# Can Social Pressure Stifle Free Speech?*

Juan S. Morales† Margaret Samahita‡

August 15, 2023

**Abstract**

This paper studies public opinion in the context of strong social norms that can induce conformity and self-censorship. We present a model that highlights how social pressure can affect the public expression of opinion either through a change in publicly stated views (*conformity*) or by inducing self-censorship (*silence*). In a series of pre-registered online experiments in the US, we elicit participants' views on two controversial topics (race and gender) and their willingness to publish these views online in an incentivized manner. The empirical patterns are consistent with the presence of ideologically left-wing social norms: participants who held left-wing views were more willing to publish their opinions, and those who were randomly made aware of the prospect of publication reported less conservative views. A priming information treatment, in which participants were informed about cancel culture and the potential negative backlash from social media posts, induced some conformity and silencing, but the results were generally weak and not statistically significant. Finally, a social information treatment, which informed respondents about high rates of others' willingness to speak up, significantly decreased self-censorship. We use our theoretical model, and empirical estimates from the experiment about the value of "speaking up", to analyze potential welfare implications. The analysis reveals that social norms which restrict freedom of expression may enhance social welfare.

**JEL Codes:** D83, P16, C90, Z13
**Keywords:** social media; spiral of silence; public opinion; cancel culture; free speech

*We thank Martin Bisgaard, Leonardo Bursztyn, Eric Dickson, Jean-Robert Tyran, Wieland Müller, Chris Roth, seminar and conference participants at the Burgundy School of Business, Collegio Carlo Alberto, London School of Economics, Lund University, University College Dublin, University of Stirling, University of Vienna, Wilfrid Laurier University, 2022 APSA meeting, 2022 ESA Special Meeting, UCD Behavioural Science Workshop on "Online Social Influence", 2022 Swiss Society of Economics and Statistics Annual Congress (Fribourg), 2022 ESA European Meeting (Bologna), 2023 Rebecca B. Morton Conference on Experimental Political Science (NYU), and the 2023 IMEBESS Meeting (Lisbon) for productive discussions and helpful suggestions. The experiments described were approved by the ethics committees at UCD (HS-E-21-43-Samahita, HS-22-50-Samahita) and WLU (8354). Funding from UCD, the Collegio Carlo Alberto and LCERPA (WLU) is gratefully acknowledged. All errors are our own.

†Department of Economics, Lazaridis School of Business and Economics, Wilfrid Laurier University. E-mail: jmorales@wlu.ca.

‡School of Economics and Geary Institute for Public Policy, University College Dublin. E-mail: margaret.samahita@ucd.ie.

# 1 Introduction

We often express public opinions which differ from our private views (Kuran, 1997). Other times, we may prefer to refrain from expressing our views altogether (Noelle-Neumann, 1974). Social norms, image concerns, and stigma are all factors which affect the public expression of opinions. With increased political polarization and the rise of social media, many prominent voices have recently argued that these social norms have become too strict and that fear of social backlash has led to a stifling of freedom of expression. In recent surveys, sixty-two percent of Americans agreed that "the political climate prevented them from saying things they believe because others might find them offensive" (Cato), and fifty-seven percent of UK residents reported that sometimes they stop themselves from "expressing their views on political and/or social issues because of fear of judgement or negative responses from others" (YouGov). These sentiments were echoed by a letter co-signed by numerous public figures, which argued that "the free exchange of information and ideas, the lifeblood of a liberal society, is daily becoming more constricted" (Harpers).

This paper studies how perceived social pressure affects the public expression of opinion. We first outline a simple model that formalizes our ideas and helps to frame the study. Individuals weigh their own private opinion against public norms. As social pressure increases and these norms become more strict, public opinion is affected both through a shift in publicly stated views towards a norm (*conformity*) and by inducing self-censorship (*silence*). Individuals trade-off espousing public views which are consistent with their true private opinions, while facing the cost of social stigma if these views differ from a socially accepted norm. Social pressure is then particularly effectual in quieting individuals with dissenting or "politically incorrect" views (further from the accepted norm). Furthermore, others' willingness to speak up also affects the utility of voicing ones' views by both providing a signal of the strictness of the norm, while potentially increasing coordination and/or free-riding among those with dissenting opinions.

We then conduct two pre-registered online experiments (total N=3,152) to study the relationship between social pressure and the public expression of opinions. In our experiments with US participants, we elicit their views on potentially contentious statements in the context of two often divisive topics: support for the participation of trans women in competitive sports (gender), and opinions about whether other people are too sensitive regarding issues of race (race). Both of these are highly polarized issues in the US and frequent sources of political conflict. Then, in an incentivized manner, we elicit respondents' willingness to "speak up": to share their stated views on social media. In particular, we ask participants whether they would be willing to let us post their opinion, and their

name, on a social media page dedicated to the experiment.

We also elicit participants' perceived norm by asking what they think the average opinion is to the stated questions, and what they think the average opinion is *among those participants willing to publish their opinion.* Both of these questions are incentivized. At the end of the survey, we inquire about subjects' concern about the political climate and freedom of speech directly, with questions such as "Are you worried about losing your job or missing out on job opportunities if your political opinions become known?" and "How often do you think social pressure causes people to misrepresent or lie about their political opinions on social media?".

Based on our model of public opinion expression, we define the social *norm* to be that where the largest majority of individuals are willing to publish their views. Our surveys reveal that these norms coincide with the liberal/progressive views on both the race and gender topic: the proportion of individuals willing to post their views online is highest among those who held left-wing opinions. These patterns are consistent with self-censorship among those who hold relatively more conservative views, who on average are less willing to post their opinions in our studies. This divergence in willingness to post relatively more conservative views is starkest among Democratic and Independent voters (and weakest for Republicans). In contrast, Republicans express the most concern about freedom of speech being restricted, as elicited in our post-survey questions. Furthermore, Independents held the highest perceptions of public views being distorted by social concerns, measured by the gap between what they believe the average opinion is, relative to that of those willing to publish their views. These descriptive findings contribute important and new stylized facts about the public expression of opinion within the currently polarized US political environment.

Our experimental interventions provide further insights into how social norms affect the public expression of opinions. In a first intervention, participants were made *aware* of the possibility of publishing their stated views on social media as the views were elicited (**Awareness** treatment). This intervention allows us to evaluate whether the prospect of their opinion being made public leads to changes in the stated views' of the respondents, and in particular, whether they conform to a more socially accepted view. Consistent with the descriptive findings of a liberal norm, participants in our Awareness treatment report on average views which are more "left-wing". This inclination to conform is driven mainly by the behaviour of Independent voters.

In our main treatment, we exposed participants to a *priming* text informing them about "cancel culture" and to examples of individuals who lost their jobs due to negative backlash over something they posted on social media (**Prime** treatment). Priming partic-

ipants to consider cancel culture and negative online backlash led to a modest (but often statistically insignificant) reduction in willingness to speak up. Independent voters are also the most sensitive to the backlash prime in a way that is consistent with the theoretical predictions: those further from the norm are most likely to self-censor as a result of the treatment. Interestingly, we also observe no conformity in the publication Awareness treatment for individuals exposed to the backlash prime.

Finally, in a third intervention, we informed participants about others' willingness to publicly express their opinions (**High/Low Peer Expression** treatments). In particular, we inform participants that for a previous group of participants "X% of them were willing to publish their opinion on the above statement with their name". We vary $X$ to be high or low. Our results reveal increased willingness to speak up when a high share of others are also speaking up. However, once again only for individuals not exposed to the backlash prime. One interpretation of these findings is that heightened attention to online backlash may provide a rationale for individuals to express dissenting views privately, while relying on others to do so publicly.

We explore the potential welfare implications of social norms that regulate speech, first by extending our baseline model to consider a setting of policy decision-making and public discourse in which public opinions are aggregated to choose a policy. Our model highlights how social pressure can improve decision making by imposing a cost on "speaking up". This cost provides a selection mechanism through which individuals with more intense preferences receive higher weight in policy decisions, and can therefore lead to improved welfare over a world of complete free speech.

To study these welfare implications empirically, we use data from our experiment which asked participants about the importance they place on the race or gender issues discussed in the survey. Two important empirical facts arise from the analysis: i) individuals are more likely to publish their views when they place higher value on these issues, and ii) individuals who hold liberal views on these topics place higher value on them. Together, these facts imply that liberal social norms which moderate public opinion can enhance policymaking and welfare, which we illustrate with direct welfare calculations using our incentivized measures of willingness to pay for public posts. At the same time, we highlight the policy trade-offs inherent in these choices: while aggregate welfare increases by restricting free speech, the welfare of Republicans decreases.

In the online appendix we examine these issues in a real social media environment. We collect data from a random sample of Twitter users who received either negative or positive comments in response to tweets they published about race or gender issues. Using a triple-difference research design, we then examine whether online social backlash,

in the form of negative comments, induces self-censorship. We find no significant changes in social media behaviour in the days after users experienced social backlash.

Our paper contributes to the literature on social conformity. Social image and reputational concerns may prevent individuals from voicing their opinion for fear of social stigma (Bursztyn and Jensen, 2017). One particular way in which individuals adapt to these concerns is by misreporting their true views (Kuran, 1997; Morris, 2001; Carlson and Settle, 2016), and such misreporting could lead to aggregate information loss. For example, Bursztyn, González and Yanagizawa-Drott (2020) document how men in Saudi Arabia misperceive the level of support for female labor force participation, and alter their behaviour after these beliefs are corrected. Braghieri (2022) documents how college students distort their true attitudes in regards to sensitive topics when answering a survey, and that the extent of this distortion is not properly accounted for by other students (leading to information loss). Our first intervention relates to this issue. Consistent with the results in Braghieri (2022), we show that awareness about opinions being made public leads people to report more liberal views. Our work complements previous related studies by providing evidence from an online survey with a broader range of participants.

Individuals holding dissenting views may also adapt to external pressures by self-censoring, or remaining silent about their opinions (Noelle-Neumann, 1974; Morales, 2020). Studying the determinants of individuals' willingness to speak up is therefore important for our understanding of broader political dynamics and collective decisions. Related to this, Bursztyn et al. (2023) document how rationales which provide social cover for holding potentially stigmatized views increases individuals' willingness to post a tweet expressing dissent. Our main intervention, which primes participants to consider cancel culture and the potential for online social backlash, allows us to study the extent to which this particular narrative affects their willingness to speak up.

Social cues can affect reported attitudes, political donations, voting, and other behaviours (Perez-Truglia and Cruces, 2017; Bursztyn and Jensen, 2017; Conzo ⓡ et al., 2023). In the context of the expression of dissent, observing others' actions provides an important motivation for ones' own voicing of opinions and participation in activism. Others' actions can provide wider signals about dissenting views and may be strategically complementary to participation in the context of collective action (González, 2020; Manacorda and Tesei, 2020), therefore increasing the expression of dissenting or socially stigmatized views. Such social signals can therefore contribute to the unraveling or erosion of social norms (Morales, 2020; Bursztyn, Egorov and Fiorin, 2020; Álvarez-Benjumea, 2023). At the same time, others' participation may also provide opportunities to free-ride and may be a strategic substitute to public dissent (Cantoni et al., 2019; Hager et al., 2023).

Our third intervention relates to this work and investigates how information about others' actions affects public expression in the context of politically contentious views, where social image concerns and the potential for online backlash are present.

A number of recent papers in the social sciences have also studied social backlash in online settings, focusing on the features of social media that can encourage moral outrage and "cancel culture" (Brady, Crockett and Van Bavel, 2020; Brady et al., 2021; Crockett, 2017). Recent work has documented how perceptions of cancel culture are highest for academics who hold minority opinions (Norris, 2023*b*,*a*) but are generally exaggerated in the wider population (Dias, Druckman and Levendusky, 2022). We show that increased awareness of negative backlash on social media increases self-censorship only modestly. Finally, our work contributes to the broad debate regarding political correctness and the value of free speech (Morris, 2001; Braghieri, 2022; Voerman-Tam, Grimes and Watson, 2023). We argue that, under some conditions, social norms which moderate public opinion may be welfare enhancing in the process of policy decision-making.

In the following section we provide our baseline theoretical framework. Section 3 explains the experimental design and our implementation. Section 4 details the results of our survey experiments. Section 5 discusses the welfare implications of social backlash, both theoretically and empirically. Section 6 concludes.

# 2 Conceptual Framework

## 2.1 Model

Assume a continuum of individuals $N$ defined over their private opinions $o_i \in [0,1]$. Consider an individual who decides whether they want to publicly express their opinion on a particular issue. Specifically, individuals choose: i) a public stance $s_i \in [0,1]$, and ii) whether to "speak-up" and voice their stance, $v_i \in \{0,1\}$, given that there are norms that dictate what an appropriate public stance is: $n \in [0,1]$.[1] Individuals get an intrinsic payoff for speaking up plus some idiosyncratic component: $\kappa + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. They maximize their utility:

$$u_i = \begin{cases} -[\alpha(s_i - o_i)^2 + \beta(s_i - n)^2] + \kappa + \varepsilon_i & \text{if } v_i = 1 \\ 0 & \text{if } v_i = 0 \end{cases} \tag{1}$$

[1] In some cases the norm $n$ may vary by individual if, for instance, Democrats and Republicans adhere to different norms or if individuals have their own perceptions of what these norms may be. However, this does not change our main predicted treatment effects. For exposition purposes we keep a common $n$ and for simplicity we treat it as exogenous.

where $\beta$ is the cost or risk from social disapproval and $\alpha$ is a "cognitive dissonance" cost from expressing a public stance which differs from the individual's private opinion.

The maximization problem can be solved by backward induction, first choosing the optimal public stance $s_i^*$, and then whether to speak up or not $v_i^*$. Optimally, individuals choose:

$$s_i^* = \frac{\beta n + \alpha o_i}{\beta + \alpha}$$

The costs of speaking up are minimized when there is no conflict between the private opinion and the social norm, $s_i^* = o_i = n$. When $o_i \neq n$, $s_i^*$ will be a weighted average between the private opinion and social norm, with weights that depend on the social disapproval and cognitive dissonance parameters $\beta$ and $\alpha$.

The utility of expressing the above optimal public stance is

$$u_i(v_i = 1) = \kappa + \varepsilon_i - \frac{\beta\alpha(o_i - n)^2}{\beta + \alpha} \tag{2}$$

Thus, the individual will speak up if

$$(o_i - n)^2 \leq \frac{\beta + \alpha}{\beta\alpha}(\kappa + \varepsilon_i) \tag{3}$$

Define $v^*(o)$ as the share of individuals who find it optimal to speak up out of those holding private opinion $o$. Assuming that $\varepsilon$ is uncorrelated with private opinion $o$, it can be shown that the social norm $n$ is the opinion $o$ whose holders are most vocal:[2]

$$n = \arg\max_o v^*(o)$$

Our experimental interventions aim to manipulate $\beta$, that is, to investigate the impact of an increase in the perceived cost of social disapproval from expressing views which differ from the norm. We think of an increase in $\beta$ as social norms regarding public expression becoming stricter. We consider two main outcomes: individuals' reported views $s$ and their willingness to publicly express those views $v$. In turn, we refer to changes in these actions respectively as increased *conformity* ($s$) and *silence* ($v$).

[2]While we only observe $s_i$ and not $o_i$, note that $o_i = \frac{s_i^*(\beta+\alpha)-\beta n}{\alpha}$, and the threshold equation 3 can be written $(s_i^* - n)^2 \leq \frac{\alpha(\kappa+\varepsilon_i)}{\beta(\alpha+\beta)}$. Thus, for $n \neq \arg\max_s v^*(s)$ information loss has to be extreme—for example, if $\varepsilon$ was highly heterogeneous across $o$.

## 2.2 Hypotheses

Based on this simple framework, we can state three hypotheses related to the public expression of opinion and social norms.

***Hypothesis 1*** *(Conformity):* As social norms become more strict, conformity increases. That is, an increase in $\beta$ reduces the distance between expressed views $s$ and the norm $n$.

To see this, note that the absolute distance between the optimal public stance ($D$) and the norm is:

$$D_i^* = |s_i^* - n| = \frac{\alpha}{\beta + \alpha}|o_i - n|$$

The effect of increasing $\beta$ on individuals' public stance $s_i^*$ is thus to move it closer to the norm $n$ (that is, $\partial D_i^*/\partial\beta < 0$).

***Hypothesis 2*** *(Silence):* As social norms become more strict, silencing increases. That is, an increase in $\beta$ reduces the likelihood that individuals choose to voice their views publicly ($v$).

If individuals choose to speak up, that is, if $v_i = 1$, then:

$$u_i(v_i = 1) = -[\alpha(s_i - o_i)^2 + \beta(s_i - n)^2] + \kappa + \varepsilon_i$$

and this utility decreases as $\beta$ increases (that is, $\partial u_i(v_i = 1)/\partial\beta < 0$). As the payoff of speaking up decreases, it increases the likelihood that individuals cross the threshold ($u_i(v_i = 1) < u_i(v_i = 0)$) at which they would rather remain silent.

Note that an increase in $\beta$ results in a decrease in $u_i(v_i = 1)$ both when i) [*H2a*] $s_i$ is fixed (in our experiment described below, the change in $\beta$ occurs after $s_i$ is reported), and ii) [*H2b*] $s_i$ changes endogenously (in our experiment, when the increase in $\beta$ occurs before the choice of $s_i$).[3]

***Hypothesis 3*** *(Differential silence):* The silencing effect of stricter social norms (as represented by an increase in $\beta$) is more pronounced for individuals whose public stance $s_i$ deviates further from the social norm $n$.

[3]The negative effect is larger in case i) as, intuitively, the individual would have chosen a less conforming $s_i$ and would experience greater social cost from speaking up compared to if their $s_i$ had conformed to the norm as per Hypothesis 1. More details and the proof are given in the Appendix.

Formally, the sensitivity of utility to a change in $\beta$ increases as the absolute difference between $s_i$ and $n$ increases. Note that:

$$\frac{\partial u_i}{\partial \beta} = -(s_i - n)^2 \tag{4}$$

That is, the effect of an increase in $\beta$ on reducing the likelihood of speaking up is stronger for individuals whose public stance $s_i$ deviates more from the social norm $n$.[4] A corollary of this is that stricter norms (an increase in $\beta$) do not affect individuals whose private opinion $o_i$ matches the social norm $n$, since this opinion is not stigmatized (and they would choose $s_i^* = n$).

## 2.3 Social participation

Now consider a simple extension of the model in which individuals observe how many others are publicly expressing their views. We consider this a signal that allows agents to learn about the social disapproval costs ($\beta$), and that could affect the intrinsic reward for speaking up ($\kappa$).

The utility of speaking up is now given by:

$$u_i = -[\alpha(s_i - o_i)^2 + \beta(v_{-i})(s_i - n)^2] + \kappa(v_{-i}) + \varepsilon_i$$

In this context, the social costs $\beta(v_{-i})$ are decreasing in $v_{-i}$, the proportion of others who speak up: when others publicly express their views, individuals can infer that social sanctioning is weak.[5] However, the effect of peer participation $v_{-i}$ on the reward for speaking up $\kappa(v_{-i})$ is ambiguous due to the possibility of both strategic complementarity or substitutability in voicing ones' views, as in the literature on collective action (Cantoni et al., 2019; Hager et al., 2023). Thus, the overall effect of higher peer participation

$$\frac{\partial u_i}{\partial v_{-i}} = -\beta'(v_{-i})(s_i - n)^2 + \kappa'(v_{-i})$$

will depend on whether any strategic substitutability ($\kappa'(v_{-i}) < 0$) is sufficient to undo the positive effect of weaker perception of social sanctioning coming from the first term.

[4]Note that the cross partial derivative of the utility function with respect to $\beta$ and $|s_i - n|$ is therefore negative. In particular, we can consider two cases: (1) If $s_i > n$, the impact of an increase in $\beta$ on the decision to speak up becomes more negative as $s_i$ increases: $\frac{\partial^2 u_i}{\partial \beta \partial s_i} = -2(s_i - n) < 0$. (2) If $s_i < n$, the impact of an increase in $\beta$ on the decision to speak up becomes more negative as $s_i$ *decreases*: $\frac{\partial^2 u_i}{\partial \beta \partial s_i} = -2(n - s_i) < 0$.

[5]In our experimental design, we do not tell individuals *what* others are expressing, so an implicit assumption is that the increase in public expression is broad.

# 3 Experimental Design

## 3.1 Motivation and context

We aimed to study public expression in the context of online social backlash and social image concerns. The US is a large democracy and has become increasingly politically polarized (Boxell, Gentzkow and Shapiro, 2022). At the same time, many Americans report that "the political climate prevented them from saying things they believe because others might find them offensive" (Cato). In particular, norms regarding race and gender issues in the US provide an appropriate context to evaluate our conceptual framework and research questions, as these constitute politically divisive and sensitive topics. Finally, the expression of political opinions on social media is relevant given high rates of social media use among Americans and the potential for real social costs (Bursztyn et al., 2023).

We conducted two experiments which attempt to manipulate the perceived social cost of publicly expressing views online. Our first experiment was conducted in 2021 and our second experiment was conducted in 2023 (the pre-registration and analysis plans for these are available at https://www.socialscienceregistry.org/trials/7905). The experiment timelines are shown in Figures 1 and 2.

## 3.2 Main variables

Our two main variables of interest are reported attitudes and participants' willingness to share their views on social media. As highlighted by our conceptual framework, these are key variables in the context of public expression of opinion (corresponding to $s$ and $v$ in the model).

### 3.2.1 Elicitation of Attitudes

Participants were asked to consider two statements in random order:

- In my opinion, trans women should be allowed to participate in women's sports competitions.
- In my opinion, many people nowadays are too sensitive about things to do with race.

Participants were asked what they think of each statement, choosing from: strongly disagree, disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, or strongly agree (coded as 1-7).[6]

[6]The statements were pre-tested in a pilot together with other questions. Based on our initial analysis we

### 3.2.2 Elicitation of Willingness to Publish

We next asked two questions: "Would you be willing to let us post on social media, anonymously, your response to the previous statement:" (for example)

> *[Participant 37]*
> *"**I somewhat disagree**" that many people nowadays are too sensitive about things to do with race."*

Participants are informed that if they select Yes, we will create a tweet containing the above response and post it on a public Twitter page created once data collection is complete.

We then asked: "Would you be willing to let us post on social media, together with your name, your response to the previous statement:" (for example)

> *[Your name here]*
> *"**I somewhat disagree**" that many people nowadays are too sensitive about things to do with race."*

We inform participants that:

- We will create a tweet containing the above response and may post it on a public Twitter page created once data collection is complete (* see below)
- *We will contact Prolific to request your first and last names. Note that while in general Prolific does not allow researchers to collect personal information, Prolific does encourage researchers to get in touch in cases such as this, where the study design requires the collection of personal data (see https://researcher-help.prolific.co/hc/en-gb/articles/360015378834-Can-I-ask-Participants-for-their-Personal-Information-Identifiers-).
- The tweet will only contain a text of your name without any hyperlink, the public Twitter page will potentially contain the names and opinions of many participants.
- The link to the public Twitter page will be made available to participants who contact the researcher to ask for it, but it will not be otherwise advertised. The public Twitter page will be deleted after 30 days.

Given that we study the effect of perceived social cost on stated opinion, it was crucial that participants seriously considered the possibility that their opinions would be shown

selected these two topics and added "In my opinion" to the start of each statement. See Online Appendix Figures A6 and A7.

to others. However, actual publication with names is not something we wanted to do given the potential for negative consequences for the participants. Additionally, we followed the standard of no deception which is the norm in experimental economics. We therefore truthfully informed participants that, if they said Yes, we would attempt to obtain their names for the purpose of publication, but that publication is conditional on an event that (as we explained in the debrief) had an extremely low chance of happening. Previous inquiries about accessing participants' names from Prolific have, as we expected, been turned down. In sum, this experimental design allows us to measure real behavioural responses to our treatments, while avoiding deception and minimizing potential harms.[7]

## 3.3 Experimental interventions

### 3.3.1 Prime treatment

In our main intervention, participants were shown the following text and image, to *prime* them with the potential risk of social disapproval from online expression of opinions:

> The public nature of social media has resulted in individuals sometimes experiencing negative consequences as a result of their posts, in a phenomenon that some people refer to as **"cancel culture"**.
>
> "those most vulnerable to harm tend to be **individuals previously unknown to the public**, like the communications director who was **fired** in 2013 after posting on social media, from her personal account, **an ill-thought-out joke** about Africa, AIDS and her own white privilege ... or the data analyst who was **fired** last spring after posting on social media, after the death of George Floyd in police custody, a study that suggested that riots depressed rather than increased Democratic Party votes"
>
> These cases highlight the risk of **public backlash from social media.**

[7] As for the publication of the anonymous responses, we published some of these in https://twitter.com/SurveyResponses, but our program to post tweets was eventually blocked by Twitter due to spam-like behaviour (understandably).


The text used was modified from a New York Times article on cancel culture (NYTimes), and highlights some of the potential risks of posting personal opinions on online social media platforms. As an alternative "control" text, participants were shown information about University College Dublin (UCD) or the University of Turin (UniTo), together with the university logo (replacing the cancel culture image).

We included an attention check after the text, asking: "To check that you are paying attention, what does the text say cancel culture can result in?" Participants select from the following alternatives: losing a job, lower voter turnout, or toppling a famous figure, and they cannot proceed unless they select "losing a job". We register whether they correctly answer on their first try. We randomized which of the two control texts was shown and also include an attention check for each text.

As highlighted in Figure 1, in Experiment 1, Treatment 1, the prime text was shown to participants *before* their attitudes regarding the topics are elicited. In Treatment 2, the prime was instead shown *after* the attitude elicitation. Before the attitude elicitation, they were shown one of the two control texts (UCD or UniTo). In Treatment 3, participants saw both control texts, one before and one after the attitude elicitation (randomized order).

**Weak and Strong Prime Treatments**

In Experiment 2 (see Figure 2), we varied the perceived strength of our prime text. As outlined in our pre-registration, we did this to examine and attempt to confirm a preliminary result which suggested that Independent voters responded to the prime text in a manner consistent with a backlash effect: surprisingly, becoming more likely to "speak up". We therefore hoped to investigate whether "cancel culture" narratives provided a rationale for this backlash. We therefore split our prime intervention in a "strong" and a "weak" version. The StrongPrime treatment was identical to our original intervention.

The WeakPrime treatment used the same text while removing bold font, the image, and the text "in a phenomenon that some people refer to as **"cancel culture"**". We again include an attention check after the text, asking: "To check that you are paying attention, what does the text say [cancel culture/social media backlash] can result in?" Participants

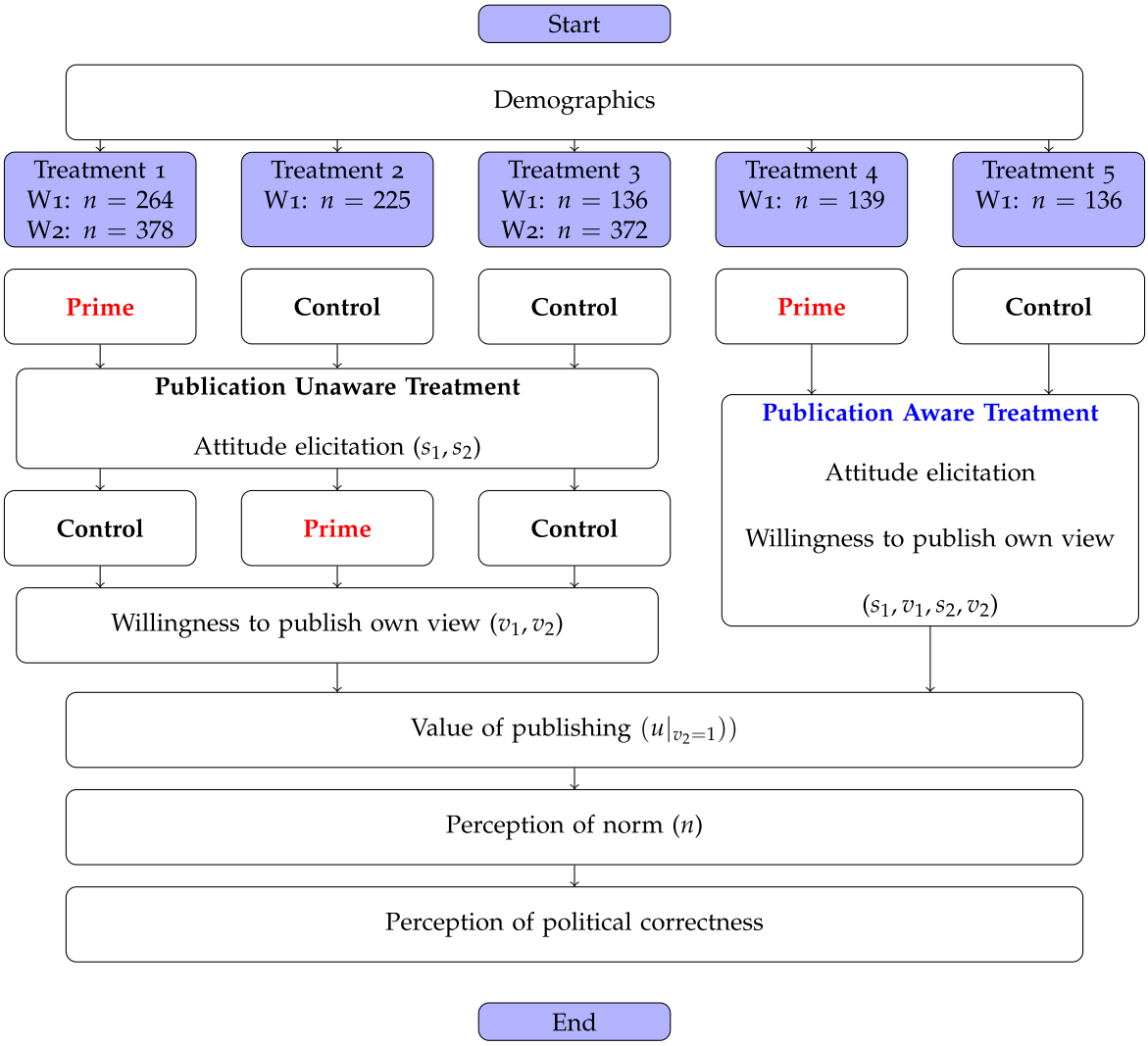



Figure 1: Experiment 1 timeline

select from the following alternatives: losing a job, lower voter turnout, or toppling a famous figure, and they cannot proceed unless they select "losing a job". As before, in the Control treatment, participants were shown a text about University College Dublin (UCD) followed by an attention check.

We found no evidence of backlash in the Strong Prime treatment and the effects of the Strong and Weak primes are not statistically different from each other. We therefore pool both treatments throughout our analyses.

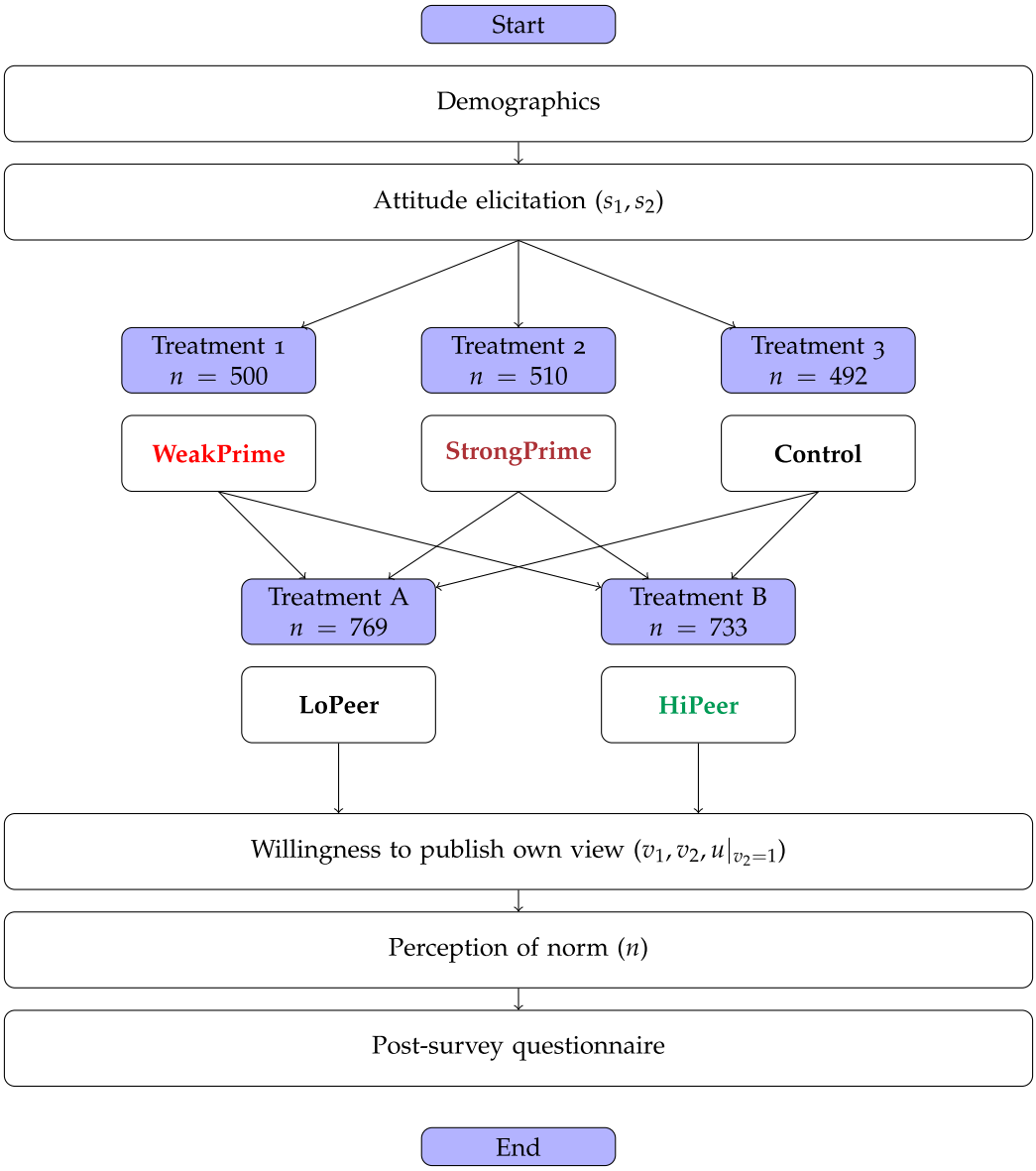



Figure 2: Experiment 2 timeline

### 3.3.2 Publication Awareness treatments

In a second treatment dimension in Experiment 1, we made participants aware of the possibility that their attitudes could be made public. Specifically, in Treatments 1-3, attitudes and willingness to publish own views were elicited at separate stages of the survey (*Publication Unaware* treatment). In contrast, participants in Treatments 4-5 were asked about their views and their willingness to publish these at the same stage (*Publication Aware* treatment). In practice, the two questions were asked sequentially on different survey pages. Participants were asked: i) what they think of the statement, and ii) their willingness to publish their response, anonymously, and then together with their name. This was then repeated for the second statement. In contrast to Treatments 1-3, participants could go back to the previous page, thus allowing them to change their answer to i) after considering their answer to ii). Furthermore, when answering the opinion question on the second topic, they would have been aware of the publication prospect.

We include this treatment branch since participants may perceive different social pressure from revealing their opinion to the experimenter (in T1-T3) rather than the public. Considering the theoretical implications, it is clear that while the model can be solved sequentially (ie. $s^*$ and then $v^*$), splitting these questions along with the treatment allowed us to think about these decisions in turn.[8] This was then followed by the elicitation of the value of publishing for a randomly chosen statement (described below).

To explore the interaction between the two interventions, in addition, in T4, participants saw the online backlash prime before the elicitation of attitude and willingness to publish, while in T5 they saw only one of the two control texts.

### 3.3.3 Hi/Low Peer Participation treatments

In a second intervention of Experiment 2, we modified information given to participants about how likely previous participants were to express their opinions. We specifically varied the percentage that was shown to the participants, which represented the proportion of earlier participants who were open to sharing their opinion on the race or gender statements. Once participants had shared their reported attitudes, and prior to asking about their willingness to publish these views, we displayed the following text:

> When we asked a previous group of participants, X% of them were willing to publish their opinion on the above statement with their name.

where X is determined by randomization into either the "LoPeer" or "HiPeer" treatments. In LoPeer, X equals [13/7] for the [gender/race] statement. In HiPeer, X equals [60/67]

[8] We elaborate on this point, including the potential hypotheses, in Appendix Section A.2.

for the [gender/race] treatment. These figures are taken from our previous surveys, selecting US states with similarly high or low proportion of participants willing to speak up. Participants are debriefed with this information at the end of the experiment.[9]

As outlined in the theoretical framework, within a context of collective action in which social sanctioning and image concerns are present, "speaking up" can be a costly action for which there are incentives to free ride. On the other hand, there may be coordination benefits by which the rewards of voicing ones' views increase with others' participation. Furthermore, learning that many others are speaking up may also lead to individuals inferring that the social costs of public expression are low, further reducing self-censorship. The overall effect is ambiguous and it is an open empirical question which we attempt to answer in our experiment.

## 3.4 Additional questions

### 3.4.1 Demographics questionnaire

In both experiments we collected data on participants' demographics, risk attitude, political preferences and social media use. Since we hypothesized that treatment effects may be heterogeneous along the latter two dimensions, we elicited these variables pre-treatment (Montgomery, Nyhan and Torres, 2018). To elicit political preference, we asked: "In political matters, people talk of 'the left' and 'the right'. How would you place your views on this scale, generally speaking?" and "Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else?" To elicit social media usage, we ask participants how much time per day they spend on Facebook, Twitter, Instagram and other social media platforms.

For Experiment 2, we also included questions for Hong psychological reactance scale (Hong and Faedda, 1996), how common participants think their first and last name are (each on a 7-point scale, 1 Extremely uncommon - 7 Extremely common), and how likely they think the opinion of those willing to publish would indeed be posted on social media (1 Extremely unlikely - 7 Extremely likely). Finally participants could give feedback on the study using a text box.

[9]For LoPeer, [13%/7%] of [Maryland/Indiana] participants were willing to publish their opinion about the [gender/race] statement. For HiPeer, [60%/67%] of [Arizona/Oregon] participants were willing to publish their opinion about the [gender/race] statement.

### 3.4.2 Value of publishing and issue importance

We also elicited a measure of participants' willingness to pay to publish their stated view with their name. Due to time constraint, we only did this for one of the two statements. We therefore randomized participants into either a Race or Gender condition, which determined which of the two statements was used for the value of publishing.

To incentivize these responses, we endowed all participants with 10 tickets for a USD 100 bonus lottery. Participants were informed at the start that their chance of winning was approximately 1 in 1000, 750 or 2000 for waves 1, 2, or 3 respectively. After participants were asked the question on willingness to publish with their name, if they selected "Yes, I would like to", they were then asked whether, in exchange for not posting, they would be willing to give up 10, 5, or 1 of their lottery tickets. These questions were asked sequentially starting from the highest value. If/when they select Yes, they moved on to the next section. If they did not ever select Yes, we code their willingness to pay (WTP) for publication as 0.

If participants stated No to the question about publication with their name, they were then asked whether they would change their mind in exchange for a higher chance of winning the USD 100 lottery. We asked if they would be willing to let us post their response if we give them 1, 5, 20, or 50 additional lottery ticket. These questions were asked sequentially starting from the lowest value. If/when they selected Yes, they moved on to the next section. If they did not ever select Yes, we coded their willingness to pay (WTP) for publication as -100 for the empirical analysis below.

Next, all participants were asked about the importance of the issue discussed in each of the two statements, they respond on a scale from 1 (Not important at all) to 5 (Extremely important).

These questions are important for our analysis about the welfare implications of social norms in regards to public expression of views, which we discuss toward the end of the paper.

Finally, in Experiment 2 we asked participants about how much knowledge they have in the topic area (1 I have little to no knowledge to 5 I am an expert in this topic area).

## 3.5 Beliefs about others' views

We also elicited beliefs about others' stated opinions. We did this for one of the two statements depending on whether participants were in the Gender or Race condition. After showing the statement, we asked participants:

- Considering ALL participants (in this US-based survey), what do you think **the**

**average** opinion is?

- Considering those participants (in this US-based survey) who stated that they WOULD be willing to let us post their opinion, together with their name, on social media (without any additional payment), what do you think **the average** opinion is?[10]

Since those whose opinions are closest to the norm are more willing to speak up, we can loosely approximate what the perceived norm is among our participants with this question. In particular, if participants report that the opinion of all participants is to the right of the opinion of participants who are willing to post their views, then their perception is that left-wing views are more publicly accepted. Furthermore, these questions allow us to measure the extent to which participants think that social norms distort public opinions on these topics. We incentivized participants by rewarding each correct answer with 5 additional lottery tickets for the USD 100 bonus.

## 3.6 Concerns about political climate and freedom of speech

Finally, we asked participants five questions regarding their views on the political climate regarding online freedom of speech and social pressure. These questions were:

- "How often do you worry that things you post on social media can be misinterpreted?" (7-point scale, Never - Always)
- "The political climate these days prevents me from saying things I believe because others might find them offensive." (7-point scale, Strongly disagree - Strongly agree)
- "Are you worried about losing your job or missing out on job opportunities if your political opinions become known?" (4-point scale, Not at all worried - Worried a lot).
- "How often do you think social pressure causes people to misrepresent or lie about their political opinions on social media?" (7-point scale, Never - Always)
- "How often do you think social pressure causes people to refrain or abstain from expressing political opinions on social media?" (7-point scale, Never - Always)

[10] In Experiment 2 we asked about the *majority* opinion instead of the average opinion (Krupka and Weber, 2013), and the results are similar (shown in Appendix Table A12).

## 3.7 Debrief

We end by debriefing participants about the purpose of the study. We inform them that we will create a public Twitter page for the study and post a tweet for each participant's opinion that they are willing to publish anonymously. We explicitly state that we do not anticipate publishing any participant's opinion with their name, even if they stated that they would like us to do this. Regardless, if the participant was willing to publish in exchange for lottery tickets, they would still get these additional tickets and the winner of the lottery would be paid after a few weeks. The full survey is provided in the Appendix.

## 3.8 Implementation

The first wave of data collection was conducted in August 2021, using the data collection platform Prolific. In order to ensure a balanced number of participants across political affiliations, we recruited 300 self-identified Democrats, Republicans and Independents, giving us a total of 900 participants (Wave 1).[11]

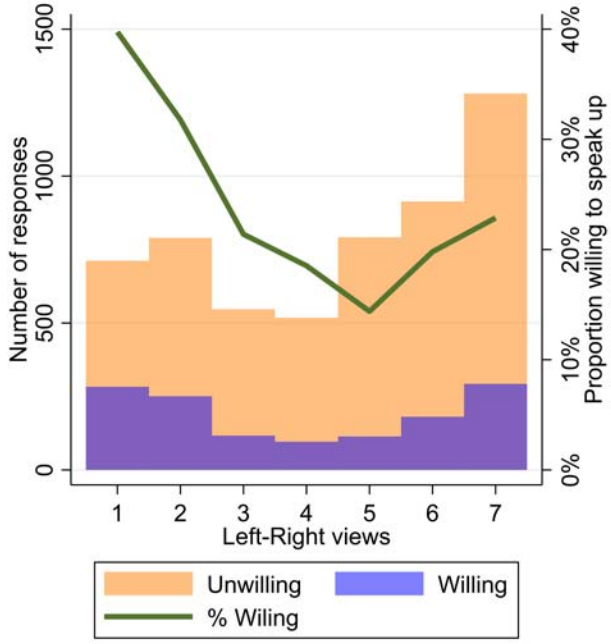
The above data collection oversampled the young female population due to unforeseen circumstances related specifically to sampling within the Prolific platform.[12] To test whether the results were affected by the unusual sample, we ran a second wave in November 2021 using a US nationally representative sample in terms of age, gender and ethnicity, recruiting 750 participants (Wave 2). We followed the original protocol with a few changes: i) due to budget constraints and the reduced sample, we only ran Treatment 1 (primed group) and Treatment 3 (control group), ii) we used only the gender question and dropped the race question, since the former resulted in more distinct norms across partisan groups, and iii) at the end of the experiment we asked respondents about their belief about the likelihood of their answer being published on a 1-7 Likert scale (extremely unlikely to extremely likely).

For Experiment 2, we again used Prolific and recruited 1502 participants, sampling 400 Democrats, 700 Independents (including Unaligned) and 400 Republicans (Wave 3). This survey was conducted in March of 2023. We over-sampled Independents to investigate the previous finding of a backlash effect among this group (which was not confirmed in this wave). We randomized participants evenly into one of the six treatments (3x2 design).

[11] Among Independents we also included those who state their political affiliation to be "None" or "Other". We allocated 27.5% of participants to each of T1-2 and 15% in T3-5 to enable us to test Hypothesis 2b (explained above) with greater power.

[12] https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/, accessed 2021-10-13.

Figure 3: Public expression and reported attitudes




Notes: Agreement to statement in experiments 1 and 2 (pooled) by willingness to publish and proportion willing to publish. Responses to Gender statement are reverse-coded such that liberal views are on the left. "Willing" participants are those who respond Yes to being willing to post their opinions on the social media account.

In a final wave of data collection (Wave 4) which was used as a robustness check, we replicated Experiment 2 but including a new treatment branch (Treatment C) in which we did not inform participants about others' participation at all (ie. as in Experiment 1).

Key descriptive statistics of our sample are provided in Appendix Table A1. As stated above, participants in the first wave are younger and consist of more women and whites compared to the second wave which is nationally representative in terms of age, gender and ethnicity. Participants in the first wave are also more active on social media, with 80% spending at least an hour a day on Facebook, Twitter, Instagram or other social media platforms, compared to 55% in wave 2. Waves 3 and 4 (Experiment 2) look much closer in terms of their demographics to Wave 2, despite not being nationally representative.

Unless otherwise stated, we pool together waves 1 through 3 in all of our results below.

# 4 Results

## 4.1 Distribution of attitudes and willingness to speak up

We begin our presentation of results by providing descriptive evidence of the distribution of attitudes regarding these topics, and individuals' willingness to share their opinions online. Figure 3 shows the distribution of attitudes among our respondents ($s$), split by their willingness to share their views together with their names. We code both questions in a 1-7 scale where 1 represents relatively more left-wing / progressive views. Not surprisingly, these are divisive topics and the distribution of opinions is bimodal.[13]

Overall, 24 percent of participants were willing to publish their views together with their names. Despite the fact that more of the respondents agree with the right-wing position for these topics, the highest share of respondents who are willing to post their views are those who report strongly agreeing with the most liberal views. For those who reported strongly agreeing with the left-most attitude, around 40 percent were willing to post their opinions (in contrast, only 23 percent of participants who held the right-most views were willing to post). Viewed through the lens of our conceptual framework, this suggests that social norms align with liberal views, and that social pressure increases as individuals consider reporting more conservative views.

Next, we present a descriptive analysis correlating willingness to speak up with other demographic characteristics. Table 1 reports these estimates. Left-Right attitudes are coded here in a 0-1 scale for ease of interpretation (as opposed to 1-7). The observation that respondents who hold more conservative views are less willing to speak up remains consistent and statistically significant after controlling for covariates (row 1), including party identifiers. This finding suggests that this relationship is not driven by across-party differences in behaviour regarding public expression. However, this relationship is not statistically significant for Republicans, suggesting that social pressure to adhere to a liberal norm is weakest among this group.

We find that age, employment, risk attitudes and social media use are positively associated with willingness to post, while higher education is negatively associated with it. Women are less willing to publish their views relative to men. In regards to ethnicity, relative to White individuals, Black individuals are more willing to publish their views, while Asian individuals are less willing. We find no statistically significant differences in willingness to post across our experimental waves (coefficients not shown) or across

[13] Appendix Figure A1 shows the distribution split by party affiliation, and Appendix Figure A2 shows the distribution split by topic.

Table 1: Correlates of willingness to publish views online with name

| | (1) All | (2) Dem | (3) Ind | (4) Rep |
|---|---|---|---|---|
| Left-Right opinion scale | -0.169*** | -0.218*** | -0.172*** | -0.069 |
| | (0.021) | (0.035) | (0.033) | (0.045) |
| Age | 0.002*** | -0.001 | 0.002** | 0.004*** |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Non-binary | -0.038 | -0.001 | -0.053 | -0.343*** |
| | (0.058) | (0.111) | (0.068) | (0.049) |
| Female | -0.051*** | -0.085*** | -0.021 | -0.047 |
| | (0.016) | (0.027) | (0.025) | (0.029) |
| Asian or Pacific Islander | -0.062** | -0.072 | -0.060 | -0.092 |
| | (0.027) | (0.044) | (0.043) | (0.058) |
| Black or African American | 0.063** | 0.074* | 0.016 | 0.132* |
| | (0.027) | (0.042) | (0.041) | (0.068) |
| Hispanic or Latino | 0.029 | 0.031 | -0.011 | 0.097 |
| | (0.031) | (0.055) | (0.047) | (0.070) |
| Other | 0.005 | -0.128 | -0.001 | 0.166 |
| | (0.053) | (0.088) | (0.066) | (0.153) |
| College degree | -0.051*** | -0.036 | -0.088*** | 0.001 |
| | (0.016) | (0.030) | (0.024) | (0.028) |
| Employed | 0.045*** | -0.027 | 0.055** | 0.108*** |
| | (0.016) | (0.031) | (0.025) | (0.028) |
| Risk attitude | 0.033*** | 0.036*** | 0.037*** | 0.025*** |
| | (0.003) | (0.006) | (0.006) | (0.006) |
| Democrat | 0.017 | | | |
| | (0.018) | | | |
| Republican | 0.007 | | | |
| | (0.018) | | | |
| Active SM users | 0.048*** | 0.030 | 0.046* | 0.071** |
| | (0.016) | (0.028) | (0.025) | (0.029) |
| Constant | 0.075* | 0.267*** | 0.036 | -0.133* |
| | (0.041) | (0.071) | (0.063) | (0.077) |
| N | 5554 | 1802 | 2233 | 1519 |
| R-sq | 0.066 | 0.077 | 0.072 | 0.081 |

Notes: OLS regressions of willingness to publish with name (0/1) as outcome. *Active SM users* indicates the response to "How much time per day do you spend on social media (Facebook, Twitter, Instagram, Tik Tok, Snapchat, etc)" (never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours) and is coded as 1 if the participant states at least 1 hour. The omitted variables for gender, ethnicity, political affiliation are, respectively, Male, White, and Independents. Fixed effects for survey waves $\times$ topic included (not shown). Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

parties.[14]

## 4.2 Conformity

We first analyze whether participants change their stated views ($s$) in response to our experimental interventions in Experiment 1. This analysis explores Hypothesis 1 from our conceptual framework. We define $s_{iq}$ as respondent $i$'s reported stance (scaled from 0 to 1) for question $q$. We then estimate the following regression:

$$s_{iq} = \theta_0 + \theta_1 Awareness_i + \theta_2 Prime_i + \theta_3 Awareness_i \times Prime_i + \delta_q + \varepsilon_{iq}$$

where $Prime_i$ is a treatment dummy that takes value 1 if subject $i$ sees the prime before the outcome variable (in this case $s$) is elicited (T1 and T4), $Awareness_i$ takes value 1 if $s$ is elicited when subjects are aware of the potential publication (T4 and T5), and $\varepsilon_{iq}$ is an individual-question specific error term. We include topic fixed effects $\delta_q$. In some specification(s) we include a vector of controls $\mathbf{X_i}$ including age, gender, race, education, employment, risk attitude, political leaning, social media use, and wave $\times$ topic fixed effects, which may increase the precision of our estimates (but should be orthogonal to our treatment since it is randomized). In all specifications we use robust standard errors clustered at the individual level (since we have two observations per subject).

Table 2: Conformity in stated views (Experiment 1)

| | All | | Q1 | | Q2 | | All | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Awareness | -0.041 | -0.047** | -0.019 | -0.025 | -0.061* | -0.069** | -0.041 | -0.046** |
| | (0.030) | (0.024) | (0.035) | (0.028) | (0.035) | (0.032) | (0.030) | (0.023) |
| Prime | | | | | | | -0.025 | -0.008 |
| | | | | | | | (0.025) | (0.020) |
| Awareness x Prime | | | | | | | 0.082* | 0.076** |
| | | | | | | | (0.043) | (0.034) |
| N | 994 | 994 | 497 | 497 | 497 | 497 | 1800 | 1800 |
| R-sq | 0.004 | 0.319 | 0.005 | 0.349 | 0.006 | 0.305 | 0.004 | 0.302 |
| Topic FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of Left-Right scale, agreement to topic statement (reverse-coded for Gender statement) and scaled to 0-1. Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, wave FE and its interaction with topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

[14]The patterns look generally similar when we use our incentivized measure of willingness to pay to publish, with the exception that we find statistical overall differences across parties, with Independents being less willing to pay for publication (see Appendix Table A2).

The results are shown in Table 2. In the first two columns, we begin by comparing only individuals who were not exposed to the Prime condition (T2, T3, and T5). We find that individuals who were aware of the possibility of publication reported on average opinions which were 0.047 (out of 1) closer to the left-most view. This result suggests that being aware of the possibility of their views becoming public induces respondents to distort their reported opinion to conform to liberal norms. The effect of the treatment is driven by question 2 (columns 5-6), consistent with the survey structure, since subjects choose $s_i$ after being informed about the possibility of publication.[15] Our finding is also consistent with recent work in Braghieri (2022), showing that public awareness induces conformity in stated opinions towards left-wing views. We also find that this propensity to conform as a result of the Awareness treatment is starkest among Independent voters (see Appendix Table A3) and in Democratic states (see Appendix Table A4).

We then expand the sample to examine the effect of the Prime on stated opinions. Overall, we observe weak effects of the prime on conformity: primed participants state opinions that are on average closer to the norm, but these were not statistically significant. Interestingly, we observe that the interaction between the two treatments also has significant effects which go in the opposite direction. Priming the subjects in the Awareness treatment results in their opinion moving further to the right.

One possible explanation for this result is that raising awareness of cancel culture among those subjects who are already in the mindset of publishing their opinion makes them more sensitive to issues of freedom of speech, thus making them more willing to state their true opinion. The overall effect of being primed and in the Awareness group is not statistically different from zero ($\theta_1 + \theta_3$, rows 1 and 3).

## 4.3 Public expression

We next study individuals' willingness to publish their opinion with their name ($v$). This analysis relates to hypotheses 2 and 3, as well as to the social participation extension, in our conceptual framework. We pool our studies and first estimate regressions of the following form:

$$v_{iq} = \theta_0 + \theta_1 Prime_i + \theta_2 HiPeer_i + \theta_3 Awareness_i + \delta_q + \varepsilon_{iq}$$

where we define $v_{iq}$ as a binary variable which takes value 1 if subject $i$ is willing to publish their opinion on question $q$ with their name (without additional lottery tickets

[15]The information is given at the end of question 1 and subjects are given the opportunity to go back and change their answer. However, only 6 subjects click the "Back" button and we do not observe an effect of publication awareness on responses to the first question.

as incentive). The parameters $\theta_1$, $\theta_2$ and $\theta_3$ capture the average treatment effect of our experimental interventions across all subjects. Since the Awareness treatment was only part of Experiment 1, and the HiPeer treatment was only part of Experiment 2, we include study wave fixed effects in all specifications (not shown). We also include topic/question fixed effects $\delta_q$. In some specifications we include additional demographic controls.

The results of this analysis are shown in Table 3. We find that on average, neither the online backlash Prime nor the Awareness treatment affect individuals' willingness to share their views online. In contrast, the HiPeer treatment, in which we inform respondents that a high share of previous participants published their opinions, has a statistically significant and positive effect on willingness to speak up.[16]

Table 3: Willingness to publish and experimental interventions

| | All | | Dem | | Ind | | Rep | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Prime | -0.022 | -0.014 | -0.032 | -0.019 | -0.013 | -0.012 | -0.020 | -0.013 |
| | (0.016) | (0.015) | (0.028) | (0.027) | (0.025) | (0.025) | (0.029) | (0.028) |
| HiPeer | 0.054*** | 0.041** | 0.058 | 0.063* | 0.038 | 0.028 | 0.077** | 0.057 |
| | (0.020) | (0.020) | (0.039) | (0.038) | (0.030) | (0.030) | (0.038) | (0.037) |
| Awareness | 0.002 | 0.009 | 0.005 | 0.005 | 0.040 | 0.044 | -0.029 | -0.001 |
| | (0.030) | (0.029) | (0.054) | (0.053) | (0.050) | (0.050) | (0.048) | (0.045) |
| N | 5554 | 5554 | 1802 | 1802 | 2233 | 2233 | 1519 | 1519 |
| R-sq | 0.004 | 0.053 | 0.011 | 0.058 | 0.003 | 0.056 | 0.007 | 0.082 |
| Topic + Wave FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of willingness to publish with name (0/1). Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Next, we investigate two empirical extensions guided by our theoretical framework and experimental design. First, we investigate whether the interaction between the main Prime treatment and our other experimental treatments affects willingness to publish (these relate to H2a and H2b). Second, as highlighted by the model, the negative effect of social pressure on public expression increases depending on the distance between individuals' views and the norm (Hypothesis 3). We therefore interact our main Prime treatment

[16] As an alternative outcome variable, we use the WTP to publish as captured by the number of lottery tickets the subject is willing to pay (or accept) to publish. We code these values as WTP of -100, -50, -20, -5, -1, 0, 1, 5, 10. The results are displayed in Table A5. The results are very similar, though less statistically precise, compared to the main results presented above.

with the Left-Right opinion scale. The extended model we analyze is the following:

$$\begin{aligned} v_{iq} = \theta_0 &+ \theta_1 Prime_i + \theta_2 HiPeer_i + \theta_3 Awareness_i \\ &+ \theta_4 Prime_i \times HiPeer_i + \theta_5 Prime_i \times Awareness_i \\ &+ \theta_6 Prime_i \times s_{iq} + \delta_q + \varepsilon_{iq} \end{aligned}$$

where $s_{iq}$ is the Left-Right opinion (scaled from 0 to 1) for individual $i$ for question $q$.

The results are shown in Table 4. While in general the HiPeer treatment makes individuals more likely to be willing to publish their opinion (row 2, $\theta_2 > 0$), this effect is largely attenuated by the online backlash Prime (row 4, $\theta_4 < 0$). The total effect ($\theta_2 + \theta_4$) is not statistically different from zero. This result is consistent with the higher cost of political expression increasing free-riding incentives among individuals who hold dissenting or conservative views. That is, the strategic complementarity induced by others speaking up is neutralized by the strategic substitutability that arises when public expression is perceived as more costly.[17]

We hypothesized (H2a and H2b) that the negative effect of the Prime would be attenuated by the Awareness treatment ($\theta_5 > 0$), in which individuals partly adjust to increased social pressure through conformity (ie. a change in $s$) before choosing whether to speak up or not ($v$). That is, in the awareness treatment $s$ is endogenous to the change in the salience of social backlash. We find that while the coefficient is indeed positive (row 5), the effect is not statistically significant.

We also hypothesized (Hypothesis 3) that the negative effect of the Prime on public expression would increase with the distance to the norm ($\theta_6 < 0$). The direction of the coefficient here is also consistent with the theoretical prediction (row 6), but again, not statistically significant. However, this interaction between the Left-right scale ($s$) and the Prime treatment is statistically significant and negative for Independent voters (columns 5 and 6). This result could be because there is more variation in $s$ for Independent voters (see Figure A1), and they may face higher perceived costs from expressing dissenting views. This finding also suggests that the model we present may be particularly useful when thinking about the behaviour of Independent voters, who we find appear to adhere to left-wing speech norms.[18] The results are also consistent with the view that Independent

[17] As further suggestive evidence of this idea, in Table A6 we interact the HiPeer treatment with the left-right opinion scale ($s$) and split the sample between Prime and Non-Primed individuals. We find that the HiPeer treatment is most effective for respondents who held conservative views but who were not Primed to consider online backlash. Furthermore, the heterogeneity in the effect of the HiPeer treatment for primed participants is starkest in Democratic states, where social pressure and the incentives to free-ride may be higher (see Appendix Table A7).

[18] We replicate this analysis for willingness to pay in Table A8. In addition, since attitudes ($s$) are measured

voters may be particularly concerned with their social image (Klar and Krupnikov, 2016).

Table 4: Willingness to publish and experimental interventions

| | All | | Dem | | Ind | | Rep | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Prime | 0.018 | 0.020 | -0.060 | -0.048 | 0.084 | 0.092* | 0.111 | 0.046 |
| | (0.032) | (0.031) | (0.046) | (0.046) | (0.051) | (0.050) | (0.092) | (0.086) |
| HiPeer | 0.091*** | 0.089*** | 0.048 | 0.063 | 0.099* | 0.105** | 0.138** | 0.120** |
| | (0.034) | (0.033) | (0.060) | (0.058) | (0.052) | (0.052) | (0.064) | (0.060) |
| Awareness | -0.016 | -0.003 | -0.024 | -0.026 | 0.028 | 0.044 | -0.031 | -0.000 |
| | (0.040) | (0.039) | (0.076) | (0.075) | (0.064) | (0.063) | (0.068) | (0.062) |
| Prime x HiPeer | -0.056 | -0.070* | 0.009 | -0.005 | -0.080 | -0.106* | -0.094 | -0.097 |
| | (0.038) | (0.037) | (0.068) | (0.066) | (0.058) | (0.058) | (0.074) | (0.070) |
| Prime x Awareness | 0.044 | 0.031 | 0.047 | 0.049 | 0.054 | 0.033 | 0.017 | 0.008 |
| | (0.052) | (0.050) | (0.097) | (0.096) | (0.088) | (0.087) | (0.085) | (0.078) |
| Prime x LR scale | -0.053 | -0.035 | 0.048 | 0.060 | -0.138** | -0.134** | -0.137 | -0.045 |
| | (0.042) | (0.041) | (0.075) | (0.074) | (0.069) | (0.068) | (0.101) | (0.096) |
| N | 5554 | 5554 | 1802 | 1802 | 2233 | 2233 | 1519 | 1519 |
| R-sq | 0.021 | 0.069 | 0.027 | 0.080 | 0.024 | 0.080 | 0.014 | 0.086 |
| Topic + Wave FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of willingness to publish with name (0/1). Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE × topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

## 4.4 Heteregeneous treatment effects

We explore whether the effect of our main Prime treatment was heterogeneous along various dimensions with specifications of the following form:

$$v_{iq} = \theta_0 + \theta_1 Prime_i + \theta_2 HiPeer_i + \theta_3 Awareness_i + \theta_4 Var_i + \theta_5 Prime_i \times Var_i + \delta_q + \varepsilon_{iq}$$

where $Var_i$ indicates the dimension of interest. In Figure 4 we report the coefficients $\theta_5$ for these interaction models.

The variables we interact with our treatment include the index of psychological reactance (Experiment 2 only), risk attitudes, active social media use, and other demographic

after the experimental interventions in some of the treatment branches, we replicate the analysis excluding all treatments in which reported attitudes are endogenous. The results are very similar and shown in Table A9.

characteristics. All variables are standardised for ease of interpretation. All of these dimensions of heterogeneity are of interest but we were generally agnostic as to the potential direction of the effects (as outlined in our pre-registration).

We find statistically significant heterogeneity along reactance, risk attitudes and age. Participants with higher reactance are more likely to backlash against the Prime treatment, by increasing their willingness to speak up, while those with lower reactance are more likely to self-censor as a result. Similarly, participants who are more willing to take risks increase their willingness to publish, while more risk averse participants become less willing to publish. Finally, older individuals are more likely to self-censor as a result of the prime, while younger individuals become more likely to speak up.[19]

## 4.5 Concerns about political climate and freedom of speech

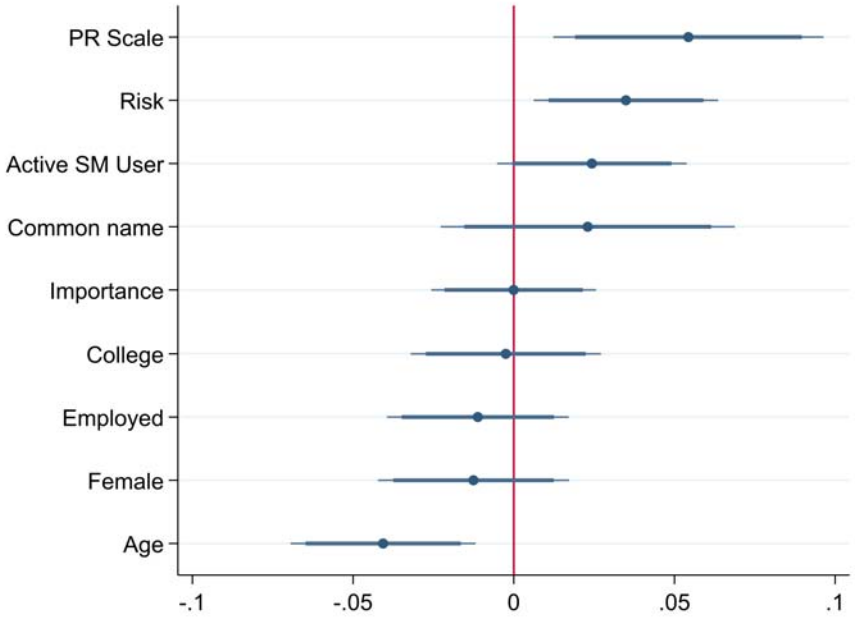
At the end of our survey we included a series of questions regarding participants' views on the current political climate and freedom of speech (Figure 5). Popular media generally report that Republicans are very concerned about issues of cancel culture and censorship. Consistent with this narrative, Republicans in our survey report the highest agreement with the statement "The political climate these days prevents me from saying things I believe because others might find them offensive", are the most worried about job loss due to their political opinion becoming known, and most frequently think social pressure causes people to misrepresent their opinion on social media or to refrain from expressing political opinion completely.

To further our analysis, we create an index of concern regarding freedom of speech using the first principal component of the responses to these questions. Based on this measure, we find that concern is higher for respondents with conservative views (Left-Right scale), Republican affiliation, active social media use, a college degree, of Asian ethnicity, and who are more willing to take risks (Table A10). Finally, we find that, on average, our experimental interventions do not appear to affect reported concern. Interestingly however, our main Prime treatment makes Democrats less concerned about freedom of speech (perhaps due to them believing that the issue is overblown), while making Republicans more concerned (Table A11).

We also measure participants' perception of political correctness using their beliefs about others' views. We create the variable perceived gap: $n_{all} - n_{pub}$ where $n_{all}$ is the participant's belief about the average (majority in Experiment 2) view of all participants in the study wave and $n_{pub}$ is their belief about the view of participants willing to pub-

[19] In Appendix Figure A3 we report the results separately by political party.

Figure 4: Heterogeneity of Prime treatment

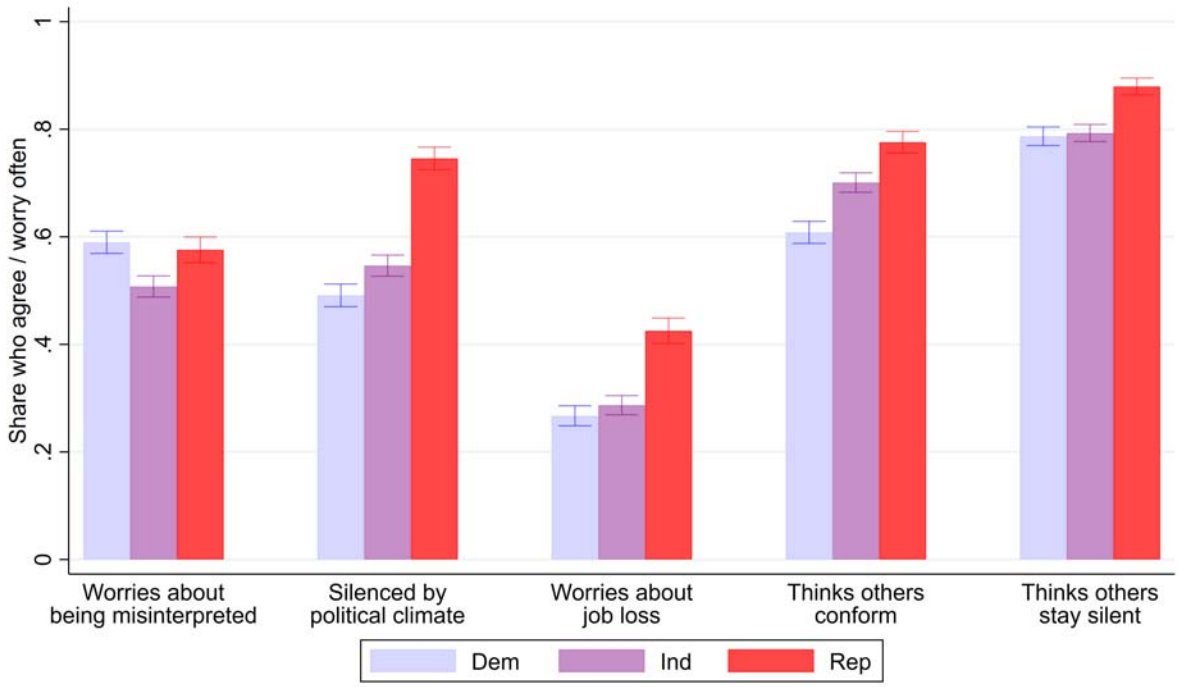



Notes: The figure shows the coefficient $\theta_5$ from the following model:

$$v_{iq} = \theta_0 + \theta_1 Prime_i + \theta_2 HiPeer_i + \theta_3 Awareness_i + \theta_4 Var_i + \theta_5 Prime_i \times Var_i + \delta_q + \varepsilon_{iq}$$

where $Var_i$ indicates the dimension of interest in exploring heterogeneous effects. *PR Scale:* responses to the Hong psychological reactance scale (Hong and Faedda, 1996) (Experiment 2 only). *Risk:* response to "Please tell us, in general, how willing or unwilling you are to take risks." (0 Completely unwilling to take risks - 10 Very willing to take risks). *Active SM user:* response to "How much time per day do you spend on social media (Facebook, Twitter, Instagram, Tik Tok, Snapchat, etc)" (never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours) and is coded as 1 if the participant states at least 1 hour. *Importance Gender/Race:* response to "How important is the issue discussed in the statement to you?" (1 Not important at all - 5 Extremely important). All variables are standardised. Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Robust standard errors are clustered at the individual level.

Figure 5: Perceptions of political climate and freedom of speech online




Notes: Respectively, the questions correspond to: "How often do you worry that things you post on social media can be misinterpreted?"; "The political climate these days prevents me from saying things I believe because others might find them offensive."; "Are you worried about losing your job or missing out on job opportunities if your political opinions become known?"; "How often do you think social pressure causes people to misrepresent or lie about their political opinions on social media?"; "How often do you think social pressure causes people to refrain or abstain from expressing political opinions on social media?".

Table 5: Summary statistics on beliefs about others' views.

| | $n_{all}$ | $n_{pub}$ | $n_{all} - n_{pub}$ |
|---|---|---|---|
| All | 4.390 | 3.788 | 0.602 |
| Dem | 4.130 | 3.687 | 0.443 |
| Ind | 4.475 | 3.727 | 0.749 |
| Rep | 4.604 | 4.013 | 0.590 |

Notes: Belief about views of all participants in the study, all participants willing to publish, and the difference between these two values. In Experiment 1 we elicit beliefs about the *average* views of other participants, in Experiment 2 we elicit beliefs about the *majority* views of other participants.

lish.[20] Summary statistics are shown in Table 5 and separately for each wave and topic in Appendix Table A12. Participants (correctly) perceive the true norm to be less progressive (more to the right) than that expressed by those willing to speak up, but Independents perceive the largest gap compared to the other parties. We also present a descriptive analysis correlating this perceived gap with other demographic characteristics in Table A13. Perceived censorship is greater the more conservative the participant's view. Consistent with the above analysis, Republicans in fact perceive the smallest censorship gap despite expressing the highest concerns about freedom of speech and the political climate.

# 5 Welfare Implications

In this section we include a broader discussion about the welfare implications of strict social norms which regulate speech, using an extension of our theoretical model as a framework, and through additional empirical analyses based on our survey experiments and the data we collected.

## 5.1 Conceptual framework

Consider the following extension of the model. Given a continuum of individuals $N$, each individual $i \in N$ has a true opinion $o_i$ and a publicly expressed opinion $s_i$. For simplicity, consider the case where $o_i \sim U(0,1)$.

The individuals collectively choose a policy $p(\beta_j) \in [0,1]$. Only individuals who choose to publicly express their opinions ("speak up") are taken into account in the policy decision process. In addition, the expressed views ($s_i^*$) may differ from the true opinions due to social pressure. Speaking up in this context acts like a vote, affecting the resulting

[20] We reverse-code the gender question so that 1 corresponds to the progressive view and 7 the conservative view.

policy. Assume the following mean policy decision rule:

$$p(\beta_j) = \int_{o_i \in [0,1]} s_i^*(\beta_j) \cdot v_i(\beta_j)\, do_i$$

where $v_i(\beta_j) \in \{0, 1\}$ indicates whether $i$ has chosen to speak up, and $s_i^*(\beta_j)$ is the expressed opinion of individual $i$, under the level of social pressure $\beta_j$.

For each individual $i$, the resulting utility from a particular policy $p(\beta_j)$ being implemented is defined as:

$$w_i(\beta_j) = -(p(\beta_j) - o_i)^2 \mu_i \tag{5}$$

where $o_i$ is their opinion regarding the issue and $\mu_i$ is a parameter which captures the weight that $i$ gives to the particular issue, that is, how much they care about this policy dimension. We assume this quadratic functional form for tractability. The utility of $i$ decreases as the implemented policy is further from $i$'s ideal point, but specially so if $i$ places higher weight on this policy dimension.

Formally, let $W_{\beta_j} = \int_{o_i \in [0,1]} w_i(\beta_j)\, do_i$ denote the total welfare in society where social pressure is $\beta_j$. We can define an absolute free speech state to be one in which there is no social pressure, such that $\beta_j = 0$, and we denote the welfare associated with absolute free speech as $W_0$. Similarly, we can define the optimal level of social pressure $\beta_j^*$ to be:

$$\beta_j^* = \arg\max_{\beta_j} W_{\beta_j}$$

and the optimal policy as $p^* = p(\beta_j^*)$.

We now illustrate how social pressure may lead to welfare improvements. To illustrate this point, we consider a simple example and highlight two important assumptions.

***Proposition 1*** *(Welfare improvement under social pressure):* Under certain conditions, social norms that regulate speech enhance overall welfare.

If it is the case that $\frac{\partial W}{\partial \beta_j}\Big|_{\beta_j = 0} > 0$, then increasing social pressure from zero (the absolute free speech condition), increases welfare.

To illustrate a simple case with our framework, we need two assumptions:

A1: Individuals who care more about a particular issue get a higher utility from speaking up, and are therefore more likely to do so. That is, $\mu_i$ is positively correlated with $\varepsilon_i$ (the idiosyncratic utility from speaking up in equation 1).[21]

[21] Another way to think about this assumption is that there is an instrumental benefit to speaking up, since individuals influence the policy in the direction of their preferences (as long as social pressure does

A2: Those closer to the norm care more about the issue: $\mu_i$ is correlated with $i's$ bliss point's distance to the norm $|n - o_i|$. Since we have a uniform distribution of citizens in this example, this assumption ensures that the optimal policy $p^*$ lies between the norm and the absolute free speech policy, ie. $n < p^* < p(0)$.

Consider now the following simple functional form for $\mu_i$ which satisfies A2 when $n = 0$:

$$\mu_i = 1 - o_i$$

We can now calculate total welfare under a setting when there is no social pressure, that is, under absolute free speech where $\beta_j = 0$. In this case, everyone chooses $s_i^* = o_i$, and $v_i = 1$. Since there is no social pressure, everyone truthfully voices their opinion.[22] In this case, $p(0) = 0.5$, given our assumption that $o_i \sim U(0, 1)$.

$$W_0 = \int_0^1 -(p(0) - o_i)^2(1 - o_i)\, do_i$$

but at this point, if we were to decrease $p$, welfare would increase:

$$\frac{dW(p)}{dp} = \int_0^1 -2(p - o_i)(1 - o_i)\, do_i$$

Substituting $p = 0.5$ into the equation gives:

$$\left.\frac{dW(p)}{dp}\right|_{p=0.5} = \int_0^1 -2(0.5 - o_i)(1 - o_i)\, do_i = -\frac{1}{6} < 0$$

This implies that the welfare function is decreasing in $p$, and that choosing a policy $p^* < p(0)$ increases welfare.

Furthermore, given our assumptions, we also know that $\frac{\partial p(\beta_j)}{\partial \beta} < 0$. As the strictness of social norms increase, i) $s_i^*$ decreases for all individuals, and ii) individuals who are closer to the norm (0) are more likely to speak up. Both of these forces pull the chosen policy closer to 0. This implies that increasing social pressure from the absolute free speech condition decreases $p$ and increases $W$, such that it is welfare enhancing. In this setting we can also show that the policy which maximizes welfare is $p^* = 1/3$ (so there exists an optimal level of social pressure $\beta^* > 0$).

not distort their $s_i$ too much; one could also consider that $\alpha$ –the cognitive cost from distorting voiced opinions– also increases with $\mu_i$ for similar reasons).

[22] For simplicity, we assume that $\kappa + \varepsilon_i > 0, \forall i$.

## 5.2 Empirical analysis

We now study empirically whether norms that regulate speech may lead to welfare gains in the context of race and gender issues in the US. For this, we use data from our survey on i) individuals' willingness to speak up ($v_i$), ii) individuals' willingness to pay for publication (as a proxy for the utility that individuals get from publication $u_i(v_i = 1)$), iii) individuals' reported attitudes ($s_i$), and iv) individuals' assigned importance to the issues in the questions (as a proxy measure for the weight they place on these particular policy dimensions $\mu_i$).

To evaluate A1, that $\mu_i$ is correlated with $\varepsilon_i$, we start by investigating whether individuals who consider the issues to be of higher importance are more likely to speak up and have a higher willingness to pay for publication. The results of a simple regression analysis studying this relationship are shown in Table 6, columns 1-6. We find that individuals who consider the issues to be more important are more likely to be willing to speak up. Furthermore, this relationship is largely not dependent on individual attitudes (controlling for attitudes only affects it slightly, columns 3 and 6).

To evaluate A2, that $\mu_i$ is higher for individuals who hold opinions closer to the norm, we do a similar exercise using the left-right attitude scale as outcome. The results (Table 6, columns 7-9) show that conservative individuals on average care less about these gender and race issues in the US. This relationship persists even when controlling for individuals' willingness to speak up.

Table 6: Topic importance, willingness to publish and attitudes

| | Publish | | | WTP | | | LR-scale | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Importance (0-1) | 0.272*** | 0.245*** | 0.220*** | 25.436*** | 22.456*** | 18.890*** | -0.226*** | -0.195*** | -0.178*** |
| | (0.021) | (0.021) | (0.021) | (2.878) | (2.865) | (2.898) | (0.018) | (0.015) | (0.015) |
| LR-scale (0-1) | | | -0.131*** | | | -18.532*** | | | |
| | | | (0.021) | | | (2.732) | | | |
| Publish (0/1) | | | | | | | | | -0.070*** |
| | | | | | | | | | (0.011) |
| N | 5554 | 5554 | 5554 | 3152 | 3152 | 3152 | 5554 | 5554 | 5554 |
| R-sq | 0.036 | 0.079 | 0.087 | 0.029 | 0.091 | 0.104 | 0.057 | 0.291 | 0.297 |
| Topic + Wave FE | X | X | X | X | X | X | X | X | X |
| Controls | | X | X | | X | X | | X | X |

Notes: OLS regressions of willingness to publish with name (0/1), willingness to pay for publication, and the Left-right attitude scale (0-1). Importance scale (1-5) is rescaled to 0-1 for ease of interpretation. Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use and and wave FE $\times$ topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

### 5.2.1 Welfare calculation

We now evaluate the welfare implications of increased social pressure directly for participants in our survey data. The exercise is meant as a simple example to illustrate our argument. We first estimate individuals' value of the policy dimension $\mu_i$ using stated willingness to pay for publication and stated importance. WTP captures the utility that individuals receive when "speaking-up", that is, equation 1 when $v_i = 1$. One of the main components of $\varepsilon_i$ in this equation is $\mu_i$, which is directly related to the importance participants place on these issues and largely orthogonal to social pressure (as highlighted in the previous section).[23] To estimate $\mu_i$, we first regress WTP on stated opinion (non-parametrically), responses to the five questions about perceptions of political climate and freedom of speech online, and demographic controls including age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. This first regression captures the components of $u_i$ which are not directly related to $\mu_i$ (including social pressure). The residuals of this first stage are then regressed non-parametrically on stated importance, and the resulting coefficients yield our estimates of $\mu_i$.

Then, we assume the functional form in equation 5 to directly estimate the welfare associated with two possible policy conditions: $p_{all}$, the absolute free speech condition using the average opinion of all participants (ie. $p(0)$); and $p_{pub}$, the restricted speech condition using the average opinion of those participants willing to publish their views.[24] These correspond to: $p_{pub} = 3.92 < 4.40 = p_{all}$. Since we do not observe $o_i$, we use stated opinion $s_i$ (which is a distorted function of $o_i$).

The resulting welfare under each policy is shown in Table 7 (with the unit being lottery tickets in the survey). Implementing $p_{pub}$ in fact leads to a welfare *improvement* relative to $p_{all}$, since those closer to the (progressive) norm place a higher value on the policy issue. It is also possible to calculate the policy that minimises welfare loss under the assumed functional form: the optimal policy is in fact somewhere in between $p_{pub}$ and $p_{all}$, $p = 4.03$. These results suggests that social pressure can have a welfare-enhancing effect with respect to the absolute free speech condition; but they also highlight if social pressure increases too much, it could lead to subsequent overall welfare losses. Finally, the results also outline the winners and losers of strict social norms in this scenario, with both Democrats and Independents experiencing gains, while –perhaps not surprisingly– Republicans suffering losses.[25]

[23] Consider $\varepsilon_i = \mu_i + \epsilon_i$.

[24] As before, we reverse-code the gender question so that 1 corresponds to the progressive view and 7 the conservative view.

[25] We show welfare calculation for all possible policy positions between 1 and 7 in Figure A4 in the Ap-

Table 7: Welfare loss from adopting $p_{all}$ and $p_{pub}$.

| | $p_{all} = 4.40$ | $p_{pub} = 3.92$ |
|---|---|---|
| All | -24.996 | -24.485 |
| Dem | -28.744 | -22.834 |
| Ind | -23.312 | -22.422 |
| Rep | -22.607 | -29.719 |

Notes: Average welfare calculated using equation 5 with two different values of $p$: $p_{all}$, the free speech policy, and $p_{pub}$, the policy chosen through speaking up after self-censorship is taken into account. See text for additional details.

These findings underscore that social norms that regulate speech can act as a selection mechanism through which political discourse incorporates the intensity of preferences of individuals. By imposing a cost on speaking-up, it allows individuals who are potentially more affected by policy outcomes to get a larger say in these collective decisions. This is an often overlooked aspect of the debate surrounding freedom of speech and cancel culture.

# 6 Conclusion

Our paper offers new insights into the public expression of opinion in the context of increasing social and political polarization, contributing to our understanding of how social pressure and online backlash can shape public discourse. The results highlight the fact that a substantial percentage of individuals are indeed hesitant to share their opinions on polarizing subjects due to fear of social backlash. These patterns are more prevalent among those holding more conservative views, and the level of concern about restricted freedom of speech is notably higher among Republicans. Our findings indicate the existence of liberal or left-wing norms which regulate public expression in this setting. We also found that awareness of one's views potentially being made public influenced reported opinions to conform to these perceived liberal social norms.

Priming participants to consider the prospect of negative social backlash had only modest effects on individuals' willingness to speak up. However, Independent voters holding conservative views displayed noticeable increases in self-censorship – suggesting that this group may be particularly sensitive to these norms of expression. Finally, our results indicate that information about others' actions can significantly increase willingness

pendix. Appendix Table A14 shows the results of the welfare calculations under alternative assumptions of the functional form and $\mu_i$, and reweighting the sample to match population shares of Democrats, Independents and Republicans.

to express views.

Our welfare analysis also raises questions about the broader implications of online social pressure. Notably, we considered how norms that regulate free speech may, under certain conditions, actually contribute to the enhancement of welfare in policy decision-making. Our results suggest that the prospect of social backlash can act as a selection mechanism, providing more weight to those with intense preferences in the policy decision process, and potentially improving societal outcomes. The analysis highlights the trade-offs of these social norms and point to the existence of "optimal" levels of social pressure.

Whether norms that regulate speech may enhance welfare will largely depend on the context and on whether these norms align with individuals' preferences, weighted by the importance that the individuals place on these policy dimensions (ie. whether $n < p* < p(0)$). Importantly, this condition is unlikely to be met in general, and in particular it will not hold in regimes where autocrats employ strategies to maintain norms that silence anti-government dissent.[26]

By embedding the public expression of opinion in a policy decision environment, we provide a way to think about free speech that is compatible with the tools of economic reasoning and political economy. This framework presents a new perspective on several contentious and often polarizing debates, and opens up several avenues for future work in topics such as: policy decision-making, regulation of social media, minority rights, cancel culture, and political conflict. We hope the analysis of these issues might be seen in a new light, potentially contributing to a more productive dialogue.

[26]See for example Morales (2020); Buckley et al. (2023); Dal, Nisbet and Kamenchuk (2023).

# References


**Álvarez-Benjumea, Amalia.** 2023. "Uncovering hidden opinions: social norms and the expression of xenophobic attitudes." *European Sociological Review*, 39(3): 449–463.

**Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro.** 2022. "Cross-country trends in affective polarization." *Review of Economics and Statistics*, 1–60.

**Brady, William J., Killian McLoughlin, Tuan N. Doan, and Molly J. Crockett.** 2021. "How social learning amplifies moral outrage expression in online social networks." *Science Advances*, 7(33): eabe5641.

**Brady, William J, Molly J Crockett, and Jay J Van Bavel.** 2020. "The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online." *Perspectives on Psychological Science*, 15(4): 978–1010.

**Braghieri, Luca.** 2022. "Political Correctness, Social Image, and Information Transmission." Working paper.

**Buckley, Noah, Kyle L Marquardt, Ora John Reuter, and Katerina Tertytchnaya.** 2023. "Endogenous popularity: How perceptions of support affect the popularity of authoritarian regimes." *American Political Science Review*, 1–7.

**Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott.** 2020. "Misperceived social norms: Women working outside the home in Saudi Arabia." *American Economic Review*, 110(10): 2997–3029.

**Bursztyn, Leonardo, and Robert Jensen.** 2017. "Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure." *Annual Review of Economics*, 9: 131–153.

**Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin.** 2020. "From extreme to mainstream: The erosion of social norms." *American Economic Review*, 110(11): 3522–48.

**Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth.** 2023. "Justifying dissent." *The Quarterly Journal of Economics*, 138(3): 1403–1451.

**Cantoni, Davide, David Y Yang, Noam Yuchtman, and Y Jane Zhang.** 2019. "Protests as strategic games: Experimental evidence from Hong Kong's antiauthoritarian movement." *The Quarterly Journal of Economics*, 134(2): 1021–1077.

**Carlson, Taylor N, and Jaime E Settle.** 2016. "Political chameleons: An exploration of conformity in political discussions." *Political Behavior*, 38(4): 817–859.

**Conzo ⓡ, Pierluigi, Laura K Taylor ⓡ, Juan S Morales ⓡ, Margaret Samahita ⓡ, and Andrea Gallice.** 2023. "Can ❤s Change Minds? Social Media Endorsements and Policy Preferences." *Social Media + Society*, 9(2).

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

**Dal, Aysenur, Erik C Nisbet, and Olga Kamenchuk.** 2023. "Signaling silence: Affective and cognitive responses to risks of online activism about corruption in an authoritarian context." *New Media & Society*, 25(3): 646–664.

**Dias, Nicholas C, James N Druckman, and Matthew Levendusky.** 2022. "Speech Norms in Contemporary America: The Realities and Misperceptions of "Cancel Culture"." *Working paper SSRN 4235680*.

**González, Felipe.** 2020. "Collective action in networks: Evidence from the Chilean student movement." *Journal of Public Economics*, 188: 104220.

**Hager, Anselm, Lukas Hensel, Johannes Hermle, and Christopher Roth.** 2023. "Political activists as free riders: Evidence from a natural field experiment." *The Economic Journal*, 133(653): 2068–2084.

**Hong, Sung-Mook, and Salvatora Faedda.** 1996. "Refinement of the Hong psychological reactance scale." *Educational and Psychological Measurement*, 56(1): 173–182.

**Jiménez-Durán, Rafael.** 2021. "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter." Working paper.

**Klar, Samara, and Yanna Krupnikov.** 2016. *Independent politics: How American disdain for parties leads to political inaction.* Cambridge University Press.

**Krupka, Erin L, and Roberto A Weber.** 2013. "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11(3): 495–524.

**Kuran, Timur.** 1997. *Private truths, public lies.* Harvard University Press.

**Manacorda, Marco, and Andrea Tesei.** 2020. "Liberation technology: Mobile phones and political mobilization in Africa." *Econometrica*, 88(2): 533–567.

**Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres.** 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science*, 62(3): 760–775.

**Morales, Juan S.** 2020. "Perceived popularity and online political dissent: Evidence from Twitter in Venezuela." *The International Journal of Press/Politics*, 25(1): 5–27.

**Morris, Stephen.** 2001. "Political correctness." *Journal of Political Economy*, 109(2): 231–265.

**Noelle-Neumann, Elisabeth.** 1974. "The spiral of silence: A theory of public opinion." *Journal of Communication*, 24(2): 43–51.

**Norris, Pippa.** 2023*a*. "Cancel culture: Heterodox self-censorship or the curious case of the dog-which-didn't-bark." Harvard Kennedy School Faculty Research Working Paper Series RWP23-020.

**Norris, Pippa.** 2023*b*. "Cancel culture: Myth or reality?" *Political Studies*, 71(1): 145–174.

**Perez-Truglia, Ricardo, and Guillermo Cruces.** 2017. "Partisan interactions: Evidence from a field experiment in the united states." *Journal of Political Economy*, 125(4): 1208–1243.

**Voerman-Tam, Diana, Arthur Grimes, and Nicholas Watson.** 2023. "The economics of free speech: Subjective wellbeing and empowerment of marginalized citizens." *Journal of Economic Behavior & Organization*, 212: 260–274.

# A Appendix (For Online Publication)

## A.1 Evidence from Twitter

In this section we provide additional evidence on our research question using non-experimental data from Twitter. All tweets were collected using the Twitter API for Academic Research. First, we collected a set of tweets with positive or negative feedback, related to race or gender issues, published in the last two years. In particular we searched for tweets that are replies and that contain one of these phrases: *'you should delete this'*, *'this is a bad take'*, *'you should be ashamed'*, for negative feedback, or *'you are so right'*, *'you win the internet'*, *'underrated tweet'*, for positive feedback, in addition to either *'race'* or *'gender'*. Next, we collected all tweets by the author of the original tweet, that is, the tweets to which these replies referred to.[27] Our final dataset contains more than 7 million tweets.

Next, we estimate changes in Twitter activity in the weeks after receiving negative feedback, relative to the weeks after receiving positive feedback. We do so in a triple-difference framework, specifically, we estimate:

$$\begin{aligned} numTweets_{ut} = \alpha + & \sum_{w=-5,-4,\ldots}^{10} \delta_w weeksSinceComment_t \\ & + \sum_{w=-5,-4,\ldots}^{10} \beta_w weeksSinceComment_t \times negativeFeedback_u + \gamma_u + \gamma_t + \varepsilon_{iut} \end{aligned}$$

for Twitter user $u$, published on day $t$. The event-dummy indicators $weeksSinceComment_t$ are dummy variables counting the weeks to the identified tweets for which users received either negative or positive feedback, and $negativeFeedback_u$ indicates whether user $u$ received positive or negative feedback.[28] We also include user ($\gamma_u$) and time ($\gamma_t$) fixed effects.

Our coefficients of interest $\beta'_w$ are normalized relative to the week before the comments and reported in Figure A5. We observe no significant changes in Twitter activity in the weeks following negative social backlash, consistent with evidence from Jiménez-Durán (2021) who find that being sanctioned on Twitter does not reduce hate speech or Twitter activity.

[27] There were many negative feedback tweets, so we randomly selected 1,000 out of these.

[28] A small set of users which were identified in both the positive and the negative feedback samples are dropped.

## A.2 Theoretical extension

In this section we consider the effect of an increase in social pressure $\beta$ (either through the prime or awareness treatment) on speaking up, distinguishing between two cases depending on whether treatment is implemented before or after the elicitation of stated opinion ($s_i$); that is, whether $s_i$ is endogenous (N) or exogenous (X). Intuitively, treatment effect will be weaker when $s_i$ is endogenous as the individual has had the opportunity to conform through their stated opinion and will be less worried about making this opinion public, compared to the exogenous case when $s_i$ would be closer to the individual's true opinion.

If individuals choose to speak up, that is, if $v_i = 1$, then:

$$u_i(v_i = 1) = \begin{cases} \kappa - \left[\beta^T (s_i^*(\beta^C) - n)^2 + \alpha (s_i^*(\beta^C) - o_i)^2\right] = u^X & \text{if } s_i^* \text{ exogenous} \\ \kappa - \left[\beta^T (s_i^*(\beta^T) - n)^2 + \alpha (s_i^*(\beta^T) - o_i)^2\right] = u^N & \text{if } s_i^* \text{ endogenous} \\ \kappa - \left[\beta^C (s_i^*(\beta^C) - n)^2 + \alpha (s_i^*(\beta^C) - o_i)^2\right] = u^C & \text{if control} \end{cases}$$

where $\beta^T > \beta^C$.

In case X, individuals choose $s$ using $\beta^C$ (before they are shown the prime), and choose $v$ using $\beta^T$ (after the prime is shown). The distance between the norm and their reported stance is then higher compared to N, that is $|s_i^X - n| > |s_i^N - n|$, individuals are more willing to report a dissenting opinion in this first stage. When they are later primed prior to choosing whether to publish their views, they become less willing to do so, $u^X < u^N$ and $v_i^X \leq v_i^N$.

***Hypothesis 2a:*** Individuals are less willing to publish their opinion when they are exposed to the prime after the elicitation of $s$ than if they are primed before.

*Proof.* Given our treatment, we assume that $\beta_T > \beta_C$.[29]

$$\begin{aligned}
\beta_T(\beta_T + \alpha) &> \beta_C(\beta_T + \alpha) \\
\beta_T^2 + \beta_T\alpha &> \beta_C\beta_T + \beta_C\alpha \\
\beta_T(\beta_T + \beta_C + 2\alpha) &> 2\beta_C\beta_T + \beta_T\alpha + \beta_C\alpha \\
\beta_T\alpha[(\beta_T + \beta_C)(\beta_T - \beta_C) + 2\alpha(\beta_T - \beta_C)] &> \alpha[2\beta_C\beta_T(\beta_T - \beta_C) + \alpha(\beta_T + \beta_C)(\beta_T - \beta_C)] \\
\beta_T\alpha(\beta_T^2 + 2\beta_T\alpha - \beta_C^2 - 2\beta_C\alpha) &> 2\beta_C\beta_T^2\alpha + \beta_T^2\alpha^2 - 2\beta_C^2\beta_T\alpha - \beta_C^2\alpha^2 \\
\beta_T\alpha(\beta_T + \alpha)^2 + \beta_C^2(\beta_T + \alpha)^2 &> \beta_T\alpha(\beta_C + \alpha)^2 + \beta_T^2(\beta_C + \alpha)^2 \\
\frac{\beta_T\alpha + \beta_C^2}{(\beta_C + \alpha)^2} &> \frac{\beta_T\alpha + \beta_T^2}{(\beta_T + \alpha)^2} \\
\kappa - (o_i - n)^2\left[\frac{\beta_T\alpha^2 + \beta_C^2\alpha}{(\beta_C + \alpha)^2}\right] &< \kappa - (o_i - n)^2\left[\frac{\beta_T\alpha^2 + \beta_T^2\alpha}{(\beta_T + \alpha)^2}\right] \\
\kappa - \beta_T\left(\frac{\alpha(o_i - n)}{\beta_C + \alpha}\right)^2 - \alpha\left(\frac{\beta_C(o_i - n)}{\beta_C + \alpha}\right)^2 &< \kappa - \beta_T\left(\frac{\alpha(o_i - n)}{\beta_T + \alpha}\right)^2 - \alpha\left(\frac{\beta_T(o_i - n)}{\beta_T + \alpha}\right)^2
\end{aligned}$$

That is, $u^X < u^N$. □

In cases C and N, $\beta$ is the same for both choices ($s$ and $v$). Since $\beta^T > \beta^C$, then $u^N < u^C$ from (2). When social pressure is increased *before* the elicitation of public opinion, we expect that willingness to publish will be lower ($v_i^N \leq v_i^C$).

***Hypothesis 2b:*** Individuals are less willing to publish their opinion when they are exposed to the prime or awareness treatment before the elicitation of $s$ than if they are not primed.

[29] Subscripts are used instead of superscripts for readability.

# B Appendix Tables

Table A1: Summary statistics

| | *Wave 1* | | | | | *Wave 2* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | Mean | SD | Min | Max | N | Mean | SD | Min | Max |
| Age | 900 | 28.48 | 9.45 | 18 | 74 | 750 | 36.36 | 13.43 | 18 | 81 |
| Male | 900 | 0.25 | 0.43 | 0 | 1 | 750 | 0.49 | 0.50 | 0 | 1 |
| White | 900 | 0.77 | 0.42 | 0 | 1 | 750 | 0.72 | 0.45 | 0 | 1 |
| College degree | 900 | 0.65 | 0.48 | 0 | 1 | 750 | 0.68 | 0.47 | 0 | 1 |
| Employed | 900 | 0.72 | 0.45 | 0 | 1 | 750 | 0.75 | 0.43 | 0 | 1 |
| Risk attitude | 900 | 6.35 | 1.84 | 0 | 10 | 750 | 5.72 | 2.14 | 0 | 10 |
| Political ideology | 900 | 4.50 | 2.89 | 0 | 10 | 750 | 3.77 | 2.85 | 0 | 10 |
| Active SM user | 900 | 0.80 | 0.40 | 0 | 1 | 750 | 0.55 | 0.50 | 0 | 1 |
| | *Wave 3* | | | | | *Wave 4* | | | | |
| | N | Mean | SD | Min | Max | N | Mean | SD | Min | Max |
| Age | 1502 | 42.08 | 13.42 | 18 | 85 | 369 | 38.74 | 13.68 | 18 | 85 |
| Male | 1502 | 0.49 | 0.50 | 0 | 1 | 369 | 0.48 | 0.50 | 0 | 1 |
| White | 1502 | 0.79 | 0.41 | 0 | 1 | 369 | 0.70 | 0.46 | 0 | 1 |
| College degree | 1502 | 0.63 | 0.48 | 0 | 1 | 369 | 0.56 | 0.50 | 0 | 1 |
| Employed | 1502 | 0.75 | 0.43 | 0 | 1 | 369 | 0.73 | 0.45 | 0 | 1 |
| Risk attitude | 1502 | 5.11 | 2.43 | 0 | 10 | 369 | 5.14 | 2.35 | 0 | 10 |
| Political ideology | 1502 | 4.47 | 2.94 | 0 | 10 | 369 | 4.33 | 2.08 | 0 | 10 |
| Active SM user | 1502 | 0.47 | 0.50 | 0 | 1 | 369 | 0.47 | 0.50 | 0 | 1 |

Notes: *Political ideology* is the response to "In political matters, people talk of 'the left' and 'the right'. How would you place your views on this scale, generally speaking?" (0-10). *Active SM users* indicates the response to "How much time per day do you spend on social media (Facebook, Twitter, Instagram, Tik Tok, Snapchat, etc)" (never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours) and is coded as 1 if the participant states at least 1 hour.

Table A2: Correlates of willingness to pay for online publication

| | (1) All | (2) Dem | (3) Ind | (4) Rep |
|---|---|---|---|---|
| Left-Right opinion scale | -21.764*** | -23.305*** | -23.110*** | -12.934** |
| | (2.703) | (4.601) | (4.044) | (6.262) |
| Age | -0.050 | -0.203 | -0.044 | 0.092 |
| | (0.070) | (0.130) | (0.110) | (0.127) |
| Non-binary | -0.989 | 10.551 | -6.318 | 42.090*** |
| | (6.369) | (11.625) | (7.655) | (5.882) |
| Female | -6.139*** | -8.025*** | -3.202 | -7.349** |
| | (1.796) | (3.066) | (2.850) | (3.604) |
| Asian or Pacific Islander | -6.357* | -5.061 | -7.738 | -13.223* |
| | (3.303) | (5.226) | (5.065) | (7.341) |
| Black or African American | 4.386 | 3.102 | -1.244 | 21.666*** |
| | (2.988) | (4.488) | (4.570) | (8.265) |
| Hispanic or Latino | -0.927 | -1.663 | -1.868 | -1.504 |
| | (3.514) | (5.950) | (5.469) | (7.827) |
| Other | 8.148 | -8.861 | 10.317 | 23.111 |
| | (5.923) | (11.535) | (7.453) | (15.349) |
| College degree | -7.665*** | -7.706** | -10.066*** | -2.509 |
| | (1.799) | (3.384) | (2.757) | (3.476) |
| Employed | 5.345*** | 3.884 | 3.214 | 10.174*** |
| | (1.911) | (3.511) | (2.958) | (3.679) |
| Risk attitude | 3.899*** | 3.895*** | 4.585*** | 2.987*** |
| | (0.385) | (0.680) | (0.623) | (0.717) |
| Democrat | 5.138** | | | |
| | (2.020) | | | |
| Republican | 3.831* | | | |
| | (2.173) | | | |
| Active SM users | 7.593*** | 6.967** | 7.115** | 8.693** |
| | (1.833) | (3.237) | (2.875) | (3.576) |
| Constant | -77.687*** | -62.838*** | -82.759*** | -85.439*** |
| | (4.958) | (8.434) | (7.635) | (10.441) |
| N | 3152 | 1085 | 1237 | 830 |
| R-sq | 0.092 | 0.095 | 0.106 | 0.092 |

Notes: OLS regressions of willingness to pay as outcome. *Active SM users* indicates the response to "How much time per day do you spend on social media (Facebook, Twitter, Instagram, Tik Tok, Snapchat, etc)" (never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours) and is coded as 1 if the participant states at least 1 hour. The omitted variables for gender, ethnicity, political affiliation are, respectively, Male, White, and Independents. Fixed effects for survey waves $\times$ topic included (not shown). Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Conformity by political affiliation (Experiment 1)

| | Dem | | | | Ind | | | | Rep | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Awareness | -0.037 | -0.051 | -0.037 | -0.035 | -0.059 | -0.093** | -0.059 | -0.077* | -0.024 | -0.022 | -0.024 | -0.024 |
| | (0.038) | (0.035) | (0.038) | (0.034) | (0.045) | (0.046) | (0.045) | (0.045) | (0.037) | (0.036) | (0.037) | (0.036) |
| Prime | | | -0.030 | -0.031 | | | 0.032 | 0.023 | | | -0.015 | -0.010 |
| | | | (0.032) | (0.030) | | | (0.042) | (0.042) | | | (0.030) | (0.031) |
| Awareness x Prime | | | 0.069 | 0.075 | | | 0.081 | 0.105 | | | 0.030 | 0.018 |
| | | | (0.057) | (0.053) | | | (0.068) | (0.068) | | | (0.052) | (0.051) |
| N | 332 | 332 | 634 | 634 | 334 | 334 | 590 | 590 | 328 | 328 | 576 | 576 |
| R-sq | 0.028 | 0.196 | 0.019 | 0.165 | 0.007 | 0.108 | 0.012 | 0.078 | 0.004 | 0.060 | 0.003 | 0.037 |
| Topic FE | X | X | X | X | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X | | X | | X |

Notes: OLS regressions of Left-Right scale, agreement to topic statement (reverse-coded for Gender statement) and scaled to 0-1. Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, wave FE and its interaction with topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Conformity by states (Experiment 1)

| | Dem states | | | | Rep states | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Awareness | -0.095** | -0.052 | -0.095** | -0.049 | -0.001 | -0.037 | -0.001 | -0.038 |
| | (0.045) | (0.039) | (0.045) | (0.038) | (0.038) | (0.030) | (0.038) | (0.030) |
| Prime | | | -0.018 | -0.014 | | | -0.029 | -0.005 |
| | | | (0.038) | (0.031) | | | (0.033) | (0.025) |
| Awareness x Prime | | | 0.072 | 0.070 | | | 0.085 | 0.068 |
| | | | (0.067) | (0.053) | | | (0.055) | (0.044) |
| N | 374 | 374 | 682 | 682 | 620 | 620 | 1118 | 1118 |
| R-sq | 0.028 | 0.304 | 0.020 | 0.290 | 0.000 | 0.340 | 0.006 | 0.322 |
| Topic FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of Left-Right scale, agreement to topic statement (reverse-coded for Gender statement) and scaled to 0-1. Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, wave FE and its interaction with topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Willingness to pay and experimental interventions

| | All | | Dem | | Ind | | Rep | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Prime | -2.963 | -2.016 | -1.371 | -0.327 | -4.128 | -3.718 | -2.670 | -1.906 |
| | (1.828) | (1.782) | (3.124) | (3.070) | (2.909) | (2.857) | (3.545) | (3.420) |
| HiPeer | 4.093* | 2.453 | 3.466 | 3.344 | 2.380 | 1.131 | 7.848* | 5.064 |
| | (2.436) | (2.377) | (4.789) | (4.696) | (3.533) | (3.425) | (4.762) | (4.685) |
| Awareness | -1.161 | -0.447 | -0.951 | -2.240 | 7.276 | 8.090 | -8.852 | -5.561 |
| | (3.574) | (3.467) | (6.228) | (6.144) | (6.134) | (6.012) | (5.961) | (5.624) |
| N | 3152 | 3152 | 1085 | 1085 | 1237 | 1237 | 830 | 830 |
| R-sq | 0.007 | 0.073 | 0.017 | 0.075 | 0.008 | 0.084 | 0.009 | 0.089 |
| Topic + Wave FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of willingness to pay for publication. Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Willingness to publish and experimental interventions

| | Prime Treatment | | | | Control (Not Prime) | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Left-Right opinion scale | -0.177*** | -0.188*** | -0.243*** | -0.246*** | -0.165*** | -0.188*** | -0.142** | -0.165*** |
| | (0.027) | (0.029) | (0.046) | (0.045) | (0.038) | (0.041) | (0.063) | (0.064) |
| HiPeer | 0.021 | -0.000 | -0.034 | -0.064 | -0.036 | -0.027 | -0.098 | -0.092 |
| | (0.043) | (0.043) | (0.067) | (0.063) | (0.062) | (0.063) | (0.093) | (0.093) |
| HiPeer x Left-Right scale | 0.040 | 0.040 | 0.118 | 0.125 | 0.178** | 0.178** | 0.286** | 0.283** |
| | (0.057) | (0.056) | (0.091) | (0.086) | (0.085) | (0.085) | (0.135) | (0.132) |
| N | 3654 | 3654 | 1478 | 1478 | 1900 | 1900 | 755 | 755 |
| R-sq | 0.026 | 0.081 | 0.035 | 0.116 | 0.019 | 0.065 | 0.024 | 0.066 |
| Sample | All | All | Ind | Ind | All | All | Ind | Ind |
| Topic + Wave FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of willingness to publish with name (0/1). Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Willingness to publish and experimental interventions, by states

| | Dem states | | | | Rep states | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Prime | -0.032 | -0.028 | 0.030 | 0.033 | -0.015 | -0.007 | 0.007 | 0.007 |
| | (0.026) | (0.025) | (0.049) | (0.048) | (0.020) | (0.019) | (0.041) | (0.041) |
| HiPeer | 0.051 | 0.040 | 0.137** | 0.130** | 0.056** | 0.042* | 0.061 | 0.060 |
| | (0.033) | (0.033) | (0.057) | (0.055) | (0.025) | (0.025) | (0.041) | (0.041) |
| Awareness | -0.058 | -0.041 | -0.069 | -0.047 | 0.037 | 0.047 | 0.018 | 0.031 |
| | (0.047) | (0.045) | (0.064) | (0.062) | (0.038) | (0.037) | (0.051) | (0.050) |
| Prime x HiPeer | | | -0.123* | -0.129** | | | -0.012 | -0.028 |
| | | | (0.063) | (0.062) | | | (0.047) | (0.047) |
| Prime x Awareness | | | 0.029 | 0.027 | | | 0.053 | 0.035 |
| | | | (0.082) | (0.079) | | | (0.067) | (0.065) |
| Prime x LR scale | | | -0.060 | -0.054 | | | -0.043 | -0.019 |
| | | | (0.066) | (0.066) | | | (0.054) | (0.053) |
| N | 2084 | 2084 | 2084 | 2084 | 3470 | 3470 | 3470 | 3470 |
| R-sq | 0.007 | 0.063 | 0.023 | 0.083 | 0.004 | 0.056 | 0.023 | 0.070 |
| Topic + Wave FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of willingness to publish with name (0/1). Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Sample is split by states based on general election vote shares. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Willingness to pay and experimental interventions

| | All | | Dem | | Ind | | Rep | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Prime | -1.487 | -1.489 | -7.646 | -7.340 | 2.930 | 4.757 | 12.337 | 3.335 |
| | (3.713) | (3.624) | (5.326) | (5.267) | (6.165) | (5.991) | (12.216) | (11.953) |
| HiPeer | 7.346* | 6.875* | 2.282 | 3.153 | 13.070** | 13.925** | 5.790 | 2.674 |
| | (3.897) | (3.855) | (6.863) | (6.546) | (5.959) | (6.057) | (7.692) | (7.477) |
| Awareness | 0.676 | 1.866 | 0.322 | -1.966 | 5.672 | 9.251 | -1.055 | 1.092 |
| | (4.972) | (4.895) | (8.911) | (9.062) | (8.021) | (7.726) | (9.063) | (8.701) |
| Prime x HiPeer | -5.117 | -6.523 | 1.440 | 0.306 | -14.413** | -17.946*** | 2.562 | 3.377 |
| | (4.410) | (4.341) | (7.815) | (7.523) | (6.702) | (6.712) | (8.803) | (8.580) |
| Prime x Awareness | -3.850 | -5.013 | -4.667 | -3.533 | 6.775 | 1.707 | -14.949 | -13.443 |
| | (6.229) | (6.098) | (11.124) | (11.111) | (10.781) | (10.435) | (10.559) | (10.047) |
| Prime x LR scale | -0.065 | 2.199 | 15.570* | 17.707* | -6.625 | -6.871 | -17.690 | -5.815 |
| | (5.120) | (5.021) | (9.337) | (9.191) | (8.582) | (8.405) | (13.734) | (13.584) |
| N | 3152 | 3152 | 1085 | 1085 | 1237 | 1237 | 830 | 830 |
| R-sq | 0.031 | 0.093 | 0.038 | 0.099 | 0.039 | 0.116 | 0.024 | 0.097 |
| Topic + Wave FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of willingness to pay for publication. Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Willingness to publish and experimental interventions (alternative sample, exogenous $s$)

| | All | | Dem | | Ind | | Rep | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Prime | 0.017 | 0.020 | -0.031 | -0.023 | 0.062 | 0.070 | 0.077 | -0.004 |
| | (0.038) | (0.037) | (0.060) | (0.059) | (0.058) | (0.057) | (0.110) | (0.104) |
| HiPeer | 0.090** | 0.090*** | 0.059 | 0.082 | 0.087 | 0.096* | 0.128* | 0.112* |
| | (0.035) | (0.035) | (0.065) | (0.062) | (0.053) | (0.054) | (0.066) | (0.062) |
| Prime x HiPeer | -0.054 | -0.071* | -0.007 | -0.026 | -0.064 | -0.088 | -0.081 | -0.077 |
| | (0.041) | (0.040) | (0.076) | (0.074) | (0.062) | (0.061) | (0.077) | (0.074) |
| Prime x LR scale | -0.055 | -0.034 | 0.019 | 0.046 | -0.129* | -0.126* | -0.111 | -0.006 |
| | (0.047) | (0.047) | (0.083) | (0.083) | (0.077) | (0.076) | (0.120) | (0.114) |
| N | 4098 | 4098 | 1234 | 1234 | 1749 | 1749 | 1115 | 1115 |
| R-sq | 0.020 | 0.067 | 0.030 | 0.088 | 0.018 | 0.079 | 0.016 | 0.078 |
| Topic + Wave FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of willingness to pay for publication. The sample excludes treatment branches in which LR-scale ($s$) is endogenous, namely: Treatments 1, 4 and 5 in Experiment 1 (see Figure 1). Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Correlates of concern about political climate and freedom of speech online

| | (1) All | (2) Dem | (3) Ind | (4) Rep |
|---|---|---|---|---|
| Left-Right opinion scale | 1.019*** | 1.144*** | 1.180*** | 0.301** |
| | (0.074) | (0.120) | (0.117) | (0.153) |
| Willing to publish with name | -0.527*** | -0.421*** | -0.577*** | -0.576*** |
| | (0.057) | (0.092) | (0.097) | (0.104) |
| Age | -0.007*** | 0.002 | -0.007** | -0.014*** |
| | (0.002) | (0.004) | (0.004) | (0.004) |
| Non-binary | 0.088 | -0.502 | 0.414* | -0.690*** |
| | (0.225) | (0.492) | (0.244) | (0.182) |
| Female | 0.058 | -0.032 | 0.145 | 0.091 |
| | (0.057) | (0.100) | (0.094) | (0.109) |
| Asian or Pacific Islander | 0.288** | 0.409** | 0.427*** | -0.365 |
| | (0.113) | (0.167) | (0.164) | (0.340) |
| Black or African American | 0.120 | 0.302** | 0.001 | -0.171 |
| | (0.099) | (0.151) | (0.169) | (0.196) |
| Hispanic or Latino | 0.039 | -0.154 | 0.217 | 0.042 |
| | (0.120) | (0.212) | (0.184) | (0.224) |
| Other | 0.154 | 0.097 | 0.113 | 0.595 |
| | (0.178) | (0.418) | (0.227) | (0.373) |
| College degree | 0.216*** | -0.046 | 0.215** | 0.440*** |
| | (0.057) | (0.103) | (0.090) | (0.107) |
| Employed | 0.044 | -0.016 | 0.082 | 0.083 |
| | (0.064) | (0.108) | (0.103) | (0.126) |
| Risk attitude | 0.051*** | 0.068*** | 0.055** | 0.016 |
| | (0.013) | (0.024) | (0.022) | (0.022) |
| Democrat | -0.116* | | | |
| | (0.064) | | | |
| Republican | 0.264*** | | | |
| | (0.067) | | | |
| Active SM users | 0.321*** | 0.357*** | 0.245*** | 0.406*** |
| | (0.059) | (0.103) | (0.093) | (0.112) |
| Constant | -0.702*** | -0.994*** | -0.804*** | 0.299 |
| | (0.151) | (0.255) | (0.242) | (0.284) |
| N | 5554 | 1802 | 2233 | 1519 |
| R-sq | 0.152 | 0.133 | 0.141 | 0.129 |

Notes: OLS regressions of the first principal component of responses to end-of-survey concern questions: "How often do you worry that things you post on social media can be misinterpreted?"; "The political climate these days prevents me from saying things I believe because others might find them offensive."; "Are you worried about losing your job or missing out on job opportunities if your political opinions become known?"; "How often do you think social pressure causes people to misrepresent or lie about their political opinions on social media?"; "How often do you think social pressure causes people to refrain or abstain from expressing political opinions on social media?". *Active SM users* indicates the response to "How much time per day do you spend on social media (Facebook, Twitter, Instagram, Tik Tok, Snapchat, etc)" (never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours) and is coded as 1 if the participant states at least 1 hour. The omitted variables for gender, ethnicity, political affiliation are, respectively, Male, White, and Independents. Fixed effects for survey waves × topic included (not shown). Robust standard errors in parentheses are clustered at the individual level. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Concerns about freedom of speech and experimental interventions

| | All | | Dem | | Ind | | Rep | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Prime | -0.105 | -0.112 | -0.338** | -0.304** | 0.027 | 0.050 | 0.633** | 0.524* |
| | (0.103) | (0.101) | (0.139) | (0.137) | (0.170) | (0.171) | (0.308) | (0.303) |
| HiPeer | -0.102 | -0.110 | -0.115 | -0.138 | -0.140 | -0.098 | -0.038 | -0.073 |
| | (0.121) | (0.119) | (0.218) | (0.215) | (0.185) | (0.182) | (0.234) | (0.230) |
| Awareness | -0.112 | -0.168 | -0.418 | -0.469* | 0.126 | 0.087 | -0.079 | -0.085 |
| | (0.137) | (0.136) | (0.261) | (0.255) | (0.217) | (0.220) | (0.238) | (0.227) |
| Prime x HiPeer | 0.119 | 0.143 | 0.112 | 0.108 | 0.194 | 0.183 | 0.049 | 0.087 |
| | (0.138) | (0.136) | (0.246) | (0.243) | (0.215) | (0.211) | (0.264) | (0.258) |
| Prime x Awareness | -0.016 | -0.004 | 0.539* | 0.544* | -0.368 | -0.331 | -0.244 | -0.272 |
| | (0.176) | (0.176) | (0.326) | (0.315) | (0.288) | (0.297) | (0.308) | (0.301) |
| Prime x LR scale | 0.046 | 0.063 | 0.199 | 0.176 | -0.249 | -0.270 | -0.575* | -0.450 |
| | (0.141) | (0.139) | (0.243) | (0.235) | (0.235) | (0.233) | (0.346) | (0.341) |
| N | 5554 | 5554 | 1802 | 1802 | 2233 | 2233 | 1519 | 1519 |
| R-sq | 0.099 | 0.134 | 0.092 | 0.125 | 0.090 | 0.121 | 0.050 | 0.112 |
| Topic + Wave FE | X | X | X | X | X | X | X | X |
| Controls | | X | | X | | X | | X |

Notes: OLS regressions of the first principal component of responses to end-of-survey concern questions: "How often do you worry that things you post on social media can be misinterpreted?"; "The political climate these days prevents me from saying things I believe because others might find them offensive."; "Are you worried about losing your job or missing out on job opportunities if your political opinions become known?"; "How often do you think social pressure causes people to misrepresent or lie about their political opinions on social media?"; "How often do you think social pressure causes people to refrain or abstain from expressing political opinions on social media?". Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Summary statistics on beliefs about others' views.

| | $n_{all}$ | $n_{pub}$ | $n_{all} - n_{pub}$ |
|---|---|---|---|
| *Wave 1, gender topic* | | | |
| All | 3.882 | 3.744 | 0.138 |
| Dem | 3.731 | 3.569 | 0.163 |
| Ind | 3.866 | 3.613 | 0.254 |
| Rep | 4.061 | 4.061 | 0.000 |
| *Wave 1, race topic* | | | |
| All | 4.499 | 3.545 | 0.953 |
| Dem | 4.459 | 3.682 | 0.777 |
| Ind | 4.490 | 3.255 | 1.235 |
| Rep | 4.553 | 3.709 | 0.844 |
| *Wave 2, gender topic* | | | |
| All | 4.027 | 3.901 | 0.125 |
| Dem | 3.851 | 3.793 | 0.057 |
| Ind | 4.241 | 3.905 | 0.336 |
| Rep | 4.121 | 4.177 | -0.057 |
| *Wave 3, gender topic* | | | |
| All | 4.868 | 4.050 | 0.819 |
| Dem | 4.641 | 3.887 | 0.754 |
| Ind | 4.805 | 3.986 | 0.819 |
| Rep | 5.208 | 4.325 | 0.883 |
| *Wave 3, race topic* | | | |
| All | 4.517 | 3.590 | 0.926 |
| Dem | 4.205 | 3.400 | 0.805 |
| Ind | 4.546 | 3.595 | 0.951 |
| Rep | 4.779 | 3.775 | 1.005 |

Notes: Belief about views of all participants in the study, all participants willing to publish, and the difference between these two values. In Experiment 1 (Waves 1 and 2) we elicit beliefs about the *average* views of other participants, in Experiment 2 (Wave 3) we elicit beliefs about the *majority* views of other participants.

Table A13: Correlates of perceived censorship

| | (1) All | (2) Dem | (3) Ind | (4) Rep |
|---|---|---|---|---|
| Left-Right opinion scale | 0.358*** | 0.307 | 0.335* | 0.717*** |
| | (0.113) | (0.204) | (0.173) | (0.238) |
| Age | -0.002 | -0.013** | 0.003 | 0.003 |
| | (0.003) | (0.005) | (0.005) | (0.005) |
| Non-binary | 0.189 | 0.207 | 0.239 | -0.619** |
| | (0.305) | (0.498) | (0.378) | (0.244) |
| Female | 0.128* | 0.181 | 0.161 | -0.071 |
| | (0.074) | (0.124) | (0.122) | (0.146) |
| Asian or Pacific Islander | -0.074 | 0.095 | -0.495** | 0.249 |
| | (0.144) | (0.217) | (0.221) | (0.349) |
| Black or African American | -0.194 | -0.434** | 0.150 | -0.376 |
| | (0.120) | (0.185) | (0.191) | (0.288) |
| Hispanic or Latino | -0.106 | -0.079 | -0.284 | 0.246 |
| | (0.143) | (0.211) | (0.231) | (0.360) |
| Other | 0.117 | 0.931 | -0.026 | -0.458 |
| | (0.255) | (0.576) | (0.323) | (0.489) |
| College degree | 0.016 | 0.014 | 0.068 | -0.025 |
| | (0.076) | (0.134) | (0.119) | (0.150) |
| Employed | -0.025 | -0.025 | -0.115 | 0.105 |
| | (0.082) | (0.142) | (0.127) | (0.164) |
| Risk attitude | -0.030* | -0.031 | -0.029 | -0.025 |
| | (0.016) | (0.027) | (0.027) | (0.032) |
| Democrat | -0.115 | | | |
| | (0.084) | | | |
| Republican | -0.220** | | | |
| | (0.090) | | | |
| Active SM users | -0.147* | -0.112 | -0.193 | -0.170 |
| | (0.077) | (0.133) | (0.124) | (0.151) |
| Constant | 0.372* | 0.631* | 0.282 | -0.371 |
| | (0.206) | (0.342) | (0.328) | (0.408) |
| N | 3152 | 1085 | 1237 | 830 |
| R-sq | 0.045 | 0.051 | 0.043 | 0.071 |

Notes: OLS regressions of perceived censorship, measured as $n_{all} - n_{pub}$ where $n_{all}$ is belief about average (majority in Experiment 2) view of all participants in the study and $n_{pub}$ is belief about those willing to publish. *Active SM users* indicates the response to "How much time per day do you spend on social media (Facebook, Twitter, Instagram, Tik Tok, Snapchat, etc)" (never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours) and is coded as 1 if the participant states at least 1 hour. The omitted variables for gender, ethnicity, political affiliation are, respectively, Male, White, and Independents. Fixed effects for survey waves $\times$ topic included (not shown). Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

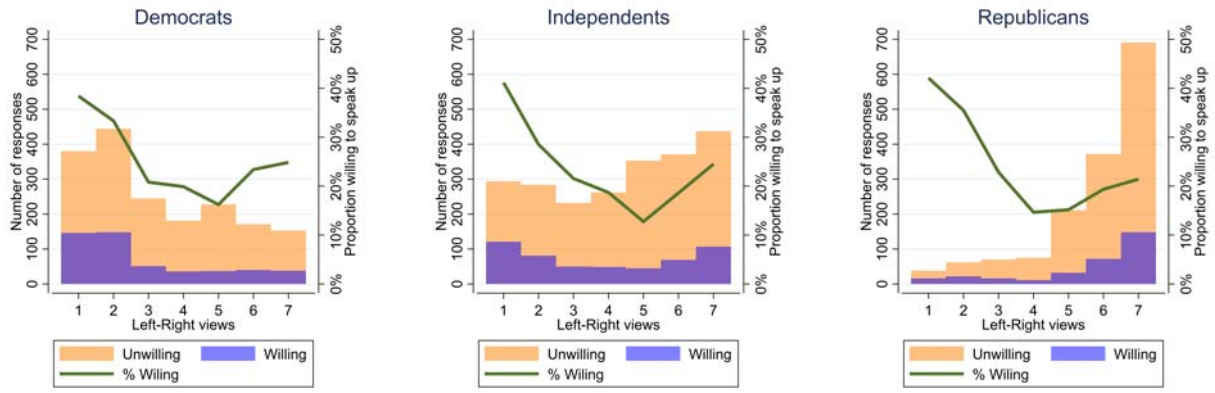
Table A14: Welfare loss from adopting $p_{all}$ and $p_{pub}$.

| | $p_{all} = 4.40$ | $p_{pub} = 3.92$ |
|---|---|---|
| *a) Stated importance* | | |
| All | -16.558 | -16.810 |
| Dem | -18.874 | -15.330 |
| Ind | -15.592 | -15.509 |
| Rep | -15.230 | -20.478 |
| *b) Implied importance,* $w_i = -\|p - o_i\|\mu_i$ | | |
| All | -9.530 | -9.465 |
| Dem | -10.424 | -9.286 |
| Ind | -8.838 | -8.724 |
| Rep | -9.390 | -10.802 |
| *c) Stated importance,* $w_i = -\|p - o_i\|\mu_i$ | | |
| All | -6.656 | -6.730 |
| Dem | -7.207 | -6.495 |
| Ind | -6.270 | -6.282 |
| Rep | -6.568 | -7.666 |
| | Weighted sample | |
| | $p_{all} = 4.53$ | $p_{pub} = 4.01$ |
| *d) Implied importance* | | |
| All | -23.393 | -22.710 |
| Dem | -24.683 | -19.112 |
| Ind | -24.138 | -22.612 |
| Rep | -20.598 | -27.562 |
| *e) Stated importance* | | |
| All | -16.764 | -16.912 |
| Dem | -17.507 | -13.843 |
| Ind | -17.381 | -16.886 |
| Rep | -14.974 | -20.590 |

Notes: Average welfare calculated using equation 5 with the following input: $p \in [1, 7]$, stated opinion $s_i$ (reverse-coded for the gender question), stated or implied importance of issue. Implied importance is calculated as follows: we regress WTP on stated opinion, responses to the five questions about perceptions of political climate and freedom of speech online and controls including age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. The residuals are then regressed on stated importance. The predicted values yield our estimate of implied importance, re-scaling it by subtracting the constant. Panels d) and e) reweigh the sample using the following proportions: 30% Democrats, 43% Independents, and 28% Republicans (source: https://news.gallup.com/poll/388781/political-party-preferences-shifted-greatly-during-2021.aspx).

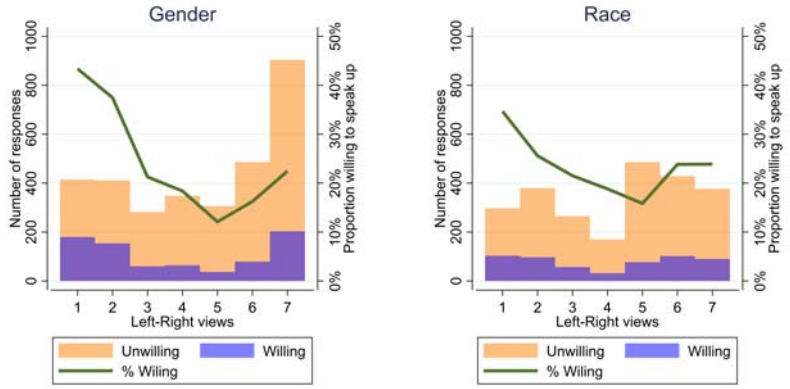
# C Appendix Figures

Figure A1: Public expression and attitudes across parties



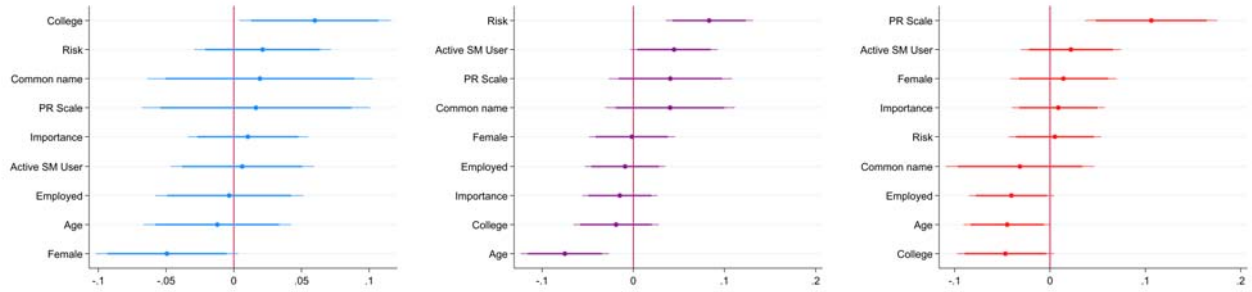

Notes: Agreement to statement in experiments 1 and 2 (pooled) by willingness to publish and proportion willing to publish. Responses to Gender statement are reverse-coded. Figures show attitudes and public expression separately by party: Democrats (left), Independents (middle) and Republicans (right).

Figure A2: Public expression and attitudes across topics




Notes: Agreement to statement in experiments 1 and 2 (pooled) by willingness to publish and proportion willing to publish. Responses to Gender statement are reverse-coded. Figures show attitudes and public expression separately by topic: Gender (left) and Race (right).

Figure A3: Heterogeneity of Prime treatment by party

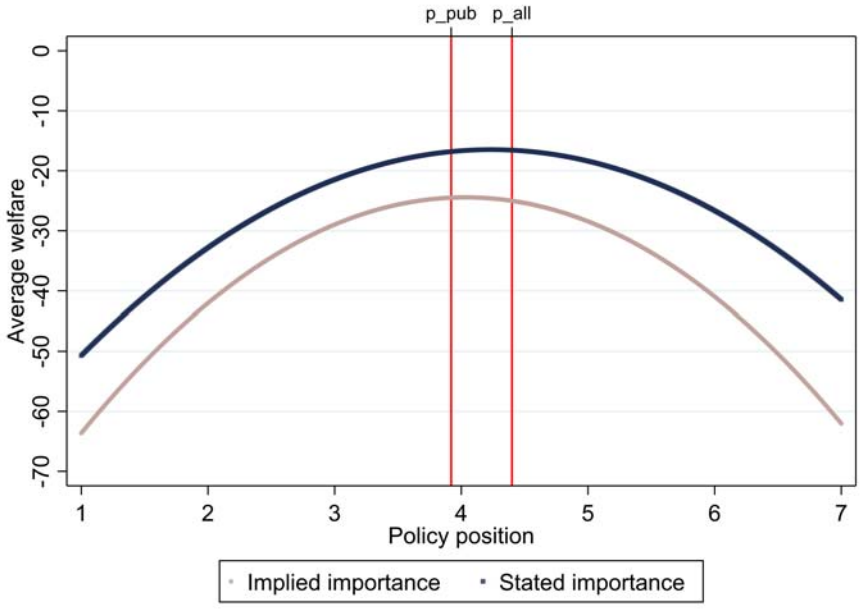



Notes: The figure shows the coefficient $\theta_5$ from the following model:

$$v_{iq} = \theta_0 + \theta_1 Prime_i + \theta_2 HiPeer_i + \theta_3 Awareness_i + \theta_4 Var_i + \theta_5 Prime_i \times Var_i + \delta_q + \varepsilon_{iq}$$

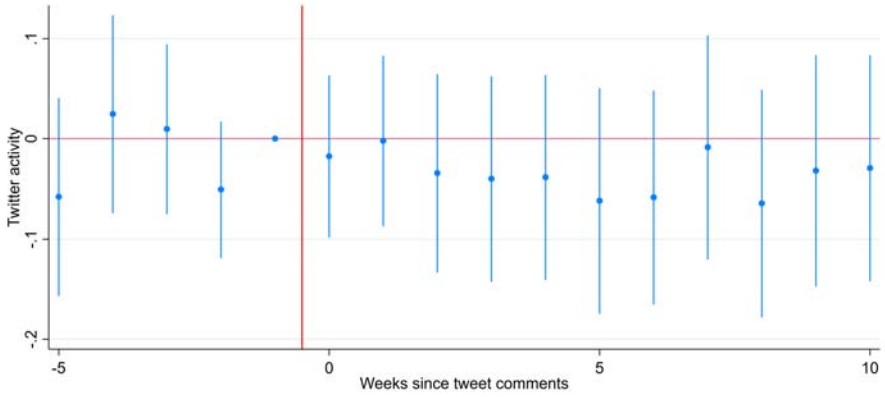
where $Var_i$ indicates the dimension of interest in exploring heterogeneous effects, estimated separately for each political party identification. *PR Scale:* responses to the Hong psychological reactance scale (Hong and Faedda, 1996) (Experiment 2 only). *Risk:* response to "Please tell us, in general, how willing or unwilling you are to take risks." (0 Completely unwilling to take risks - 10 Very willing to take risks). *Active SM user:* response to "How much time per day do you spend on social media (Facebook, Twitter, Instagram, Tik Tok, Snapchat, etc)" (never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours) and is coded as 1 if the participant states at least 1 hour. *Importance Gender/Race:* response to "How important is the issue discussed in the statement to you?" (1 Not important at all - 5 Extremely important). All variables are standardised. Controls include age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. Robust standard errors are clustered at the individual level.

Figure A4: Average welfare under different policy positions




Notes: Average welfare calculated using equation 5 with the following input: $p \in [1,7]$, stated opinion $s_i$ (reverse-coded for the gender question), stated or implied importance of issue. Implied importance is calculated as follows: we regress WTP on stated opinion, responses to the five questions about perceptions of political climate and freedom of speech online and controls including age, gender, race, education, employment, risk attitude, political affiliation, social media use, and wave FE $\times$ topic FE. The residuals are then regressed on stated importance. The predicted values yield our estimate of implied importance, re-scaling it by subtracting the constant. $p_{all} = 4.40$ is the average opinion of all participants and $p_{pub} = 3.92$ is the average opinion of those participants willing to publish.

Figure A5: Changes in Twitter activity after negative comments




Notes: The figure reports the coefficients from the triple-difference event-study specification comparing changes in Twitter activity (log(number of tweets+1)) after negative comments, relative to positive comments.

Figure A6: Attitudes in Pilot 1

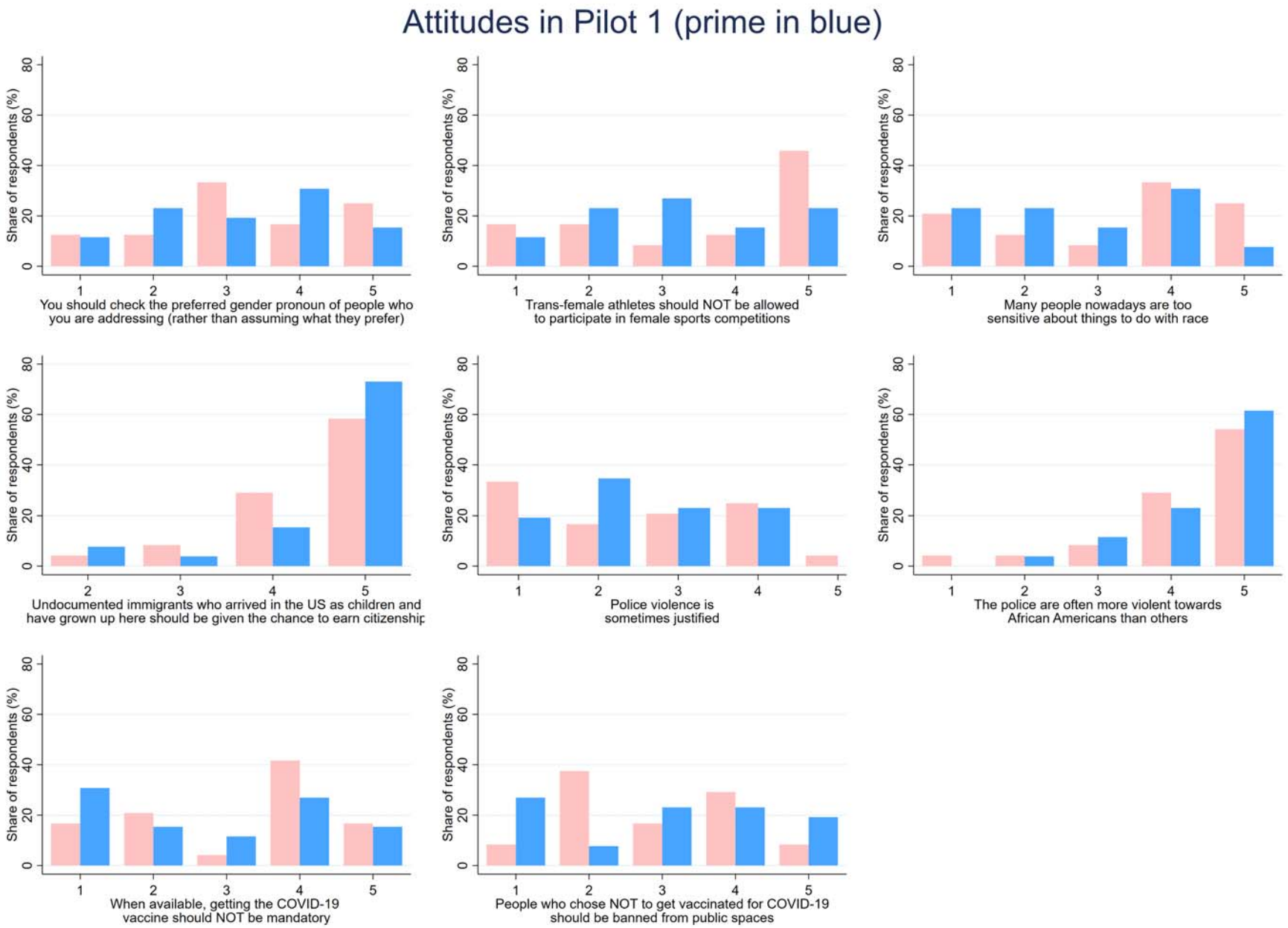
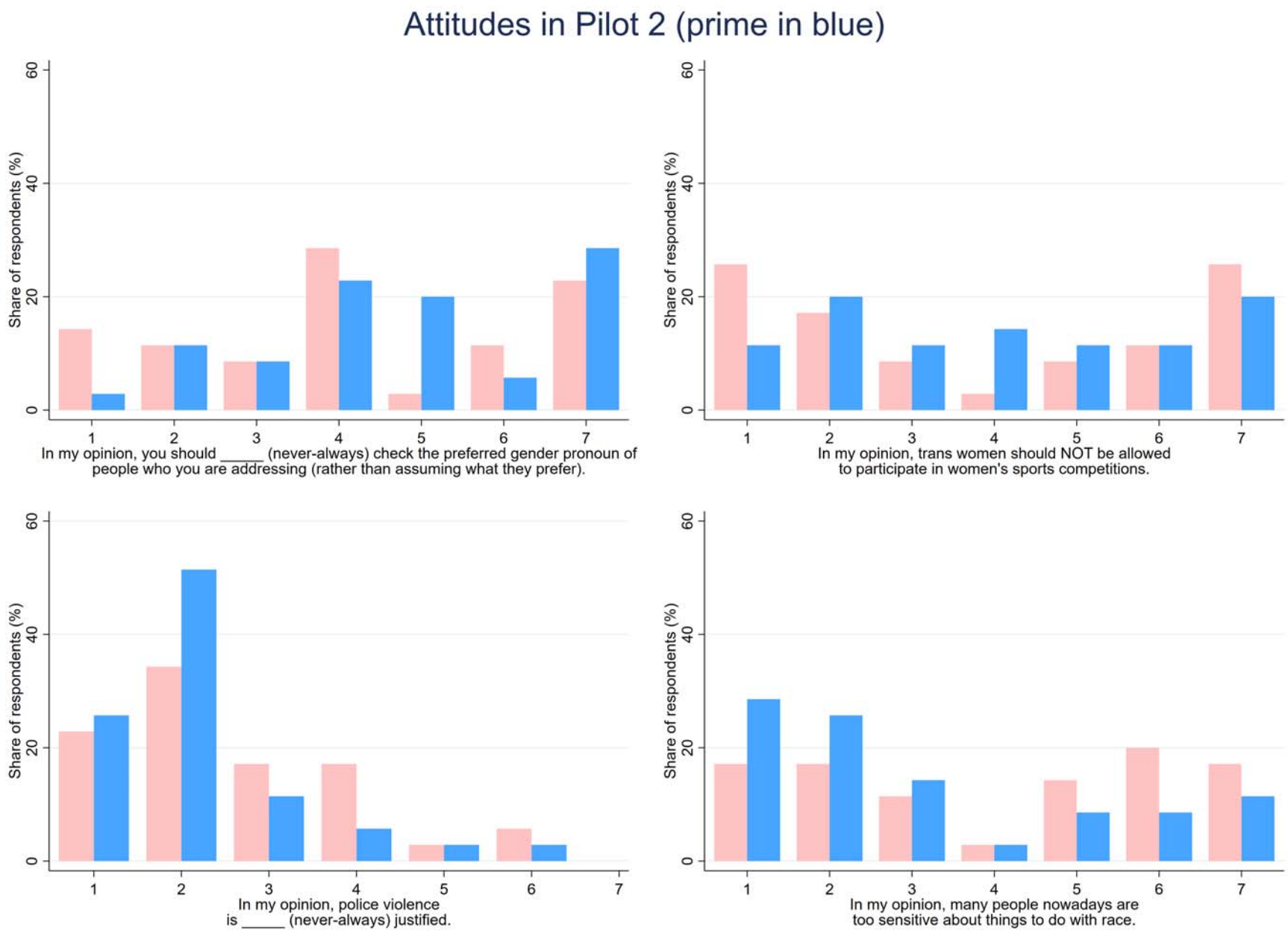


Figure A7: Attitudes in Pilot 2



# D Full Survey

The full survey for Wave 1 is provided on the next page. Modifications for other waves are detailed in the main text and available with our pre-registration.

***[Horizonal lines indicate page break. Unless otherwise specified, all options are presented as radio buttons.]***

**Introductory Statement**
This study is conducted by Dr Margaret Samahita from the School of Economics, University College Dublin.

**What is this research about?**
This study is part of a research project to study the opinions of Americans.

**Why have you been invited to take part?**
You have been invited to take part since you meet the research requirement: you are an adult aged over 18 years living in the US.

**How will your data be used?**
Unless otherwise noted, your data will be analysed and aggregate results will be reported in a future research paper for publication in an academic journal.

**What will happen if you decide to take part in this research study?**
You will fill out a 10-15 minute survey through Prolific using your desktop computer.

**How will your privacy be protected?**
Unless otherwise noted, we will collect your Prolific participant ID as is standard procedure, ensuring the data is anonymous.

**What are the benefits of taking part in this research study?**
Your responses will help researchers better understand the opinions of Americans and how these are formed. You will be paid a participation fee as is standard on Prolific. You will also have the possibility of earning an additional $100 bonus payment through a lottery. You start this survey with 10 tickets and your chance of winning is approximately 1 in 1000.

**What are the risks of taking part in this research study?**
There are no foreseeable risks to taking part in this study beyond that arising from everyday activities. However, if you have any concern and wish to withdraw at any point, simply close the survey window.

**Can you change your mind at any stage and withdraw from the study?**
Yes, if you wish to withdraw at any point, simply close the survey window.

**How will you find out what happens with this project?**
Future updates to the project will be available by contacting the researcher.

**Contact details for further information**
margaret.samahita@ucd.ie

If you consent to the above information sheet, please select Yes below.
I have read and understood the above and want to participate in this study.
o Yes

o No

---

Please enter your Prolific ID _____

---

What is your age (in years)? _____

What is your gender?
o Man
o Woman
o Non-binary/Other _____
o Prefer not to say

Please specify your ethnicity.
o White
o Hispanic or Latino
o Black or African American
o Native American or American Indian
o Asian / Pacific Islander
o Other _____

In which state do you currently reside? -Dropdown menu containing 50 US states]

---

What is the highest level of school you have completed or the highest degree you have received?
o Less than high school degree
o High school graduate (high school diploma or equivalent including GED)
o Some college but no degree
o Associate degree in college (2-year)
o Bachelor's degree in college (4-year)
o Master's degree
o Doctoral degree
o Professional degree (JD, MD)

Which statement best describes your current employment status?
o Working (paid employee)
o Working (self-employed)
o Not working (temporary layoff from a job)
o Not working (looking for work)
o Not working (retired)
o Not working (disabled)
o Not working (other) _____
o Prefer not to answer

Please tell us, in general, how willing or unwilling you are to take risks. [0-10 Likert scale, 0 Completely unwilling to take risks to 10 Very willing to take risks]

---

In political matters, people talk of 'the left' and 'the right'. How would you place your views on this scale, generally speaking? [0-10 Likert scale, 0 The Left to 10 The Right]

Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else?

o Republican
o Independent
o Democrat
o Other _____
o No preference

How much time per day do you spend...
-On social media (Facebook, Twitter, Instagram, Tik Tok, Snapchat, etc) [never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours]
-Watching, reading or listening to news about politics and current affairs [never/no account, less than 30 minutes, from 30 minutes to 1 hour, from 1 hour to 2 hours, more than 2 hours]

---

*[Control text UNITO]*

Please read the following text.

---
The University of Turin is one of the most ancient and prestigious Italian Universities. Hosting over 79,000 students and with 120 buildings in different areas in Turin and in key places in Piedmont, the University of Turin can be considered as “city-within-a-city”, promoting culture and producing research, innovation, training and employment.

Facilities include 22 libraries spread over 32 locations, the Botanic Garden and several University Museums such as "Cesare Lombroso" - Criminal Anthropology Museum and "Luigi Rolando" - Human Anatomy Museum.


To check that you are paying attention, how many museums are named in the text?

o 22
o 2
o 32

---

*[Control text UCD]*

Please read the following text.

---

University College Dublin (commonly referred to as UCD) is a research university in Dublin, Ireland, and a member institution of the National University of Ireland. With 33,284 students, it is Ireland's largest university. Five Nobel Laureates are among UCD's alumni and current and former staff. UCD's main campus is located on a 133-hectare (330-acre) campus at Belfield, four kilometres to the south of the city centre. In 1991, it purchased a second site in Blackrock. This currently houses the Michael Smurfit Graduate Business School.

A report published in May 2015 showed the economic output generated by UCD and its students in Ireland amounted to €1.3 billion annually.

To check that you are paying attention, where does the text say UCD's main campus is located?

- o Smurfit
- o Belfield
- o Blackrock

---

*[Prime text]*

Please read the following text.

---

The public nature of social media has resulted in individuals sometimes experiencing negative consequences as a result of their posts, in a phenomenon that some people refer to as **"cancel culture".**

*"Those most vulnerable to harm tend to be **individuals previously unknown to the public**, like the communications director who was **fired** in 2013 after posting on social media, from her personal account, **an ill-thought-out joke** about Africa, AIDS and her own white privilege ... or the data analyst who was **fired** last spring after posting on social media, after the death of George Floyd in police custody, a study that suggested that riots depressed rather than increased Democratic Party votes."*

These cases highlight the risk of **public backlash from social media.**


To check that you are paying attention, what does the text say cancel culture can result in?

o losing a job
o lower voter turnout
o toppling a famous figure

---

*[OPINION ELICITATION---for a description of the survey logic, please see the experimental design]*

You will now be asked to state your opinion on a number of questions.

---

Please consider the following statement.

**People who have been vaccinated against COVID-19 should be allowed to travel without testing and quarantine requirements.**

What do you think of the above statement? [1-7 Likert scale, 1 Strongly disagree to 7 Strongly agree]

---

Please consider the following statement.

**In my opinion, trans women should be allowed to participate in women's sports competitions.**

What do you think of the above statement? [1-7 Likert scale, 1 Strongly disagree to 7 Strongly agree]

---

Please consider the following statement.

**In my opinion, many people nowadays are too sensitive about things to do with race.**

What do you think of the above statement? [1-7 Likert scale, 1 Strongly disagree to 7 Strongly agree]

---

*[PUBLICATION ELICITATION---for a description of the survey logic, please see the experimental design]*

Would you be willing to let us post on social media, anonymously, your response to the previous statement:

---

*[Participant 37]*
*"I [pipe selected choice] that many people nowadays are too sensitive about things to do with race."*

---

[In T4-5 only] Note that if you would like to change your response, you can simply click the Back (left arrow) button below to go back to the previous page.

If you select Yes, **we will create a tweet** containing the above response and post it on a public Twitter page created once data collection is complete. Participant numbers (eg, 37 in the above) are randomly assigned and not linked to your identity in any way.

If you select No, we will NOT create a tweet containing the above.

o Yes
o No

---

Would you be willing to let us post on social media, <u>together with your name</u>, your response to the previous statement:

---

*[Your name here]*
*"I [pipe selected choice] that many people nowadays are too sensitive about things to do with race."*

---

[In T4-5 only] Note that if you would like to change your response, you can simply click the Back (left arrow) button below to go back to the previous page.

-**We will create a tweet** containing the above response and may post it on a public Twitter page created once data collection is complete (* see below)
-***We will contact Prolific to request your first and last names.** Note that while in general Prolific does not allow researchers to collect personal information, Prolific does encourage researchers to get in touch in cases such as this, where the study design requires the collection of personal data (see https://researcher-help.prolific.co/hc/en-gb/articles/360015378834-Can-I-ask-Participants-for-their-Personal-Information-Identifiers-).
-The tweet will only contain **a text of your name without any hyperlink**, the public Twitter page will potentially contain the names and opinions of many participants.
-The link to the public Twitter page will be **made available to participants** who contact the researcher to ask for it, but it will not be otherwise advertised. The public Twitter page will be **deleted after 30 days.**

o Yes, I would like to
o No, I'd rather not

---

*[WILLINGNESS TO PAY ELICITATION for subjects who chose "Yes, I would like to" above---for a description of the survey logic, please see the experimental design]*

You stated that you would like us to post on social media, <u>together with your name</u>, your response to the previous statement:

---

*[Your name here]*
*"I [pipe selected choice] that many people nowadays are too sensitive about things to do with race."*

---

In exchange for this post, we want to know if you would be willing to **give up some of your lottery tickets** for the $100 bonus (remember that you start with 10 tickets).

*Would you be willing to give up **all 10 lottery tickets** in exchange for this public post? [This question is repeated with 5 lottery tickets and 1 lottery ticket. If Yes is selected, the subject moves on to the next section.]*

o Yes
o No

If you select Yes,
-**We will contact Prolific to request your first and last names.** Note that while in general Prolific does not allow researchers to collect personal information, Prolific does encourage researchers to get in touch in cases such as this, where the study design requires the collection of personal data (see https://researcher-help.prolific.co/hc/en-gb/articles/360015378834-Can-I-ask-Participants-for-their-Personal-Information-Identifiers-).
-Note, we will only reduce your lottery tickets if we do publish the above text with your name.

---

*[WILLINGNESS TO ACCEPT ELICITATION for subjects who chose "No, I'd rather not" above---for a description of the survey logic, please see the experimental design]*

You would rather not let us post on social media, <u>together with your name</u>, your response to the previous statement:

---

*[Your name here]*
*"I [pipe selected choice] that many people nowadays are too sensitive about things to do with race."*

---

We would now like to ask whether you would be willing to **change your mind** in exchange for a **higher chance of winning the $100 lottery**. Remember that you start with 10 tickets.

*Would you be willing to let us post the above if we give you **1 additional lottery ticket?** [This question is repeated with 5, 25, and 50 lottery tickets. If Yes is selected, the subject moves on to the next section.]*

o Yes
o No

If you select Yes,
-You will get 1 additional ticket in the lottery.
-**We will contact Prolific to request your first and last names.** Note that while in general Prolific does not allow researchers to collect personal information, Prolific does encourage researchers to get in touch in cases such as this, where the study design requires the collection of personal data (see https://researcher-help.prolific.co/hc/en-gb/articles/360015378834-Can-I-ask-Participants-for-their-Personal-Information-Identifiers-).

---

Please consider the following statement.

**In my opinion, trans women should be allowed to participate in women's sports competitions.**

How important is the issue discussed in the statement to you? [1-5 Likert scale, 1 Not important at all to 5 Extremely important]

---

Please consider the following statement.

**In my opinion, many people nowadays are too sensitive about things to do with race.**

How important is the issue discussed in the statement to you? [1-5 Likert scale, 1 Not important at all to 5 Extremely important]

---

*[NORM ELICITATION---for a description of the survey logic, please see the experimental design]*

As earlier mentioned, you have the chance to win an additional bonus of $100 through a lottery.

You will now see **2 questions**. You will earn **5 additional lottery tickets** for each question you answer correctly, in addition to your existing tickets.

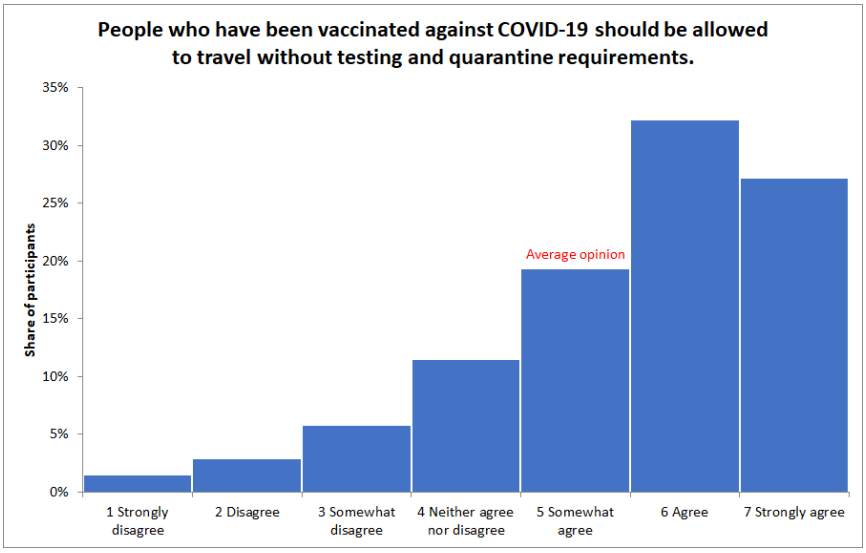
Therefore, please consider your answers carefully since each correct answer will increase your chance of winning the $100 bonus.

---

You will now be asked what you think about the **average** opinion out of other participants in this study.

Here is an example using the COVID-19 question. Suppose that the share of participants who state a particular opinion (between 1 to 7) is as shown in the graph below.

**The average opinion** is calculated by summing up everyone's opinion and dividing by the total number of participants. In this example, the average opinion is **5 - Somewhat agree.**




Please consider the following statement.

**In my opinion, many people nowadays are too sensitive about things to do with race.**

Remember, you will earn 5 additional lottery tickets for each correct answer, so please consider your answers carefully.

Considering ALL participants (in this US-based survey), what do you think the **average** opinion is? [1-7 Likert scale, 1 Strongly disagree to 7 Strongly agree]

Considering those participants (in this US-based survey) who stated that they WOULD be willing to let us post their opinion, together with their name, on social media (without any additional payment), what do you think the **average** opinion is? [1-7 Likert scale, 1 Strongly disagree to 7 Strongly agree]

---

How often do you worry that things you post on social media can be misinterpreted? [1-7 Likert scale, 1 Never to 7 Always]

The political climate these days prevents me from saying things I believe because others might find them offensive. [1-7 Likert scale, 1 Strongly disagree to 7 Strongly agree]

Are you worried about losing your job or missing out on job opportunities if your political opinions become known? [Not at all worried, Not very worried, Worried a little, Worried a lot]

How often do you think social pressure causes people to **misrepresent or lie about** their political opinions on social media? [1-7 Likert scale, 1 Never to 7 Always]

How often do you think social pressure causes people to **refrain or abstain from expressing** political opinions on social media? [1-7 Likert scale, 1 Never to 7 Always]

---

Thank you for participating in our study.

This study aims to investigate the impact of cancel culture on self-expression. We are interested in how willing you would be to let us post your opinion on social media.

You were shown some of the following three texts:

- The text about UCD was modified from https://en.wikipedia.org/wiki/University_College_Dublin and serves as a filler.
- The text about UNITO was modified from https://en.unito.it/about-unito/unito-glance and serves as a filler.
- The text about cancel culture was modified from https://www.nytimes.com/2020/12/03/t-magazine/cancel-culture-history.html

As data collection is ongoing, we would like to ask you not to talk about this study with others for now.

If you win the bonus payment, it will be paid through Prolific in the next few weeks.

Regarding the publication of your opinion on social media:

- We will create a public Twitter page for the study.
- We will create an anonymous tweet for each participant's opinion that they are willing to publish.
- Previous requests to Prolific asking for participant's names in a similar study design have been turned down; so we do not anticipate that we will publish your opinion with your name, even if you stated that you would like us to do this. [For subjects who were willing to accept extra tickets for publishing:] Regardless, if you stated that you were willing to publish the opinion with your name in exchange for lottery tickets, you will still get these additional lottery tickets.

If you have any questions about the study, please feel free to contact Margaret Samahita (margaret.samahita@ucd.ie).