Cancelled: conformity or silence?
Pre-analysis plan[*]

Juan S. Morales[†] Margaret Samahita[‡]

August 18, 2021

[*]All errors are our own.

[†]Collegio Carlo Alberto and ESOMAS Department, University of Turin. E-mail: juan.morales@carloalberto.org.

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

1 Motivation

We often express opinions which in reality differ from our private views (Kuran, 1997). Other times, we may prefer to avoid expressing our views altogether (Noelle-Neumann, 1974). Social norms, social image concerns, and social stigma are all factors which can shape the public expression of opinion. With increased political polarization and the rise of social media, many prominent voices have recently argued that some of these social norms have become too strict and that fear of social backlash has led to a stifling of freedom of expression. A letter co-signed by many famous figures stated that "the free exchange of information and ideas, the lifeblood of a liberal society, is daily becoming more constricted" (Harpers). Are they right?

This paper studies how perceived social pressure affects the public expression of opinion. How sensitive are individuals to current concerns of social backlash? To what extent may public views differ from private opinion? Do these dynamics vary depending on individual characteristics like political affiliation or social media use? To investigate these questions, we first outline a simple model that formalizes these ideas and helps to frame our study. We highlight how social pressure can affect public opinion either through a change in publicly stated views towards a norm (*conformity*) or by inducing self-censorship (*silence*). We then propose to conduct an online survey experiment that allows us to disentangle these two effects and to better understand these social dynamics in the context of two current debates: Gender and LGBTQ issues, and Race.

2 Conceptual framework

Individuals choose: i) a public stance $s_i \in [0,1]$, and ii) whether to "speak-up" or voice their stance, or not $v_i \in \{0,1\}$. In addition, individuals hold private views/opinion $o_i \in [0,1]$, and there are norms that dictate what an appropriate public stance is $n \in [0,1]$.[1] Individuals get a social reward for speaking up κ>0. They maximize their utility:

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

[1] In some cases the norm n may vary by individual, if for instance Democrats and Republicans adhere to different norms, or 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 β is the cost or risk from social disapproval and α 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 utility of expressing the above optimal public stance is

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

When there is no conflict between the private opinion and norms, $s_i^* = o_i = n$ and the individual will always choose to speak up, since $\kappa > 0$. When $o_i \neq n$, s_i^* will be a weighted average between private opinion and norms with weights depending on the social disapproval and cognitive dissonance parameters β and α. Speaking up decreases with the distance between o_i and n.

3 Design

The experiment timeline is shown in Figure 1. We start by collecting data on participants' demographics, including political preferences and social media use. Since we hypothesise that treatment effects would be heterogeneous along the latter two dimensions, we elicit these variables pre-treatment (Montgomery, Nyhan and Torres, 2018). To elicit political preference, we ask: "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.

3.1 Treatments

Following the demographic questionnaire, participants are randomised into one of five treatments. As shown in Figure 1, in Treatments 1-3, attitudes and "willingness to express publicly" are elicited separately while in Treatment 4-5 they are elicited at the same stage. The assigned treatment also determines which two texts participants are

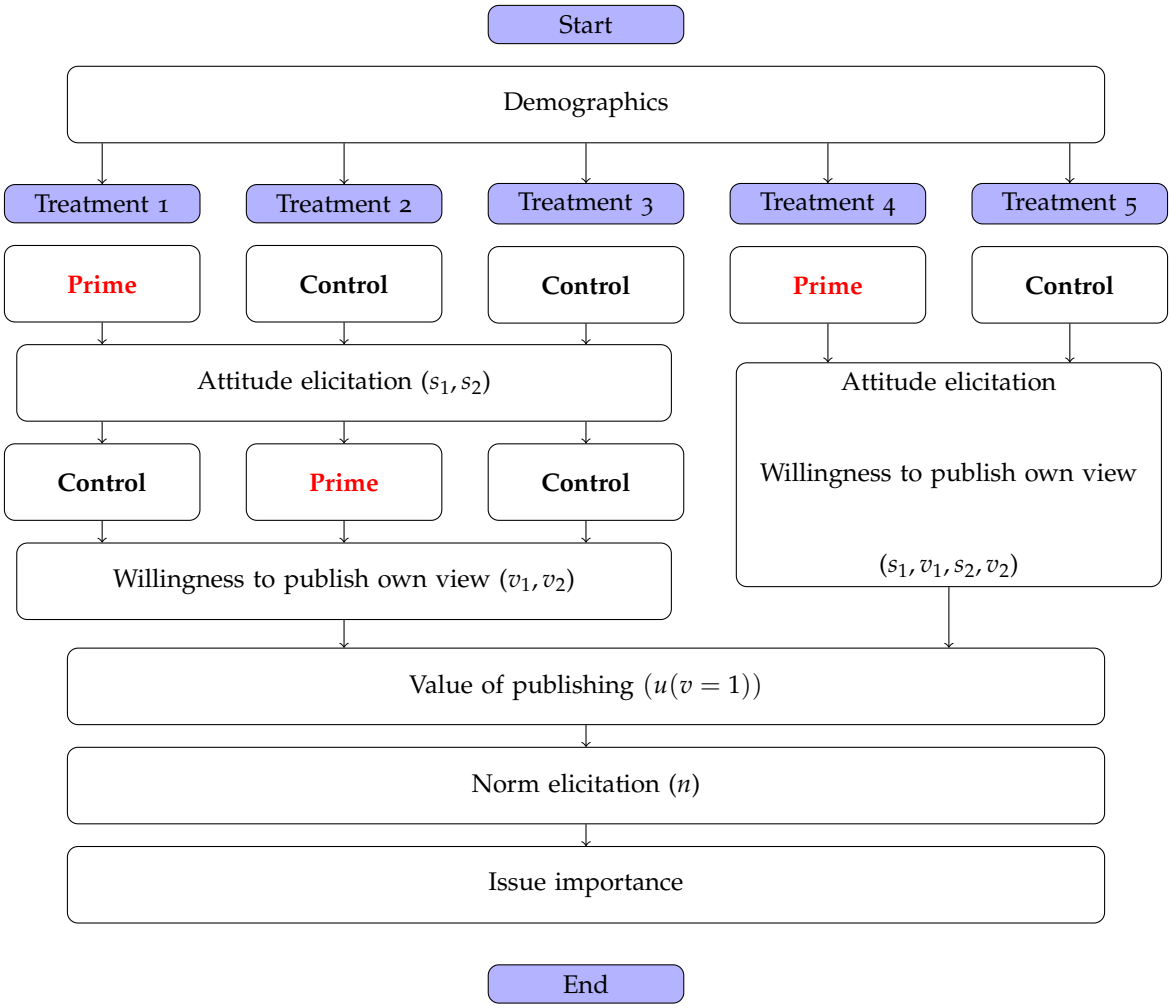
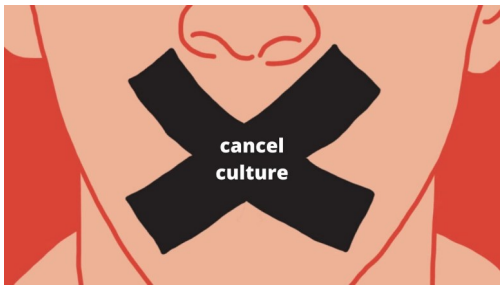


Figure 1: Experiment timeline

shown *before* (and, in Treatments 1-3, *after*) the attitude elicitation. In Treatment 1 (T1), before the attitude elicitation participants are shown the following text and image to prime them to the risk of social disapproval:

> 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.**



We include 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.

After the attitude elicitation, participants in T1 are shown a control text about University College Dublin (UCD) or the University of Turin (UniTo) together with the university logo. We randomise which of the two texts is shown and include an attention check for each text.

In T2, the social disapproval prime is instead shown *after* the attitude elicitation. Before the attitude elicitation, they are shown one of the two control texts (UCD or

UniTo). In T3, participants see both control texts (one before and one after) the attitude elicitation (the order is randomised). In T4, participants see the social disapproval prime before the elicitation of attitude and willingness to publish, while in T5 they see one of the two control texts (UCD or UniTo).

3.2 Elicitation of attitude and willingness to publish

3.2.1 Attitude

Participants are 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 are 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).

3.2.2 Willingness to publish

We next ask 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 ask: "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 the following:

- 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 is crucial that participants seriously consider the possibility that their opinions will be shown to others. However, actual publication with names is neither something we want to nor can do given the potential for negative consequences for the participants (and previous attempts at accessing participants' names from Prolific have been turned down). Additionally, we seek to follow the standard of no deception in experimental economics. We therefore truthfully inform participants that, should they say Yes, we will attempt to obtain their names for the purpose of publication, but that publication is conditional on an event that (as we explain in the debrief) has an extremely low chance of happening.

3.2.3 Value of publishing

We also elicit a quantitative measure of participants' willingness to publish the above response with their name for one of the two statements by randomising participants into either a Race or Gender condition. We first endow all participants with 10 tickets for a USD 100 bonus lottery. Participants are informed at the start that their chance of winning is approximately 1 in 1000. After participants are asked the above question on willingness to publish with their name, if they select "Yes, I would like to", they are then asked whether, in exchange for this post, they would be willing to give up 10, 5, or 1 of their lottery tickets. These questions are asked sequentially starting from the highest value. If/when they select Yes, they move on to the next section.

If participants state "No, I'd rather not" to the question about publication with their name, they are then asked whether they would change their mind in exchange for a higher chance of winning the USD 100 lottery. We ask if they would be willing to let us post their response if we give them 1, 5, 25, or 50 additional lottery ticket. These questions are asked sequentially starting from the lowest value. If/when they select Yes, they move on to the next section.

In Treatments 1-3, we first elicit participants' attitude to each statement before showing them the second prime/control text. This is then followed by the two questions about willingness to publish, which are immediately repeated for the second statement. Finally, we ask about the value of publishing for the randomly chosen statement.

In Treatments 4-5, participants are asked: i) what they think of the statement, and ii) their willingness to publish their response, anonymously and then together with their name. This is then repeated for the second statement. In contrast to Treatments 1-3, participants can go back to the previous page thus allowing them to change their answer to i) after considering their answers to ii). We include this treatment branch since participants may perceive different social pressure from revealing their opinion to the researchers (in T1-T3) rather than the public. This is then followed by the quantitative elicitation of willingness to publish for a randomly chosen statement.

Next, all participants are 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).

3.3 Norm elicitation

We proceed by eliciting the descriptive norm: beliefs about others' stated opinion, denoted n_i. We do this for one of the two statements participants were asked to consider depending on whether participants are in the Gender or Race condition. After showing the statement, we ask 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?

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. We incentivise participants by rewarding each correct answer with 5 additional lottery tickets for the USD 100 bonus.

3.4 Issue importance

Finally, we ask participants five questions to check how important social disapproval is considered to be. These questions are:

- "How often do you worry that things you post on social media can be misinterpreted?" (Never - Always)
- "The political climate these days prevents me from saying things I believe because others might find them offensive." (Strongly disagree - 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 - Worried a lot).
- "How often do you think social pressure causes people to misrepresent or lie about their political opinions on social media?" (Never - Always)
- "How often do you think social pressure causes people to refrain or abstain from expressing political opinions on social media?" (Never - Always)

3.5 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.

3.6 Implementation

Participants are recruited using the data collection platform Prolific. In order to ensure a balanced number of participants across political affiliations, we recruit 300 self-

identified Democrats, Republicans and Independents, giving us a total of 900 participants.[2] We allocate 27.5% of participants to each of T1-2 and 15% in T3-5 to enable us test Hypothesis 2b with greater power.

4 Hypotheses

Our experiment investigates the impact of an increase in the perceived cost of social disapproval (an increase in β) on individuals' reported views (the conformity effect) and their willingness to publicly express those views (the silencing effect).

4.1 Conformity

The absolute distance between the optimal public stance and the perceived norm is

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

The effect of increasing β on individuals' public stance s_i^* is thus to move it closer to the norm n.

Hypothesis 1. *Priming with social disapproval increases β. Therefore, the distance between an individual's public stance and the perceived norm is smaller when the individual is primed with social disapproval.*

4.2 Silence

We measure individuals' willingness to publish their opinion on a public website. We make a distinction between three cases: no priming about social disapproval (C), the prime is shown *before* the individuals' public stance is elicited (T1), and the prime is shown *before* asking about willingness to publish (T2).

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

$$u_i(v_i = 1) = \begin{cases} \kappa - \left[\beta^T(s_i^*(\beta^T) - n)^2 + \alpha(s_i^*(\beta^T) - o_i)^2\right] = u^{T1} & \text{if primed before } s_i^* \\ \kappa - \left[\beta^T(s_i^*(\beta^C) - n)^2 + \alpha(s_i^*(\beta^C) - o_i)^2\right] = u^{T2} & \text{if primed between } s_i^* \text{ and } v_i^* \\ \kappa - \left[\beta^C(s_i^*(\beta^C) - n)^2 + \alpha(s_i^*(\beta^C) - o_i)^2\right] = u^{C} & \text{if not primed} \end{cases}$$

[2] Among Independents we also include those who state their political affiliation to be "None" or "Other".

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

In cases C and T1, individuals choose s optimally. That is, β is the same for both choices (s and v). Since $\beta^T > \beta^C$, then $u^{T1} < u^C$ from (2). When individuals are primed *before* the elicitation of public opinion, we expect that willingness to publish will be lower ($v_i^{T1} \leq v_i^C$).

Hypothesis 2a. *Priming with social disapproval increases β. Therefore, individuals are less willing to publish their opinion when they are primed with social disapproval before the elicitation of s than if they are not primed.*

In case T2, individuals choose s using β^C (before the prime is shown), and choose v using β^T (after the prime is shown). The distance between the norm and their reported stance is then higher, that is $|s_i^{T2} - n| > |s_i^{T1} - n|$, 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, $u^{T2} < u^{T1}$ and $v_i^{T2} \leq v_i^{T1}$.[3]

Hypothesis 2b. *Priming with social disapproval increases β. Therefore, individuals are less willing to publish their opinion when they are primed with social disapproval after the elicitation of s than if they are primed before.*

Consequently,

Hypothesis 2c. *Individuals are less willing to publish their opinion when they are primed with social disapproval after the elicitation of s than if they are not primed.*

Moreover, since

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

it is easy to see that:

Hypothesis 3. *Priming with social disapproval increases β. Therefore, the treatment effect increases with the distance between the individual's public stance and the perceived norm.*

On the other hand, the prime may also have the effect of increasing κ—speaking up when few others do may yield a greater reward for the individual. Therefore,

Hypothesis 4. *Priming with social disapproval increases κ (in addition to β). Therefore, individuals closest to the norm are more willing to publish their opinion when they are primed with social disapproval.*

Consider individuals for whom $s_i^* = o_i = n$. For them, $u_i = \kappa$, and the prime should only increase the likelihood of speaking up if there is indeed an increase in κ.

[3]The proof is given in the appendix.

4.3 Comparative statics

- Increasing α, the cognitive dissonance cost, directly affects s_i^* by bringing it closer to o_i. Additionally, increasing α decreases the utility from speaking up. Hence, individuals' public stance s_i^* is closer to their true opinion and they become less willing to speak up.
- Increasing κ, the social reward for speaking up, does not directly affect s_i^* but makes individuals more willing to speak up.

5 Analyses

5.1 Conformity

To test Hypothesis 1, we first define the outcome variable $Dist_{iq}$ as the absolute distance between a respondent i's reported public stance (from the 1-7 agreement scale, our measure of s_i) and *majority* opinion among those who speak up in the whole sample (our proxy measure of n) for question q.[4]

We then estimate the following:

$$Dist_{iq} = \alpha + \beta_1 Prime_i + \beta_2 Simultaneous_i + \delta_q + \varepsilon_{iq}$$
$$Dist_{iq} = \alpha + \beta_1 Prime_i + \beta_2 Simultaneous_i + \beta_3 Prime_i \times Simultaneous_i + \delta_q + \varepsilon_{iq}$$

where $Prime_i$ is a treatment dummy that takes value 1 if subject i sees the prime before s is elicited (T1 and T4), $Simultaneous_i$ takes value 1 if s is elicited simultaneously as the choice to publish (T4 and T5), and ε_{iq} is an individual-question specific error term. Our hypotheses indicate that $\beta_1 < 0$. We also expect that subjects' stated opinion will be closer to the perceived norm when s is elicited together with the choice to publish, hence we hypothesise that $\beta_2 < 0$. This second margin helps us to measure the extent to which individuals report different views to the researchers than to the "public" (which may also increase the disapproval cost but also the social reward from reporting). We expect that $\beta_3 < 0$ since the effect of the prime should be greater when individuals are asked about their stated opinion with publication in mind.

We include question fixed effects δ_q. In some specification(s) we can include a vector of controls $\mathbf{X_i}$ including age, gender, race, education, employment, risk attitude,

[4] Recall that in the model those who speak up are individuals for whom o (and therefore s) is close to n. As an alternative, we will also structurally estimate n using iterative nonlinear least squares estimation.

political leaning, political leaning squared and social media use, 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 will have two observations per subject).

5.1.1 Other outcomes

We may also use the time taken to choose a public stance, and the number of time a participant goes back to revise their answer in T4-5, as additional outcome variables (predicted to be higher if primed, and lower in the second question in T4-5). We will also use s_i as an outcome variable to check for directional effects in the stated opinion of our subjects.

Finally, we investigate whether individuals in the primed group think others are more likely to misrepresent or lie about their political opinions on social media due to social pressure.

5.2 Silencing

Our main outcome in this section is individuals' willingness to publish their opinion with their name. We define v_i as a binary variable which takes value 1 if subject i is willing to publish their opinion (without additional lottery tickets as incentive).

We then estimate the following regression:

$$v_{iq} = \alpha + \beta_1 Prime_i + \beta_2 Prime_i \times LatePrime_i + \beta_3 Simultaneous_i + \delta_q + \varepsilon_{iq}$$
$$v_{iq} = \alpha + \beta_1 Prime_i + \beta_2 Prime_i \times LatePrime_i + \beta_3 Simultaneous_i + \beta_4 Prime_i \times Simultaneous_i + \delta_q + \varepsilon_{iq}$$

where $Prime_i$ takes value 1 if subject i receives the prime before v is elicited (T1, T2 and T4) and $LatePrime_i$ takes value 1 if subject i receives the prime after s is elicited (T2). We hypothesise that $\beta_1 < 0$ and that $\beta_2 < 0$. We hypothesise that $\beta_3, \beta_4 > 0$ since s should be closer to the norm when elicited together with v.

To test Hypotheses 3 and 4, we interact the treatment dummy with $Dist_{iq}$, for example:

$$v_{iq} = \alpha + \beta_1 Prime_i + \beta_2 Prime_i \times LatePrime_i + \beta_3 Simultaneous_i + \theta_1 Prime_i \times Dist_{iq} + \theta_2 Prime_i \times LatePrime_i \times Dist_{iq} + \gamma Dist_{iq} + \delta_q + \varepsilon_{iq}$$

hypothesising that H3: $\theta_1 < 0$ and H4: $\beta_1 > 0$, $\beta_2 = 0$. We are aware that $Dist_{iq}$ is an endogenous regressor in the specification but we are interested in this relationship as predicted by Hypothesis 4. Due to this concern, in an additional specification we include only individuals in T2 (those who were primed after the attitude elicitation) and T3. Finally, by construction (and assumption derived from our model), $\gamma < 0$.

We include question fixed effects δ_q. In some specification(s) we can include a vector of controls $\mathbf{X_i}$ including age, gender, race, education, employment, risk attitude, political leaning, political leaning squared and social media use, 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).

5.2.1 Other outcomes

We also use individuals' willingness to publish their opinion anonymously. Additionally, we use the number of lottery tickets the participant is willing to pay/accept for publication as another outcome variable. We define $lottery_i$ to be the value at which the participant responds Yes to publishing. For example, if they are willing to pay 10 tickets, $lottery_i = 10$, while if they are willing to accept 5 tickets, $lottery_i = -5$. Note that these values are potentially the lower bound: a participant who answers Yes to paying 10 tickets may have been willing to pay a higher amount. For those unwilling to accept 50 tickets, we impute a value of -100 for the analysis. In addition, we will define categorical variables for each of these groups.

We may also use the time taken to decide willingness to publish, and the number of time a participant goes back to revise their answer in T4-5, as an additional outcome variable (predicted to be higher if primed, and lower in the second question in T4-5).

Finally, we investigate whether individuals in the primed group think others are more likely to refrain or abstain from expressing political opinions on social media due to social pressure.

5.3 Norms analysis

First, we will test whether conformity to the global norm (defined above as the majority view among those willing to publish their opinion in the whole sample) varies across political background (left-right scale or party identification). We will also include a triple interaction term with the participant's state, predicting that treatment effect will

be even stronger for participants identifying with a political party that is popular in their state (Bursztyn, Egorov and Fiorin, 2020).

Second, we will split the sample by party identification and redefine a new (local) norm defined as the majority view among those willing to publish their opinion in each group (Democrats, Independents and Republicans). We then repeat the above analyses for Conformity to check i) whether different groups conform to their own norm, and ii) whether there is heterogeneity in conformity to own group's norm.

To see if the prime affects individuals' perception of conformity and silencing across other participants, we use our measures of n_i. In particular, we use $n_{all} - n_{pub}$ as an outcome variable. We predict that this reported attitude gap will be bigger for the primed group. If individuals think that others conform to a majority norm due to social pressure, then this difference will be affected by the treatment. As a complementary descriptive analysis, we also investigate whether $n_{all} - n_{pub}$ is heterogeneous across party identification, and in particular, whether this gap is larger for Republicans (who often express worry about cancel culture in popular media).

5.4 Heterogeneous treatment effects

In addition, we test for heterogeneous effects along other dimensions by interacting the treatment dummy as described above with different variables. For instance, and importantly, for active social media users as:

$$\begin{aligned} Dist_{iq} = {} & \alpha + \beta_1 Prime_i + \beta_2 Simultaneous_i + \beta_3 Prime_i \times Simultaneous_i \\ & + \theta_1 Prime_i \times ActiveSMuser_i + \theta_2 Prime_i \times Simultaneous_i \times ActiveSMuser_i \\ & + \gamma ActiveSMuser_i + \delta_q + \varepsilon_{iq} \end{aligned}$$

where $ActiveSMuser_i$ is an indicator equal to 1 if the individual spends more than 60 minutes daily on social media. While the reward for speaking up is expected to be higher for active users who are conditioned to chase after "likes" and "retweets", thus increasing κ, they may also be more sensitive to negative backlash and perceive β to be higher than passive/non-users. The overall effect on conformity is to lower $Dist_{iq}$ while the effect on speaking up is ambiguous.

Margins of heterogeneity we will explore include:

- Political leaning (as described in section 5.3)
- Active social media users (dummy equal to 1 if individual spends more than 60

minutes daily on social media): treatment effects are expected to be greater

- Importance of topic to participant (1-5 Likert scale): treatment effects are expected to be smaller
- Gender: treatment effects are expected to be greater for females (Croson and Gneezy, 2009) than males, we will also check for heterogeneity if answering "Non-Binary/Other" or "Prefer not to say"
- News consumption (dummy equal to 1 if individual spends more than 60 minutes daily watching/reading/listening to news about politics and current affairs): treatment effects are expected to be greater
- Other demographic variables: age, education, employment, race, risk attitudes
- Topic: in the Gender statement, a participant may disagree for fairness considerations but afraid to be perceived as transphobic. This motive is not present in the Race statement.
- Concern of social disapproval

5.5 Control variables

Our baseline specification includes:

- Age: coded continuously
- Gender: coded as a dummy for Man, Woman, Non-binary/Other ("Prefer not to say" as the omitted category)
- Race: coded as a series of dummies for White, Hispanic or Latino, Black of African American, Native American or American Indian, Asian/Pacific Islander ("Other" as the omitted category)
- Education: coded as a dummy for having at least a 2-year college degree
- Employment: coded as a dummy
- Risk attitude (Falk et al., 2018): coded on a 0-10 Likert scale and standardised
- Political leaning: coded on a 0-10 Likert scale and standardised

- Social media use: coded as a dummy for spending more than 60 minutes daily on social media
- Topic fixed effect

5.6 Robustness checks

We will check the robustness of our results to:

- Dropping subjects who do not answer the attention check correctly in the first attempt. We will also check whether the proportion of subjects correctly answering in the first attempt is significantly different depending on the first prime shown.

5.7 Other descriptive analyses

One important aspect of the topic that we can investigate with our data is whether "willingness to speak up" is correlated with how important individuals think a specific topic is. In particular, we study the following descriptive model:

$$v_{iq} = \alpha + \beta_1 TopicImportance_{iq} + \varepsilon_{iq}$$

where $TopicImportance_{iq}$ is the standardized measure of the reported importance that individual i assigns to question q. Studying this relationship may help us understand the broader welfare implications of cancel culture/social backlash.

In addition, we will study whether, consistent with narratives in popular media, Republican participants are more concerned about issues of freedom of speech and cancel culture as elicited by our last set of questions.

References

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

Croson, Rachel, and Uri Gneezy. 2009. "Gender differences in preferences." *Journal of Economic literature*, 47(2): 448–74.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global evidence on economic preferences." *The Quarterly Journal of Economics*, 133(4): 1645–1692.

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

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.

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

Appendices

A Proof of Hypothesis 2b

Proof. Given our treatment, we assume that $\beta_T > \beta_C$.[5]

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

That is, $u^{T2} < u^{T1}$. □

B Full Survey

Begins on the next page.

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

[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.

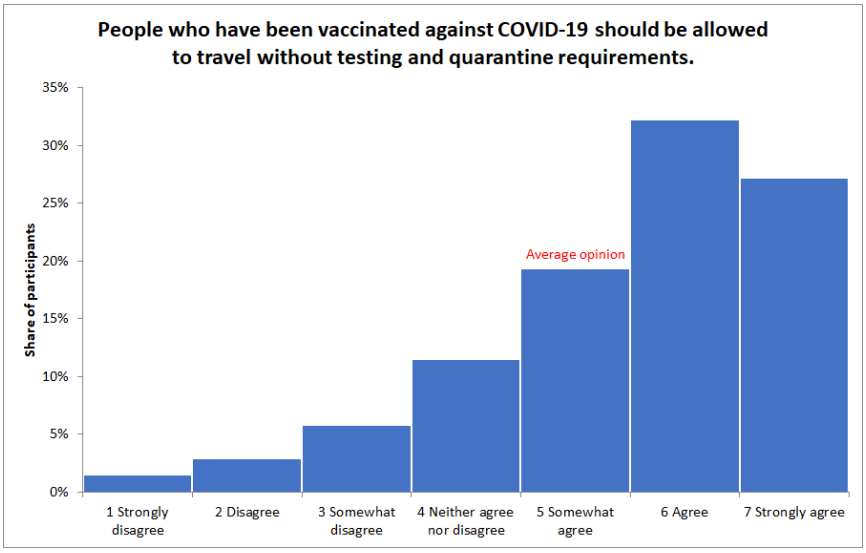
You will now see **2 questions**. You will earn **5 additional lottery tickets** for each question you answer correctly, in addition to your existing tickets.

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).