentiveness (prior to random assignment)[18]. We referred to

[18] We stopped recruiting when we reached 600 participants who passed the attention check and completed the entire study. We set our target sample to 200 per cell for Experiments 4a-b to increase power and account for binary DVs.

this game within the study as "The Reliability Game" and to the choice options (i.e., cooperate or defect) as "Rely" and "Avoid."

Participants were then instructed that they had been partnered with another Prolific worker and that they would participate in a belief signaling exercise before playing the Reliability Game for real bonus money. In actuality, all participants were assigned to signal their beliefs first, and we randomly assigned how their partners responded. To avoid deception, partners were recruited from a separate sample (n = 150), and their responses were randomly matched with our main study participants for payment purposes.

Participants selected between the following two choice options: "I believe ordinary citizens should be allowed to own guns" and "I believe ordinary citizens should NOT be allowed to own guns" in a forced choice (69% and 31% chose each option, respectively), Participants were then told that their beliefs had been shared with their partner, and that their partner had been given the option either to signal beliefs back or to decline to take a side on the issue of gun ownership. Note that participants were fully aware that their partners were given the freedom to decline to take sides by the experimenter. Partner responses were randomly assigned to be either agreement (responding by selecting the same statement as the participant), opposition (responding by selecting the opposite statement as the participant), or not-taking-sides (responding by selecting the option "decline to take sides"). After seeing their partner's response, participants reported trust on the same scale used in Experiment 3 (α = .92) and selected whether to cooperate or not in an incentive-compatible Prisoner's Dilemma. Participants

were also asked to make belief inferences, attributing to their partner convictions about gun ownership on a scale from -3 "My partner definitely believes that ordinary people should be allowed to own guns" to +3 "My partner definitely believes that ordinary people should NOT be allowed to own guns." 0 indicated "My partner's beliefs on this issue are neutral." In Experiments 4a-4b, we collected an additional exploratory item capturing the extent to which participants thought reasonable people could disagree about the issue of gun control but did not find any consistent interactions. Participants leaned liberal in their overarching political beliefs (M = -.65, SD = 1.72).

4a Results.

Belief inferences. Following our pre-registered plan, this time we recoded participants' belief inferences onto a -3 to +3 'perceived opposition' scale, such that numbers less than 0 always indicated inferred agreement and numbers greater than 0 always indicated inferred opposition. To do this, we simply reverse-coded responses from participants who indicated that ordinary people should be allowed to own guns.

In this experiment, participants were randomly assigned to see their partner indicate either outright agreement, outright opposition, or opt not to take sides. Unsurprisingly, a one-way ANOVA detected a significant main effect of condition on perceived opposition ($F(2, 597) = 659.92$, $p < .001$, $\eta^2_G = .69$, 90% CI [.66, .71]) . Participants in the agreement condition inferred that their partner agreed with their viewpoint (M = -2.30, SD = 1.16), while those in the opposition condition inferred that their partner opposed it (M = +2.42, SD = 1.21). As predicted, participants in the not-

taking-sides condition also inferred that their partner opposed their viewpoint (M = +.60, SD = 1.54; one-sample $t(199) = 5.46$, $p < .001$, $d = .39$, 95% CI [.24, .53], vs. the scale's midpoint) albeit much less strongly than in the opposition condition ($t(396) = 13.18$, $p < .001$, $d = 1.32$, 95% CI [1.10, 1.54]).

Trust. We pre-registered predictions that participants would trust their partner most in the agreement condition, that trust would be weaker in both the not-taking-sides and opposition conditions, and that not taking sides would garner no more trust than outright opposition. The results corroborated these predictions. A one-way ANOVA detected a significant omnibus effect of condition ($F(2,597) = 44.88$, $p < .001$, $\eta^2_G = .13$, 90% CI [.09, .17]). Participants trusted their partner more in the agreement condition (M = 5.27, SD = 1.10) than in the not-taking-sides condition (M = 4.20, SD = 1.28; $t(400)$ = 8.96, $p < .001$, $d = .89$, 95% CI [.69, 1.10]). They also trusted their partner more in the opposition condition (M = 5.07, SD = 1.22) than in the not-taking-sides condition ($t(396)$ = 6.98, $p < .001$, $d = .70$, 95% CI [.50, .90]). Interestingly, in terms of attitudinal trust, agreement and opposition differed only marginally ($t(398) = 1.67$, $p = .097$, $d = .17$, 95% CI [-.030, .36]). See Figure 2.

Cooperation. We pre-registered predictions that cooperation would be strongest in the agreement condition, that both not taking sides and opposition would see less cooperation, and that not taking sides would engender no more cooperation than outright opposition. The results again corroborated our predictions.

A Chi-Square test of independence detected an omnibus effect of condition on cooperation (χ^2(2 df) = 26.84, p < .001). Rates of cooperation were highest in the agreement condition (90.1%), followed by the opposition condition (72.7%), followed by the not-taking-sides condition (70.5%). Differences in cooperation between the agreement condition and both the opposition (χ^2(1 df) = 20.01, p < .001, OR = 3.41, 95% CI [1.95, 5.96]) and not-taking-sides conditions (χ^2(1 df) = 24.45, p < .001, OR = 3.81, 95% CI [2.19, 6.62]) were significant. There was no difference in cooperation between the not-taking-sides and opposition conditions (χ^2(1 df) = .24, p = .62). If anything, participants were less likely to cooperate with a partner who declined to take sides on a hot-button political issue than with one who opposed them outright (OR = .90, 95% CI [.58, 1.39])[19]. See Figure 3.

4b Method.

Experiment 4a found that both declining to take sides and outright opposition harmed trust and cooperation relative to agreement. We next examined how such behaviors impact partner choice: How would participants respond when given the opportunity to choose between a partner who opposed them outright and one opted not to take sides?

Experiment 4b followed a broadly similar design. 402 MTurkers (mean age = 39.8, SD = 12.4, 43.8% female) learned the rules of a Prisoner's Dilemma (again, "The

[19] Interestingly, rates of cooperation were relatively high across all conditions. *Post hoc*, we believe that this reflects the specific payout amounts we chose for the game. This pattern does not impact our interpretation of the results.

Reliability Game") and completed the same comprehension check questions with the same exclusion criteria. This time, participants learned that they had been randomly grouped with two other MTurkers, and that they would select which of the two to partner with for The Reliability Game.

To inform their choice, participants again engaged in a belief exchange exercise. Participants signaled their beliefs about gun control, communicating either that ordinary citizens should or should not be allowed to own guns (69% and 31% chose each option, respectively). One potential partner responded with outright opposition (selecting the opposite statement from the participant) while the other stayed out of it (by declining to take sides). The answer choices and language used for this exercise were identical to that used in 4a.

Participants rated the same trust items used previously for both partners ($\alpha_{opposition\text{-}partner}$ = .89, $\alpha_{neutral\text{-}partner}$ = .95) and, critically, selected which of the two partners to play the Reliability Game with. Partner choice served as our primary dependent variable. Participants then chose whether to cooperate or defect with their selected partner and completed the usual demographic questions. Participants leaned liberal in the overarching political beliefs (M = -.51, SD = 1.81).[20]

4b Results.

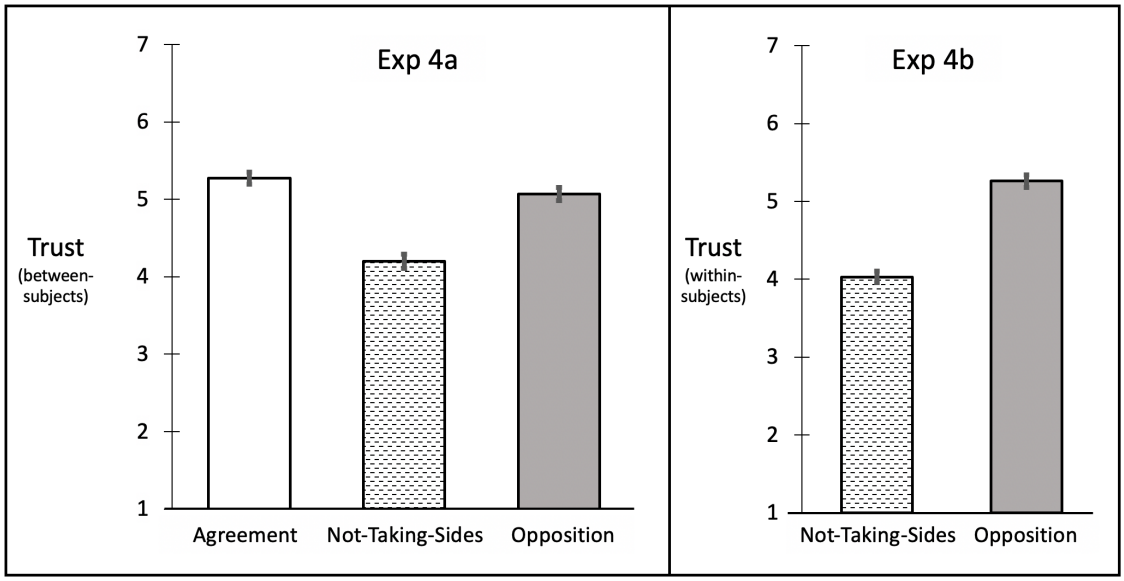
[20] We omitted belief inferences in this study only to ensure that making belief inferences salient by measuring them is not a necessary ingredient for our trust effects. We suspect that participants make the same general attributions whether we measure them or not.

Trust. Once again, participants indicated greater trust toward the partner who opposed their view on gun control outright ($M = 5.26$, SD = 1.17) than the one who did not take sides ($M = 4.03$, $SD = 1.46$; paired $t(401) = 13.17$, $p < .001$, $d = .66$, 95% CI [.55, .77]). See Figure 10.

Partner Choice. 61.2% of participants opted to play the prisoner's dilemma game with the partner who opposed their views on gun control outright, while only 38.8% opted to play with the partner who declined to take sides. A proportion test comparing this preference for outright opposition to chance revealed a significant effect ($\chi^2(df = 1) = 19.7$, $p < .001$). Moreover, logistic regression revealed that the preference for opposition (over not-taking-sides) was predicted by participant-level differences in attitudinal trust between the two responses ($B = .92$, SE = .11, Wald $Z = 8.51$, $p < .001$).[21] See Figure 11.

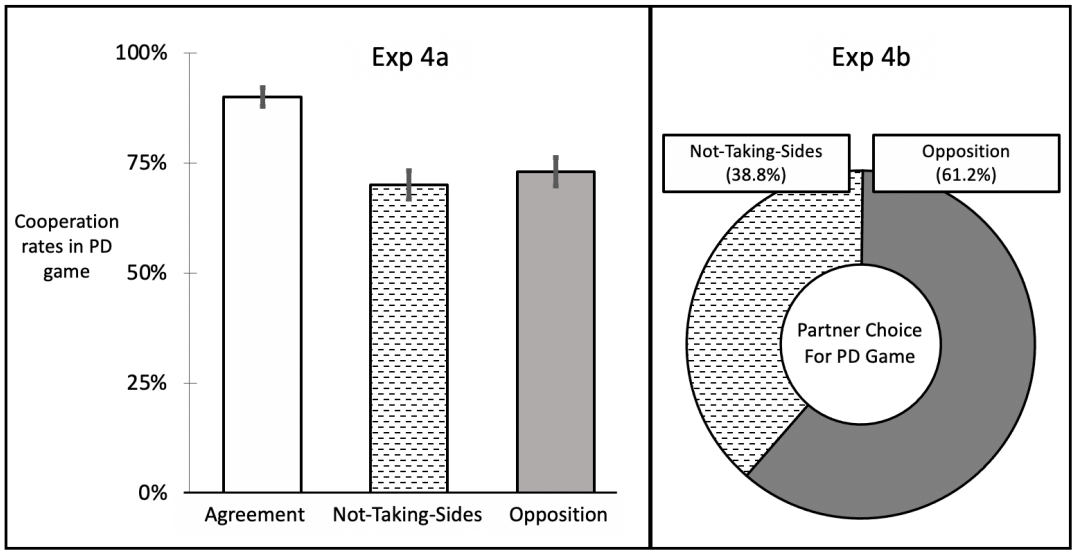
[21] Conditional on their choice of partner, participants' subsequent cooperation rates were similar whether playing the PD game with the opposition partner (71%) or the not-taking-sides partner (72%; $\chi^2(df = 1) = .001$, $p = .98$).

Figure 10. (Chapter 3) Experiments 4a and 4b: Attitudinal trust between- and within-subjects, in response to signals of agreement (4a only), opposition, or not-taking-sides



Note. Error bars represent standard errors.

Figure 11. (Chapter 3) Experiments 4a and 4b: Incentivized cooperation and partner choice for a prisoner's dilemma game, in response to signals of agreement (4a only), opposition, or not-taking-sides



Note. Error bars represent standard errors.

Discussion.

Experiments 4a and 4b provide further evidence for our belief inference and trust effects, and extend them in two ways. First, they show that our effects on trust are not restricted to public figures giving public interviews: When an anonymous partner's decision not to take sides signaled tacit disagreement with a participant's viewpoint, trust suffered. Second, they extend our results to an incentive-compatible cooperation paradigm: Declining to take sides provokes distrust that influences *behavior*, not just attitudes and intentions.

Note that in 4a, although participants trusted partners who opposed their views outright more than those who did not take sides, cooperation rates were similar across the two conditions. Post hoc, we suspect that cooperation decisions in these two conditions may be driven by different judgment processes: Whereas actors who stay out of it may see lower rates of cooperation because they are distrusted, ideological opponents may see lower rates of cooperation because they are disliked. Still, the results from 4b paint a more definitive picture. When asked to decide which they would rather have as a cooperative partner, participants preferred those who endorse views that they themselves eschew to those who opt not to take sides.

Experiment 5. The role of justifications for staying out of it

Our experiments focus on how observers respond to actors who opt not to take sides, but as we have noted, such public stances may reflect a variety of underlying dispositions. For example, an actor may choose not to take sides because they wish to avoid opposing their audience, because they feel sympathetic to both sides of an argument, or because they lack any opinion at all. Experiment 5 tested how observers would respond to different potential justifications for staying out of it. Broadly, we expected our prior results to be robust, but that the size of our effect might be influenced by the type of justification offered. In discussions of heated political issues, the presence of impression management incentives might activate suspicions about the motives behind

an avoidant stance, regardless of how it is justified. Nevertheless, we were curious whether certain justifications would soften distrust more than others.

Experiment 5 also sought to further generalize our effects beyond high stakes declarations made by public figures: Participants made judgments about the side-taking behavior of a new friend at an informal social gathering.

Method.

Five hundred forty-eight MTurkers (mean age = 39.6, SD = 13.1, 48.7% female) were recruited for a short scenario study. Participants began by rating their beliefs about gun control by selecting between two statements: "I am typically AGAINST most gun control efforts" or "I am typically IN FAVOR of most gun control efforts." (26% and 74% chose each option, respectively). We used this measure to ensure participants were presented with a scenario in which the actor's audience shared *their* beliefs. By doing so, we could more easily compare staying out of it to outright opposition.

Participants were then randomly assigned to one of five versions of a scenario about a social gathering among friends. Participants imagined meeting a new acquaintance at a small group hang-out. In the scenario, the group begins discussing the state of US politics and the issue of gun control specifically. All participants imagined that their friends by and large shared their views about gun control, but that they did not yet know what the new acquaintance believed.

We manipulated how the acquaintance responded when asked to weigh in. Four conditions depicted actors staying out of it and the fifth depicted outright opposition for

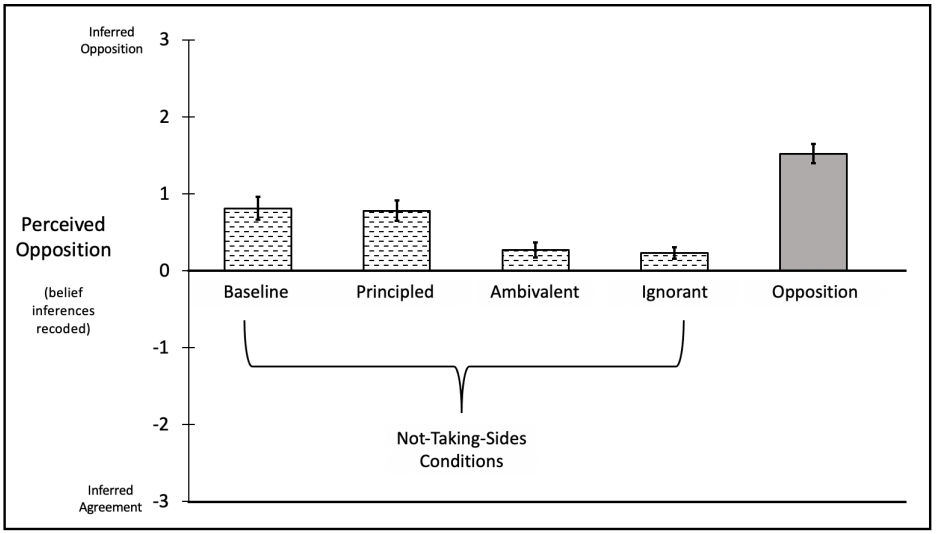
comparison. In the not-taking-sides-*baseline* condition, the acquaintance simply says: "You know, I'd really rather not take sides on that issue right now." In three additional conditions, the acquaintance says the same thing but also provides further justification for staying out of it (not-taking-sides-*ignorant* condition: "I just don't know enough about guns or gun policy to have an opinion"; not-taking-sides-*ambivalent condition*: "I think there are good arguments on both sides of the gun debate"; not-taking-sides-*principled*; "I try not to talk about politics with friends"). We note that in our follow-up study (which appears after Study 6b), these were the four most common categories of response from real people staying out of it in social scenarios, with the plurality providing no justification at all. A fifth condition depicted outright opposition to the participant and their friend group: "You know, all things considered, I personally think (do not think) more gun control would be a good thing for our country."

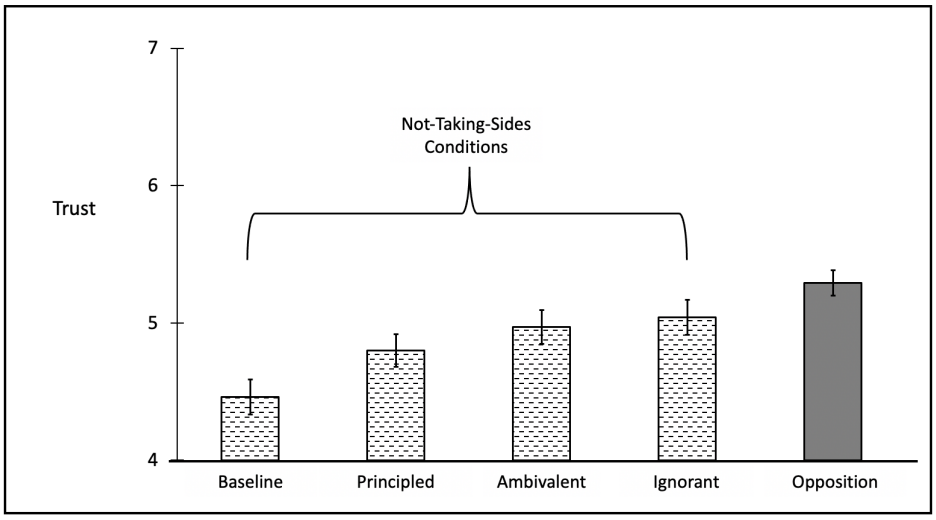
After reading, participants completed our usual belief inference measure (-3: "Strongly supports gun control efforts"; 0: "Is neutral on this issue"; +3: "Strongly opposes gun control efforts) and the same trust items used previously ($\alpha = .91$). In this study, we also included two measures of perceived competence as exploratory measures ($r = .93$); "This person is informed about the issue of gun control" and "This person is knowledgeable about the issue of gun control." Participants then answered a basic comprehension check by selecting whether the scenario was about gun control, police protests, or green energy. One additional participant failed and was excluded. Participants leaned liberal in their overarching attitudes (M = -.55, SD = 1.8).

Results.

Results for belief inferences and trust across all conditions from Experiment 5 are displayed in Figure 12. In the interest of brevity, we focus here on the most critical tests of our theory (primarily comparing different justifications for staying out of it to outright opposition), but significance tests for all comparisons are available in the supplement.

Figure 12. (Chapter 3) Experiment 5: Perceptions of opposition (-3 to +3; top panel) and attitudinal trust (1-7; bottom panel) from potential justifications for staying out of it





Note. Participants continue to perceive opposition and discount trust (relative to outright opposition) in the presence of explicit reasons for not taking sides. Error bars represent standard errors.

Belief inferences. Following our pre-registered plan, we recoded participants' belief inferences onto a perceived opposition scale (-3 →+3), with higher numbers indicating ascriptions of oppositional beliefs. A one-way ANOVA predicting perceived opposition detected a significant omnibus effect of condition ($F(4, 543) = 19.58$, $p < .001$, $\eta^2_G = .13$, 90% CI [.081, .16]). As previously, the outright opposition condition provoked the strongest attributions of opposition (M = 1.52, *SD* = 1.33) compared to all other conditions. It was followed by the not-taking-sides-baseline (M = .81, SD = 1.53; $t(218) = 3.67$, $p < .001$, $d = .50$, 95% CI [.23, .76]) and -principled (M = .78, SD = 1.38; $t(217) = 4.04$, $p < .001$, $d = .55$, 95% CI [.28, .82]) conditions, which differed significantly from opposition but not from one another; and then by the -ambivalent (M = .27, SD = 1.05; $t(216) = 7.17$, $p < .001$, $d = 1.04$, 95% CI [.76, 1.33]) and -ignorant (M =

.23, SD = .78; $t(219) = 8.8$, $p < .001$, $d = 1.19$, 95% CI [.90, 1.47]) conditions, which differed from all other conditions but not from each other. In line with our theory and prior results, participants inferred opposition from all not-taking-sides conditions relative to the scale's neutral midpoint (*ps* < .001).

Trust. A one-way ANOVA predicting trust detected a significant omnibus effect of condition ($F(4, 543) = 6.80$, $p < .001$, $\eta^2_G = .048$, 90% CI [.018, .075]). Participants trusted outright opposition the most ($M = 5.29$, SD = .96), with opposition differing from three of the four not-taking-sides conditions: baseline ($M = 4.46$, SD = 1.34; $t(218) = 5.31$, $p < .001$, $d = .72$, 95% CI [.44, .99]), principled ($M = 4.80$, SD = 1.25; $t(217) = 3.31$, $p = .001$, $d = .45$, 95% CI [.18, .72]), and ambivalent ($M = 4.97$, SD = 1.30; $t(216) = 2.10$, $p < .037$, $d = .28$, 95% CI [.017, .55]). Ignorance ($M = 5.04$, SD = 1.35; $t(219) = 1.62$, $p = .11$, $d = .22$, 95% CI [-.047, .48]) differed directionally but non-significantly from outright opposition. Both the ambivalent and ignorance conditions saw significantly more trust than baseline where no justification was provided. These results suggest that some forms of additional justification may shrink, but not necessarily close, the gap in trust between not taking sides and outright moral opposition. Even pleading ignorance earns no more trust than outright opposition.

Competence. A one-way ANOVA predicting perceived competence detected a significant omnibus effect of condition ($F(4, 543) = 37.66$, $p < .001$, $\eta^2_G = .22$, 90% CI [.16, .26]). Participants viewed the new friend as similarly knowledgeable and informed in the opposition condition ($M = 4.19$, SD = 1.3) relative to the principled ($M = 4.10$, SD = 0.94; $t(217) = .55$, $p = .58$) and baseline ($M = 4.33$, SD = 1.14; $t(218) = .88$, $p = .38$)

conditions. Unsurprisingly, the ignorance condition, in which distrust was weakest, also entailed a rather large competence penalty relative to opposition (M = 2.79, SD = 1.44; $t(219) = 7.57$, $p < .001$, d = -1.02, 95% CI [-1.30, -.74]) and to all other conditions (ps < .001). Unexpectedly, the friend who justified their opting not to take a position in terms of sympathy for both sides was seen as more informed and knowledgeable relative to outright opposition (M = 4.69, SD = 1.31; $t(216) = 2.87$, p = .004, d = .39, 95% CI [.12, .66]).

Discussion.

Experiment 5 generalizes our prior effects across a host of potential justifications in the context of a meeting a new friend: Outright opposition earned stronger trust than any of the justifications for staying out of it we tested. Appealing to ignorance as a reason for not taking sides came closest to eliminating the trust penalty documented in prior studies, but it also exacted a substantial hit to perceived competence.

Still, the results of this experiment provide reason to hope that adequately explaining one's reasons for staying out of it might mitigate its costs to some extent. For example, although simply noting the merits of both sides was still seen as less trustworthy than outright opposition, perhaps one could more articulately explain the nuances of one's moderate position in a way that might diminish distrust and signal knowledgeability. However, it may prove difficult to explain one's views effectively in the heat of a charged conversation about a contentious topic. Moreover, such a strategy would likely require the actor to engage in precisely the discussion they may have hoped

to avoid by staying out of it in the first place. As shown in Studies 2a-2b, even for public figures trained in public relations, actual attempts to stay out of it are often glib, off-the-cuff declinations to take sides, and such expressions, we demonstrate, often backfire.

Experiments 6a-6b. Strategic attributions drive negative responses to those who try not to take sides

We have argued that opting not to take sides can undermine trust and harm cooperation when it resembles strategically concealed opposition. However, not all contexts will lead to strategic attributions. Experiments 6a and 6b sought to demonstrate boundary conditions which contextualize our theory and rule out multiple potential alternative explanations. Experiment 6a investigated whether perceptions of opposition and associated distrust would be lessened if the actor lacks incentives for impression management. Experiment 6b provided observers with an explicit private signal of the actor's moderate beliefs and tested whether distrust would be attenuated as a result.

These moderations would be consistent with our theory but inconsistent with other reasons people may dislike and distrust attempts to stay out of it. For example, if people dislike not taking sides because they find non-responsiveness uncooperative (e.g., Grice, 1975), orthogonal manipulations of actor incentive or private signals of moderate belief should have no effect across cases in which what the actor actually says is held constant. If people dislike staying out of it because it signals moral conflict or apathy (Critcher, Inbar, & Pizarro, 2013; Zlatev 2019), a truly moderate position should, if anything, provoke *greater* distrust than seemingly concealed convictions. By contrast, we

predicted that removing social incentives (6a) or wiping out the inference of concealed opposition (6b) should attenuate the costs of staying out of it.

6a Method.

Experiment 6a was designed to test whether distrust in an actor who opts not to take sides would be lessened if the actor lacks incentives for impression management. 499 MTurkers (mean age = 39.3, SD = 12.8, 46.1% female) participated in a scenario study. At the outset, we asked participants for their beliefs about the issue of gun control, specifically whether ordinary citizens should be allowed to own assault weapons (40% said yes, 60% said no). As in Experiment 5, we used this question to ensure that participants always read a scenario in which the neutral actor's audience shared *their* overarching view on the issue.

Participants then read a short vignette about an elected county judge approaching the end of his term. In a public interview with either a predominantly liberal or conservative audience (according to the participant's own view), the judge is asked about his opinion on gun control. In all conditions, the judge says, "This is a very important and complex issue, and I'm not sure what I think. Consider me neutral." Participants were assigned to one of two versions of this scenario. In the high-incentives condition, the judge has decided to seek re-election and is giving an interview two weeks before the election. In the low-incentives condition, the judge has decided not to seek re-election and is giving an interview two weeks before retirement.

Participants again attributed beliefs on a 7-point scale, from -3: "Believes strongly that ordinary citizens should be allowed to own assault weapons"; to +3 indicating, "Believes strongly that ordinary citizens should NOT be allowed to own assault weapons." 0 indicated "is neutral on this issue." Participants also completed our usual trust items (α = .96). We predicted that participants would see the judge in the high-incentives condition, who is seeking reelection, as more strategically motivated than the judge in the low-incentives condition, who is planning to retire. Consequently, we predicted that the former would seem to be strategically concealing oppositional beliefs, but the latter might not, and that the former would seem less trustworthy as a result. Participants leaned liberal in their overarching political attitudes (M = -.39, SD = 1.77).

Results

Belief inferences. We again recoded the belief inference measure onto a 7-point perceived opposition scale, which captured the extent to which the neutral actor's beliefs seemed to oppose those of the participant (and those of the audience in the scenario that they read). We predicted that, in the eyes of observers, opting not to take sides would more strongly resemble strategically concealed opposition in the presence of incentives for impression management versus in the absence of such incentives. In line with this prediction, participants treated the judge's statement as a stronger signal of opposition when he had high vs. low incentives for impression management (M = +.45, SD = 1.50 vs. M = -.05, SD = 1.34; $t(497) = 3.92$, $p < .001$, $d = .35$, 95% CI [.17, .53]).

Trust. Moreover, not taking sides seemed less trustworthy in the high vs. the low incentives condition (M = 3.19, SD = 1.70 vs. M = 3.86, SD = 1.70; $t(497) = 4.42$, $p < .001$, $d = .40$, 95% CI [.22, .57]).

Process evidence. Finally, we predicted that the effect of incentive-condition on trust would be mediated by perceived opposition. This, too, was substantiated. Mediation analysis with 10,000 bootstrapped samples revealed a significant indirect effect of condition (0 = low incentives, 1 = high incentives) on trust via perceived opposition ($B = -.17$, SE = .05, 95% CI = [-.28, -.08]).

6b Method.

On our account, if an actor's public decision not to take sides seems to align with a privately held moderate or neutral position (i.e., if it signals no underlying commitment one way or the other), it should seem neither strategic nor untrustworthy. Experiment 6b was designed to test this prediction.

Five hundred and twenty-five MTurkers (mean age = 36.1, SD = 11.6, 41.9% female) read a short scenario about a prominent businesswoman running for office in a large diverse state. On the campaign trail, the businesswoman is asked to weigh in as to whether she supports increased spending on security at the U.S./Mexico border. We again wanted to compare to outright opposition, so, as in Experiments 5 and 6a, participants always read about a campaign rally in a district where voters predominantly shared *their* viewpoint. Participants indicated their personal view on a 6-point scale from "Strongly support increased spending on border security" to "Strongly oppose increased spending

on border security." Dichotomizing this variable, 46% supported and 54% opposed increased spending on border security. Participants who supported increased spending then read a scenario in which the potential supporters at the rally were predominantly conservative, while participants who opposed increased spending read a scenario in which the potential supporters at the rally were predominantly liberal.

Participants were assigned to one of three conditions. In all cases, the businesswoman is asked for her view on border security. She always responds by saying that the issue of border security is 'very important.' In the opposition condition, she opposes her audience (and the participants' personal viewpoint), by also saying either 'I oppose increased spending on border security' (when addressing conservatives) or 'I support increased spending on border security' (when addressing liberals). In the not-taking-sides condition, she instead says that she 'can't take a side on it at this time.' To tease apart whether participants object to middle-ground positions per se or to the attribution of strategic concealment, we also included a not-taking-sides *neutral beliefs* condition, which was identical to the other not-taking-sides condition, except that participants were also given private information from a conversation that the businesswoman later had with her husband where she says, "you've known me for years, and you know that I've always been neutral in the debate over border security and immigration." Although providing observers with an unambiguous private signal of truly neutral beliefs (i.e., no commitment one way or the other, a middle-ground position) is rarely possible in the real world, it is theoretically illuminating to know whether

participants would display a similar distrust if they did not attribute to them strategically concealed attitudes.

Participants were then asked to make inferences about the businesswoman's beliefs on a scale from -3 "strongly opposes increased spending on border security" to +3 "strongly supports increased spending on border security." 0 again indicated "is neutral on this issue." Participants answered the same trust items ($\alpha = .97$) and the same voting intentions items from Experiment 3. Participants leaned liberal in their overarching attitudes (mean = -.44, SD = 1.73).

6b Results.

Belief inferences. Following our pre-registered plan, we again recoded participants' belief inferences onto a -3 to +3 'perceived opposition' scale, such that larger numbers always indicated greater inferred opposition. A one-way ANOVA predicting perceived opposition detected a significant omnibus effect of condition ($F(2, 522) = 114.44, p < .001, \eta^2_G = .30$, 90% CI [.25, .35]) . Unsurprisingly, the businesswoman was believed to oppose her audience more strongly in the outright opposition condition ($M = 1.93$, SD = 1.41) than in either of the other conditions (not-taking-sides $M = .78$, SD = 1.40; $t(347) = 7.66, p < .001, d = .82$, 95% CI [.60, 1.04]; neutral beliefs $M = .03$, SD = .55; $t(350) = 16.70, p < .001, d = 1.78$, 95% CI [1.53, 2.03]). Comparing those two conditions, the businesswoman seemed to oppose her audience more strongly in the not-taking-sides condition than in the neutral beliefs condition ($t(347) = 6.62, p < .001, d = .71$, 95% CI [.49, .92]). In both the outright

opposition condition and the not-taking-sides condition, perceived opposition differed from the neutral midpoint (*p*s < .001). As we predicted, a private signal of neutral beliefs wiped out the inference that not taking sides concealed underlying opposition. We next examined whether appearing to hold genuinely neutral beliefs would eliminate the negative downstream effects of not taking sides. See Figure 13.

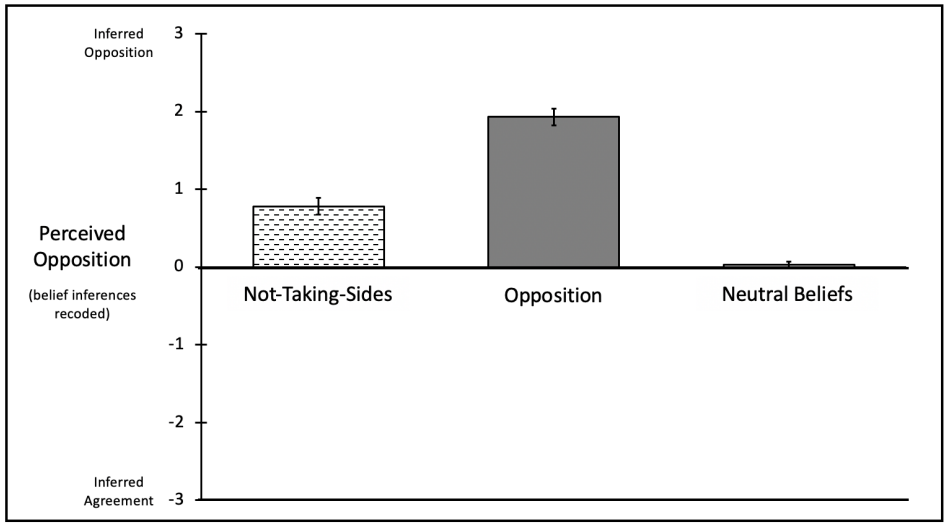
Trust. A one-way ANOVA detected an omnibus effect of condition ($F(2, 522) = 93.45$, $p < .001$, $\eta^2_G = .26$, 90% CI [.21, .31]). As previously, the businesswoman was seen as more trustworthy for siding against her audience (and the participant) than for staying out of it ($M = 5.21$, SD = 1.57 vs. $M = 3.04$, SD = 1.71; $t(347) = 12.32$, $p < .001$, $d = 1.31$, 95% CI [1.09, 1.55]). As predicted, adding a signal of privately neutral beliefs increased trust ($M = 5.02$, SD = 1.63; $t(347) = 11.05$, $p < .001$, $d = 1.18$, 95% CI [.95, 1.41]). The difference between opposition and neutral beliefs conditions was not significant ($t(350) = 1.13$, $p = .26$, $d = .12$, 95% CI [-.09, .33]). Both the opposition and neutral beliefs conditions saw greater trust than the not-taking-sides condition (in which participants attributed concealed opposition). See Figure 13.

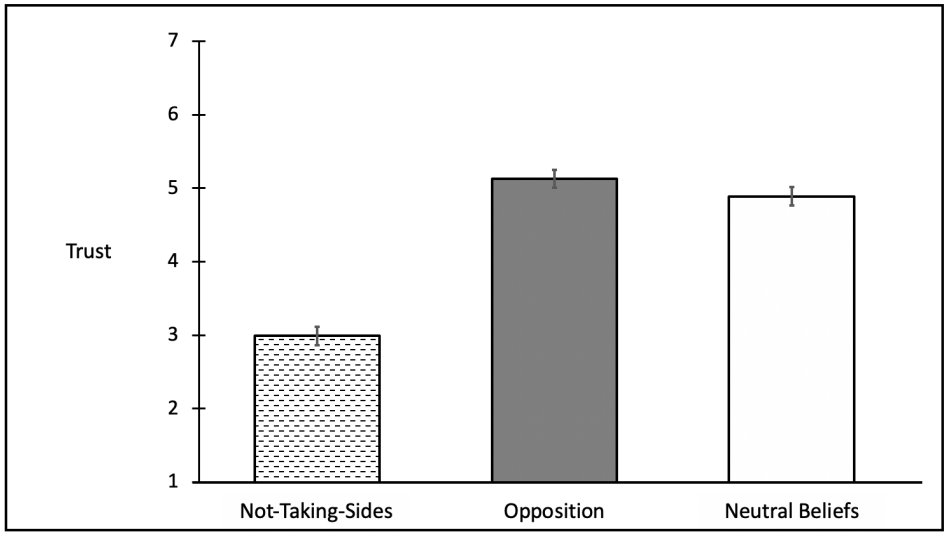
Voting Intentions. A one-way ANOVA detected a significant omnibus effect of condition ($F(2, 522) = 21.0$, $p < .001$, $\eta^2_G = .075$, 90% CI [.041, .11]). In this case, the businesswoman was more likely to receive voting support in the neutral beliefs condition ($M = 3.97$, SD = 1.57) than in either the not-taking-sides condition ($M = 2.86$, SD = 1.64; $t(347) = 6.43$, $p < .001$, $d = .69$, 95% CI [.47, .90]) or the opposition condition ($M = 3.07$, $SD = 1.85$; $t(350) = 4.87$, $p < .001$, $d = .52$, 95% CI [.31, .73]). We did not detect a difference between the not-taking-sides condition and outright opposition to the

participant's view ($t(347) = 1.14$, $p = .26$, $d = .12$, 95% CI [-.09, .33]). In the absence of a private signal of truly neutral beliefs, staying out of it garnered no more voting support than opposing the participant's view outright.

Process Evidence. To examine our proposed process further, we focused on the comparison between not-taking-sides and neutral beliefs conditions and fit an additional, exploratory serial mediation model. This model treated neutral beliefs vs. not-taking-sides as the independent variable (coded 0 and 1, respectively), inferred opposition as mediator 1, trust as mediator 2, and voting intentions as the outcome variable. This indirect pathway was significant ($B = -.25$, SE = .06, 95% CI [-.37,-.15]), consistent with our theorizing that staying out of it backfires specifically when it resembles concealed opposition, which in turn harms trust and erodes voting support.

Figure 13. (Chapter 3) Experiment 6b: Perceptions of opposition (-3 to +3; top panel) and attitudinal trust (1-7; bottom panel) by condition





Note. Error bars represent standard errors.

Discussion.

In Experiment 6a, when incentives for impression management were weakened, and in Experiment 6b, when not taking sides seemed to align with private neutral beliefs, distrust of those who opt to stay out of it was diminished. These moderations are consistent with our claim that opting not to take sides draws particular scorn when it seems to conceal hidden opposition. However, they are inconsistent with alternative accounts of our effects based on dislike of moral ambivalence or of non-responsive statements *per se*.

Although these boundary cases are theoretically important, we note that in the real world, actors who try to stay above the fray may struggle to effectively signal an *absence* of underlying beliefs, or to communicate an *absence* of reputational incentives, especially in high-stakes social contexts. Indeed, in practice, actors may find it difficult to send

unambiguous private signals of 'true' neutrality to observers, and they may be often assumed to have strategic motives and judged more negatively as a result.

Follow-up study: Actors may misunderstand the costs of staying out of it

We have suggested that, despite its costs, opting not to take sides seems intuitively appealing. To substantiate this claim, and to provide evidence that people may misunderstand the reputational consequences of staying out of it, we conducted a follow-up study. Shifting from the observer's perspective to the actor's, we sought to show that when facing ideologically hostile audiences, people often prefer not to share their opinions, and, more specifically, that they expect simply expressing a preference not to take sides to provoke less distrust than outright opposition. This follow-up study was also designed to explore what sorts of expressions people spontaneously generate when trying to avoid taking sides and to ensure that the stimuli used in our prior experiments resemble what people actually say when they attempt to stay out of it.

Method.

We ran two identical versions of this study: One with participants from a business school's behavioral lab (n = 203, mean age = 23.5, 67.5% female) and the other with MTurk workers (n = 292, mean age = 40.4, 42.5% female). Each survey was separately preregistered but with identical analysis plans. As results were similar across these two populations, we report analyses pooling across them (total N = 495). Separate analyses can be found in our supplemental materials and yield the same conclusions.

At the outset of the survey, participants were asked to list a hot-button social issue currently up for debate in America (e.g., on the news, on social media, in the workplace). Answers covered a variety of contemporary issues including COVID-19 policy, racial justice, abortion rights, etc. Participants then read a short workplace scenario designed to probe their intuitions about taking (or not taking) sides in a contentious discussion, given the explicit goal of building trust. The text of the scenario read:

> *You are being considered for a promotion to a position of leadership at work. In order to be selected, you need the people you work with to like and trust you. But you also know that you disagree with your coworkers about a hot-button political topic (the one you listed on the prior page). One day, you are sitting at lunch with your coworkers and the issue comes up. Your coworkers are talking about their opinions, and although you haven't said anything, you know you disagree with them. At some point, one of your coworkers turns to you and says, "Well, what do you think about all of this?"*

After reading, participants were reminded to consider that they needed others in this setting to trust them, and then asked what they would be most likely to do in this situation: (a) *Disagree with the group*, (b) *Try to stay out of it*, (c) *Agree with the group*. Participants were also asked to report exactly what they would say. The multiple-choice question probed people's intuitive preferences for taking (or not taking) sides, while the written-response question explored how people might verbalize preferences to stay out of it. Next, and on a separate page, we asked all participants to imagine that they had in fact decided to try to stay out of it and to select which of the following two strategies would be more effective for building liking and trust: (a) *Try to keep it short: Just say I prefer to stay out of it*, or (b) *Try to explain more deeply why I prefer to stay out of it*. This third

question captured people's general sense of whether staying out of it requires deeper justification to be considered socially acceptable.

Finally, participants reported their age, gender, and political attitudes (Mean = -.51, SD = 1.69). Three additional participants were excluded for giving nonsensical answers to the written-response question, as determined by hypothesis-blind RAs.

Results.

Looking first at answers to the "what would you do" question, 62.8% of participants indicated that they would *try to stay out of it*, a larger proportion than both other choice options combined (*disagree with the group* 32.3%, *agree with the group* 4.8%; $X^2(df = 1) = 32.1, p < .001$). People seem to see staying out of it as an intuitively appealing strategy for building trust with an audience hostile to their viewpoint.

When people opt to "stay out of it," what might they say? To investigate, we turned next to participants' written reports, enlisting the help of two hypothesis-blind RAs to code data and a third to arbitrate disagreements. As an initial check for data quality, RAs coded whether each participant wrote a response which matched their selection on the prior multiple-choice question. Did those who indicated a preference to take (or not to take) sides write responses which shared (or withheld) their opinion? 94% matched. Further coding focused on written responses which tried to stay out of it, including both responses from participants who indicated an explicit preference to stay out of it and wrote responses to match (n = 290), as well as responses from participants

who indicated a preference to take sides but wrote responses which clearly avoided doing so (n = 9).

Next, RAs sorted all such not-taking-sides responses (N = 299) into one of four categories identified *a priori* as possible ways people might express and justify their choice to stay out of it. These categories matched the conditions used in Experiment 5. The first captured simple expressions of preference not to take sides without further justification (e.g., "I'd rather not get involved in this issue"). The second captured expressions justified by a general principle or rule (e.g., "I'd rather not get involved in this issue because as a rule I don't talk politics at work"). The third captured expressions justified by ignorance or insufficient information (e.g., "I'd rather not get involved in this issue because I don't know enough about it to have an opinion"). The fourth captured expressions justified by a belief that both sides have merit (e.g., "I'd rather not get involved in this issue because on one hand I believe X, but on the other hand I also believe Y"). Finally, RAs could choose to indicate that no category fit the spirit of the response in question by selecting a fifth "other" category to be explored informally. Results of this coding are displayed in Table 5 below:

Table 5. (Chapter 3) Participant-generated responses opting not to take sides, categorized by type of justification provided

Justification for staying out of it	Count	% of all "staying out of it" responses
None	147	49.2%
General principle	69	23.1%
Ignorance of issue	44	14.7%
Both sides have merit	18	6.0%
Other	21	7.0%

The most common strategy by far for those who preferred not to take sides was to simply say so without further justification (49.2%), followed by appealing to a general principle of public neutrality (23.1%), ignorance of relevant information (14.7%), or the merits of both sides (6.0%). Only 7.0% of responses were identified as fitting none of these categories. This coding suggests that when given the goal of maintaining trust, many people may view simply stating that they would rather not to get involved (vs. providing deeper justification) as an effective strategy.

Importantly, this preference for short and simple expressions does not seem attributable to participants trying to speed through our study. Indeed, although participants were free to use curt expressions like "no comment" or "leave me alone," very few actually did, with over 80% of not-taking-sides responses being 10 words or longer. At the same time, not-taking-sides responses (M = 17.6) were 32% *shorter* on average compared to taking-sides responses (M = 26.0 words; $t(493) = 6.21$, $p < .001$), which likely reflects a more general intuition that politely staying out of it requires less explaining than sharing one's position. Finally, turning to the last question in our survey,

when asked explicitly whether providing a deeper explanation for staying out of it would be helpful for building liking and trust, 74.7% indicated that a short answer ("just say I'd rather stay out of it") would prove more effective. That is, even when endorsing the longer expression took no additional time, participants still had the intuition that short and simple was the way to go.

Discussion.

Our follow-up study tested people's intuitions about (not) taking sides from the perspective of the actor. The data suggest that when given the explicit goal of building trust and faced with the prospect of siding against their audience, people prefer not to take sides, suggesting that actors may miscalculate the costs of staying out of it. Moreover, participants who indicated a desire to stay out of it typically used short, straightforward expressions, often saying things like "I prefer to keep my political opinions to myself" or "You know, I just don't like to talk politics." In other words, participants put on the spot to take sides often generated just the sorts of not-taking-sides responses we have previously demonstrated backfire.

General Discussion.

Taking the wrong side on a hot-button political issue in public can have serious interpersonal consequences. Yet our results suggest that refusing to take sides carries its own risks. Across our experiments, we find that choosing not to take sides is often interpreted as strategically concealed opposition to the audience's prevailing position. As

a result, staying out of it often seems less trustworthy than outright opposition; it makes one a less desirable cooperative partner; and it fails to win additional voting support or engender increased cooperation even among those it seems most likely to placate. Indeed, because staying out of it also entails failing to side *with* either side, staying out of it can substantially undermine trust and support overall. Importantly, these effects persist across a host of paradigms, contexts, actors, issues, and ways to articulate one's reticence to take sides.

However, if opting not to take sides can be attributed to non-strategic motives, these patterns of inference and distrust attenuate, suggesting that our results do not reflect distrust of moderate positions or non-responsiveness *per se*, but concerns that staying above the fray may represent a deceptive attempt at impression management. Accordingly, we find that observers respond less negatively to attempts to stay out of it when they seem genuine, either because they are divorced from reputational incentives or accompanied by a private assurance that the speaker does not harbor partisan convictions one way or the other. These boundary conditions corroborate and clarify our account, and they also differentiate it from multiple potential alternatives.

For example, prior work on political apathy and moral indecision (Critcher, Inbar, & Pizarro, 2013; Zlatev, 2019) might predict distrust in some of our cases, but such accounts should, if anything, predict *stronger* distrust of actors who seem genuinely uninformed or torn between sides (as compared to those who say they prefer not to take sides but seem to harbor private convictions). This is the opposite of what we find in Experiments 5 and 6b: When observers can attribute an actor's decision not to take sides

to ignorance or genuine long-term indecision, distrust is weakened. Relatedly, our results cannot be explained by people's general distrust of those who hold abhorrent views or who abet moral injustice (Baron & Ritov, 2004). Such accounts would predict that actively endorsing the 'wrong' position (as in our opposition conditions) should seem worse than tacitly supporting them (as in our not-taking-sides conditions, where perceived opposition was present but weaker). Here, too, our studies find the opposite: Outright opposition was more trustworthy than taking no side at all.

One might reasonably wonder how our results relate to work on conversational norms (e.g., Grice, 1975). In our estimation, such work would not necessarily predict our effects, nor could it obviously explain them *ex post*. For example, although conversational norms prescribe that speakers should be responsive to direct questions where possible, they also prioritize politeness, dictating that information which might offend or cause conflict should often be kept private (Yoon, Tessler, Goodman, & Frank, 2020). Thus, it is not clear that trying to stay out of it, and seeming unresponsive, should seem a worse violation than opposing one's audience outright. But even if non-responsiveness were the greater sin, conversational norms might not obviously predict attenuations of our effect across cases which hold conversational behavior constant (e.g., Experiments 6a-b) or which entail no conversation at all (Experiments 4a-4b). An account based on negative responses to perceived impression management affords us sharper predictions and clearer explanations. Stepping back, while Gricean norms are typically thought to govern utilitarian exchanges of facts and information (Grice, 1975), discussions of political *opinion* may serve other social purposes (coalition building,

impression formation, persuasion), which complicate the situation and merit more targeted study.

Here, we focused on how observers respond to actors who use simple and straightforward language to express their choice to "stay out of it," as compared to actors who oppose observers' moral beliefs outright. This approach is not without realism. As we show in our follow-up study, when facing audiences with whom they disagree, people spontaneously offer statements like "I'd rather not talk about my political views right now" and endorse them as effective for building trust specifically. In fact, when asked explicitly whether providing deeper justifications for choosing to stay out of it would help make a better impression, people did not seem to think that it would. Even public figures, who are presumably well-versed in dealing with media relations and public perception, sometimes employ similar tactics, and we replicate our results with two such naturalistic cases in Experiments 2a and 2b.

Of course, our studies do not investigate every manner in which the choice to stay out of it might be expressed or justified, and some conversational strategies for avoiding a strong position may prove more effective than others. Building on our initial exploration of justification tactics in Experiment 5, future work can broaden the range of conversational approaches tested as a means of identifying boundary strategies that might work better. At one end of the spectrum, it is not hard to imagine glibber responses like "no comment" faring even worse than those we tested, as these might seem not only deceptive and untrustworthy per our account, but also ruder and more abrupt. At the other, it seems possible that, especially for those whose beliefs really do fall somewhere

in the middle, a more in-depth discussion of one's moderate or pragmatic preferences may help to soften the penalties associated with staying out of it. Indeed, emerging work suggests that effectively balancing the interests of multiple moral perspectives in pursuit of solutions that work for all sides can signal authentic moral character and garner respect (Puryear & Gray, 2022). And, more broadly, deeper and more connective conversations about personal experience can sometimes lesson interpersonal political hostility and bring people together (e.g., Kubin, Puryear, Schein, & Gray, 2021; Kardas, Kumar, & Epley, 2021). Still, while intriguing, such strategies are also more effortful, and they may not come readily to mind for those put on the spot to take sides and worried about saying the wrong thing. Moreover, the potential benefits of these strategies come with an ironic cost: If to avoid distrust one needs to carefully explain the nuanced beliefs behind one's choice not to take sides, there is an important sense in which one is not really free to stay out of it.

Zooming out, we hope future scholarship will continue to explore how observers respond to strategies for dealing with nuance and finding common ground in discussions of polarized issues. What sorts of inferences do people make about someone who plays devil's advocate, or who surfaces ideologically-inconvenient evidence in the name of impartial fact-finding, or who admits that their group's perspective may sometimes be biased? Unfortunately, we suspect that for touchy two-sided issues, such strategies, however well-intended, may also elicit skepticism and provoke distrust. In a polarized environment with clear social incentives, sending a credible signal that one harbors no strategic agenda may prove difficult.

Theoretical implications and future directions

Our findings contribute to emerging research on the psychology of side-taking (DeScioli & Kurzban, 2013; Shaw, DeScioli, Barakzai, & Kurzban, 2017) and specifically highlight the nuanced inferences people make about those who try to avoid two-sided issues in the political sphere. Although interest in political conflict and affective polarization has exploded in recent years (Finkel et al., 2020; Van Prooijen & Krouwel, 2019; Westfall, Van Boven, Chambers, & Judd, 2015), little is known about people's judgments of those who try to cut a middle path by trying not to get involved. This gap is important given that discussions of contentious political issues ensnare friends, family members, co-workers, businesspeople, celebrities, and politicians, with serious interpersonal and societal consequences. Here, we provide the first demonstration that observers make sophisticated belief attributions and character judgments from ostensibly signal-less choices not to take sides. As this was the first investigation into this issue, we chose to focus on expressions directed at relatively homogenous audiences (i.e., groups holding a common prevailing opinion, as is often the case in polarized environments). However, future research can explore how our effects play out with undecided or mixed audiences. Broadly, we suspect that observers may interpret staying out of it as strategic concealment in such cases, too, although they may struggle to pinpoint exactly which beliefs and opinions actors are trying to conceal.

Furthermore, this research advances a practical understanding of how actors deal with difficult questions in public contexts and how strategies for doing so are interpreted by observers. Previous work has explored a slew of evasive rhetorical tactics which allow actors to respond to direct questions without offering any substantive answers (Bitterly & Schweitzer, 2019; Rogers et al., 2017; Rogers & Norton, 2011). Although we agree that strategies like dodging and paltering are prevalent and fascinating, attempts to earnestly avoid take sides have received scant attention. Our central result – that such strategies provoke stronger distrust than outright opposition - may seem surprising, since disagreement over divisive issues is known to provoke anger, prejudice, and even violence (Skikta, 2010) and that maintaining impartiality is often a virtue (Shaw, Barakzai, & Keysar, 2019), particularly for those in positions leadership (Everett, Faber, Savulescu, & Crockett, 2018). Yet, at least in the contexts we examined, penalties associated with staying out of it were often steeper.

Nevertheless, our follow-up study suggests that when faced with the prospect of fragmenting support by taking a controversial position, staying out of it seems intuitively attractive. Why might this be so? Perhaps in context, actors underweight the indirect cost of seeming deceptive against the salient risk of directly angering observers, or perhaps they fail to realize they might inadvertently portray themselves as evasive or inauthentic at all. Another possibility is that actors might choose to stay out of it because they assume that doing so will garner less attention (e.g., social gossip, news coverage) than taking a strong stand one way or another. Yet another possible explanation is that staying out of it provides some cover for actors to change their position if public opinion later shifts (i.e.,

without seeming hypocritical; Effron, O'Connor, Leroy, & Lucas, 2018). Building on the results from our follow-up study, future authors can investigate further the intuitive pull of staying out of it by asking *why* actors choose to stay out of it and whether any specific expectations about doing so are warranted. Our data thus far seem to align with the 'impression mismanagement' thesis (Steinmetz, Sezer, & Sedikides, 2017), that people sometimes adopt self-presentation strategies which actually portray them more negatively.

We certainly do not mean to imply that not taking a side is always the wrong choice. Penalties for standing on the *wrong* side of a contentious political issue in public are well-documented and sometimes severe. In our studies, opting to stay out of it softened perceived opposition relative to outright disagreement and, in some cases, this alone may be worth the costs of seeming less trustworthy. Seeming evasive and untrustworthy may in some cases be preferable to seeming to hold an overly extreme view, particularly if the issue in question is important to one's peers, constituents, or customers. More generally, there remain open questions about other dimensions of evaluation which may prove relevant. How might staying out of it impact perceptions of confidence, moral conviction, or fitness for specific group roles? More data here will likely offer a broader lens on the costs and benefits of (not) taking sides.

A further question concerns whether certain leadership roles in society are protected from the patterns of inference and judgment documented here. For example, judges, high-level bureaucrats, and even journalists are sometimes expected to maintain political impartiality in fulfilling their professional responsibilities. We suspect that for

these sorts of actors, choosing not to take sides may not harm trust to the same extent because observers may attribute it to role-specific norms of conduct rather than to impression management motives. Roles that explicitly require neutrality may thus represent an interesting potential boundary to our effects, and they may serve an important social function more broadly – allowing actors to occupy middle-ground positions for the purposes of impartial information gathering and arbitrating disagreement.

For actors who do anticipate and fear the repercussions of conspicuously staying out of it, the backfire effects we document here seem to incentivize taking sides. In certain cases, this dynamic may compel actors to endorse positions on contentious issues they know little about or to express convictions they actually lack. Without clear and convincing communication about nuanced or moderate positions, the psychology we document may lead observers to sort those who try to avoid conflict according to a "with me or against me" mindset, leaving little room to surface nuance or deescalate conflict. Importantly, such effects may not be limited to the domain of political discourse: Staying out of it may provoke distrust and dislike across myriad issues in family feuds, workplace disputes, disagreements on social media, negotiations, or even legal proceedings – anywhere reputational incentives are on the line and avoiding the conversation might appear strategic. Future researchers should look to generalize our effects to other domains of public disagreement and to incorporate the social costs of staying out of it into broader explanatory theories of polarization and intergroup conflict.

Conclusion

Open-minded discussion of complex and consequential issues is a hallmark of a well-functioning society. But a healthy public debate also affords its participants the freedom to avoid taking sides. Our results suggest that, in practice, trying to exercise that freedom can backfire. We present evidence that choosing to stay out of it is often imbued by observers with predictable patterns of social meaning and that it can provoke skepticism and distrust as a result. These findings advance our understanding of social incentives for side-taking in moral conflict, and we hope they open avenues for mitigating ideological disagreement more broadly.

References

Alford, J. R., Hatemi, P. K., Hibbing, J. R., Martin, N. G., & Eaves, L. J. (2011). The politics of mate choice. *The Journal of Politics*, *73*, 362-379.

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, *24*, 3-25.

Baum, S. M., & Critcher, C. R. (2019). The Costs of Not Disclosing. *Current opinion in psychology*.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*, 74-85.

Bonica, A., Rosenthal, H., Blackwood, K., & Rothman, D. J. (2020). Ideological Sorting of Physicians in Both Geography and the Workplace. *Journal of Health Politics, Policy and Law*. Forthcoming.

Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The braggart's dilemma: On the social rewards and penalties of advertising prosocial behavior. *Journal of Marketing Research*, *52*, 90-104.

Bitterly, T. B., & Schweitzer, M. E. (2020). The economic and interpersonal consequences of deflecting direct questions. *Journal of personality and social psychology*, *118*, 945.

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science*, 23, 27-34.

Byrne, D. (1971). The attraction paradigm. New York, NY: Academic Press.

Chambers, J. R., Schlenker, B. R., & Collisson, B. (2013). Ideology and prejudice: The role of value conflicts. *Psychological Science*, *24*, 140-149.

Chambers, J. R., & Melnyk, D. (2006). Why do I hate thee? Conflict misperceptions and intergroup mistrust. *Personality and Social Psychology Bulletin*, *32*, 1295-1311.

Chen, M. K., & Rohla, R. (2018). The effect of partisanship and political advertising on close family ties. *Science*, *360*.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*, 308-315.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*, 769-771.

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological bulletin*, *139*, 477.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., ... & Zinger, J. F. (2019). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, *14*(2), 273-291.

Effron, D. A., O'Connor, K., Leroy, H., & Lucas, B. J. (2018). From inconsistency to hypocrisy: When does "saying one thing but doing another" invite condemnation? *Research in Organizational Behavior*, *38*, 61-75.

Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200-216.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., ... & Skitka, L. J. (2020). Political sectarianism in America. *Science*, *370*, 533-536.

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*, 1-12.

Gantz, J. & Murray, V. V. (1980). The experience of workplace politics. *Academy of Management Journal*, *23*, 237-251.

Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics 3: Speech acts* (pp. 41-58).

John, L. K., Barasz, K., & Norton, M. I. (2016). Hiding personal information reveals the worst. *Proceedings of the national academy of sciences*, *113*, 954-959.

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. *Psychological Perspectives on the Self*, *1*, 231-262.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological science*, *28*, 356-368.

Jost, J. T. (2017). The marketplace of ideology: "Elective affinities" in political psychology and their implications for consumer behavior. *Journal of Consumer Psychology*, *27*, 502-520.

Kardas, M., Kumar, A., & Epley, N. (2021). Overly shallow?: Miscalibrated expectations create a barrier to deeper conversation. *Journal of Personality and Social Psychology*.

Kubin, E., Puryear, C., Schein, C., & Gray, K. (2021). Personal experiences bridge moral and political divides better than facts. *Proceedings of the National Academy of Sciences*, *118*(6).

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88-106.

Marks, J., Copland, E., Loh, E., Sunstein, C. R., & Sharot, T. (2019). Epistemic spillovers: Learning others' political views reduces the ability to assess and use their expertise in nonpolitical domains. *Cognition*, *188*, 74-84.

Parker, M. T., & Isbell, L. M. (2010). How I vote depends on how I feel: The differential impact of anger and fear on political information processing. *Psychological Science*, *21*, 548-550.

Pew Research Center (2019). The public's level of comfort talking politics and Trump. appears in: *Public highly critical of state of political discourse in the US*.

Puryear, C., & Gray, K. (2021). Using "Balanced Pragmatism" in Political Discussions Increases Cross-Partisan Respect. Pre-print.

Rogers, T., & Norton, M. I. (2011). The artful dodger: Answering the wrong question the right way. *Journal of Experimental Psychology: Applied*, *17*, 139.

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, *112*, 456.

Rogowski, J. C., & Sutherland, J. L. (2016). How ideology fuels affective polarization. *Political Behavior*, *38*, 485-508.

Schlenker, B. R. (1980). *Impression Management*. Monterey, CA: Brooks/Cole Publishing Company.

Shaw, A., DeScioli, P., Barakzai, A., & Kurzban, R. (2017). Whoever is not with me is against me: The costs of neutrality among friends. *Journal of Experimental Social Psychology*, 71, 96-104.

Shaw, A., Barakzai, A., & Keysar, B. (2019). When and Why People Evaluate Negative Reciprocity as More Fair Than Positive Reciprocity. *Cognitive Science*.

Silver, I., Newman, G., & Small, D. A. (In Press). Inauthenticity aversion: Moral reactance toward tainted actors, actions, and objects. *Consumer Psychology Review*.

Silver, I., & Shaw, A. (2018). No harm, still foul: Concerns about reputation drive dislike of harmless plagiarizers. *Cognitive science*, *42*, 213-240.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, *4*, 267-281.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*, 359-393.

Steinmetz, J., Sezer, O., & Sedikides, C. (2017). Impression mismanagement: People as inept self-presenters. *Social and Personality Psychology Compass*, *11*(6).

Thomas, A. R. & Duncan, C. (1999). Chapter 7: The Law of Neutrality. In *The Commander's Handbook on the Law of Naval Operations*. International Law Studies, 73, 365-400.

Toner, K., Leary, M. R., Asher, M. W., & Jongman-Sereno, K. P. (2013). Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority. *Psychological Science*, *24*, 2454-2462.

Van Prooijen, J. W., & Krouwel, A. P. (2019). Psychological features of extreme political ideologies. *Current Directions in Psychological Science*, *28*, 159-163.

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science*, *10*, 145-158.

Wosinka, W., Dabul, A. J., Whetstone-Dion, R., & Cialdini, R. B. (1996). Self-presentational responses to success in the organization: The costs and benefits of modesty. *Basic and Applied Social Psychology*, *18*, 229-242.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite Speech Emerges From Competing Social Goals. *Open Mind*, *4*, 71-87.

Zlatev, J. J. (2019). I may not agree with you, but I trust you: Caring about social issues signals integrity. *Psychological Science*.

APPENDIX NOTE

Note that each of the three chapters has a project-specific online repository where relevant additional materials are publicly available. Preregistrations, data, study materials, and supplementary analyses can be found here. Details on the specific content that can be found at each link is also highlighted in the relevant chapter.

Chapter 1:

https://researchbox.org/105&PEER_REVIEW_passcode=MKFQMJ

Chapter 2:

https://researchbox.org/311&PEER_REVIEW_passcode=BXQOFM

Chapter 3:

https://researchbox.org/118&PEER_REVIEW_passcode=OJNSOY