Conformity and Silence:

Experimental Evidence on Social Pressure and Free Speech*

Juan S. Morales† Margaret Samahita‡

September 24, 2025

Abstract

This paper studies the expression of public opinions in the context of social norms that induce conformity and self-censorship. In a series of pre-registered online experiments in the US in 2021 and 2023, we elicit participants' views on two controversial topics (gender and race) and their willingness to publish these views online. We document the existence of ideologically left-wing norms: participants with progressive views were more willing to publish, and awareness of potential publication reduced expressed conservatism. A priming information treatment, in which participants were informed about the potential negative backlash from social media posts, induced some conformity and silencing, but the results were weak and not statistically significant. A social information treatment, which informed respondents about high rates of others' willingness to speak up, significantly decreased self-censorship.

JEL Codes: D70, D83, P00, C90, Z13
Keywords: social media; spiral of silence; public opinion; cancel culture; free speech

*This paper was previously titled "Can Social Pressure Stifle Free Speech?". We thank Martin Bisgaard, Luca Braghieri, Leonardo Bursztyn, Eric Dickson, Edoardo Grillo, Caroline Le Pennec, Jean-Robert Tyran, Wieland Müller, Chris Roth, seminar and conference participants at the Burgundy School of Business, Collegio Carlo Alberto, Durham University, London School of Economics, Lund University, Trinity College Dublin, University College Dublin, University of Nottingham, University of Stirling, University of Vienna, Wilfrid Laurier University, Workshop on "Online Social Influence" (UCD), Workshop on "Directions of Polarization, Social Norms, and Trust in Societies" (MIT), 2022 APSA Meeting, 2022 ESA Special Meeting, 2022 SSES Annual Congress (Fribourg), 2022 ESA European Meeting (Bologna), 2023 Rebecca B. Morton Conference on Experimental Political Science (NYU), 2023 IMEBESS Meeting (Lisbon), 2023 EEA Meeting (Barcelona), and the 2023 CPEG Meeting (Queen's University) 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), and pre-registered as AEARCTR-0007905 (Samahita, 2021). Funding from UCD, the Collegio Carlo Alberto, LCERPA (WLU) and the Institute for Humane Studies (grant no. IHS017631) 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). Understanding when and why individuals choose to self-censor or conform is vital to diversity of thought and strong democratic institutions (Benson, 2024).

This paper studies how perceived social pressure affects the public expression of opinion. To do so, we conduct two pre-registered online experiments (total N=3,152) to study the relationship between social pressure and the public expression of opinions, in a period of time when the prevalent norms on social media tended to be liberal. 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" by sharing 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 public Twitter page dedicated to the experiment.

We also elicit perceived norms by asking participants to guess others' opinion regarding the stated questions and the opinion *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?".

Our surveys confirm that the prevailing norms at the time of our experiment matched the liberal/progressive views on both the race and gender topic: the proportion of individuals willing to post their views online was highest among those who held left-wing opinions. These patterns are consistent with self-censorship among those who held relatively more conservative views, who on average were less willing to post their opinions in our studies. This negative relationship between willingness to post and conservative views was starkest among Democratic and Independent voters (and weakest for Republicans). In contrast, Republicans expressed the most concern about freedom of speech being restricted, as elicited in our post-survey questions. Furthermore, participants perceive overall opinion as more right-leaning than their perception of the views of those willing to publish, suggesting that left-wing views are seen as more publicly accepted. Independents showed the strongest sense of distortion, measured by the gap between what they believe others' opinion is, relative to the opinion of those willing to publish. These descriptive findings contribute important and new stylized facts about the public expression of opinions within a polarized US political environment.

Our experimental interventions provide further insights into how social norms affect public speech. 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 allowed us to evaluate whether the prospect of their opinion being made public led to changes in the stated views' of the respondents, and in particular, whether they conformed to a more socially accepted view. Consistent with the descriptive findings of a liberal norm, participants in our Awareness treatment reported on average views which were more left-leaning. This inclination to conform was driven mainly by the behaviour of Independent voters.

In our second 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 participants to consider cancel culture and negative online backlash led to a modest (but often statistically insignificant) reduction in willingness to speak up. Independent voters were also the most sensitive to the backlash such that those further from the norm were 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 Participation** 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.

At the end of the experiment, we asked participants in an open text box about their reasons for choosing whether to speak up or not. Using a large language model (LLM) we analyze this text and classify these rationales into relevant categories. The analysis provides additional insights, as well as evidence consistent with our other findings. Liberals are less likely to express fear of social backlash as a reason not to post, moderates appear less interested in the issues and are more disengaged from social media, and conservatives are more likely to emphasize the importance of free speech as a reason to speak up.

Our paper makes several important contributions. Unlike prior studies focusing on university or academic contexts (Norris, 2023; Braghieri, 2024; Ho and Huang, 2024), we demonstrate that conformity to liberal social norms extended broadly into a more general US adult population (of online respondents), contributing new evidence to debates surrounding cancel culture. Second, we identify Independent voters as especially sensitive to social pressure, a group which plays a critical but often overlooked role in policy discussions and is increasingly important for electoral outcomes (Klar and Krupnikov, 2016). That Independent voters are particularly sensitive to social pressure may have implications for policy decision-making and democratic processes. Third, we document how willingness to speak up significantly increases when individuals learn that a large proportion of previous participants are also publicly expressing their views, highlighting the critical role of social cues in overcoming self-censorship.

Our work relates to the large body of literature on social image and reputational concerns (Bernheim, 1994; Bénabou and Tirole, 2006; Andreoni and Bernheim, 2009). Individuals often misreport their true opinion (Kuran, 1997; Braghieri, 2024; Valentim, 2024) or change their behavior (Andreoni and Petrie, 2004; Rege and Telle, 2004; Friedrichsen and Engelmann, 2018) in the presence of an audience due to social image concerns and a desire to conform to social norms. Fear of social stigma may also result in individuals self-censoring their opinion (Noelle-Neumann, 1974; Morales, 2020; Gibson and Sutherland, 2023; Bursztyn et al., 2023) or hiding behaviours that deviate from the perceived

social norm (DellaVigna et al., 2016; Carlson and Settle, 2016; Holm and Samahita, 2018; von Siemens, 2020). However, the specific role of cancel culture in shaping social norms and behavior remains relatively underexplored in the existing literature.

Social image concerns are particularly relevant in online settings, where backlash and moral outrage, often linked to cancel culture, can intensify these pressures (Brady et al., 2021; Crockett, 2017; Forestal, 2024). Recent work has documented how perceptions of cancel culture are highest for academics who hold minority opinions (Norris, 2023, 2025) but are generally exaggerated in the wider population (Dias, Druckman and Levendusky, 2025). These issues also relate to the broad debate regarding political correctness and the value of free speech (Morris, 2001; Voerman-Tam, Grimes and Watson, 2023). Our priming intervention, which exposes participants to consider cancel culture and the potential for online social backlash, allows us to study the extent to which this particular form of social pressure affects their willingness to speak up. We show that increased awareness of negative backlash on social media increases self-censorship only modestly.

Our work also contributes to the literature on political participation, particularly how social pressure and social cues influence individuals' willingness to engage in public discourse and activism. Social cues can affect reported attitudes, political donations, voting, and other behaviours (Bond et al., 2012; Perez-Truglia and Cruces, 2017; Bursztyn and Jensen, 2017; Conzo ⓡ et al., 2023; Isler and Gächter, 2022). In the context of dissent, observing others speak up can motivate individuals to voice their own opinions and participate in activism (González, 2020; Manacorda and Tesei, 2020), thereby eroding social norms (Morales, 2020; Bursztyn, Egorov and Fiorin, 2020; Álvarez-Benjumea, 2023; Albornoz, Bradley and Sonderegger, 2022; Dinas, Martínez and Valentim, 2024; Apffelstaedt, Freundt and Oslislo, 2022). Relatedly, Ho and Huang (2024) show that highlighting the presence of self-censoring individuals leads to higher levels of speaking up among those who hold dissenting views. At the same time, others' participation can also lead to free-riding and act as a strategic substitute for individual dissent (Cantoni et al., 2019; Hager et al., 2023). Our study shows that individuals are more likely to speak up when *many* others do so, even without knowing what views are being expressed. These findings complement evidence that political polarization distorts social behavior through norm-based mechanisms (Dimant, 2024), by showing how such norms influence individuals' willingness to express opinions publicly.

In the following section we explain the experimental design and our implementation. Section 3 details the results of our survey experiments. Section 4 documents empirical extensions. Section 5 concludes.

2 Experimental Design

2.1 Motivation and context

We study public expression amid online backlash and social image concerns during a period of prevalent discussions about cancel culture. The US provides a relevant setting given rising political polarization (Boxell, Gentzkow and Shapiro, 2024; Draca and Schwarz, 2024; Dimant, 2024) and an increase in self-censorship (Cato). Sensitive topics like race and gender norms amplify these pressures, especially on social media, where expression of opinion could carry social costs. To examine these issues, we conducted experiments in 2021 and 2023 (pre-analysis plans available here; timelines in Figures 2 and 3).

2.2 Main variables

Our outcomes of interest are reported attitudes (which we refer to as s) and participants' willingness to share their views on social media (v).

2.2.1 Elicitation of Attitudes (s)

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.

Responses range from *strongly disagree* to *strongly agree* (coded as 1-7).[1]

2.2.2 Elicitation of Willingness to Publish (v)

We asked two questions about sharing participants' responses on social media, in order:

1. **Posting anonymously:** "Would you be willing to let us post on social media, anonymously, your response to the previous statement:" (for example)

[1]Statements were pre-tested in a pilot together with other statements, see Appendix Figures A9 and A10. In Experiment 1, we also included a placebo statement—"People who have been vaccinated against COVID-19 should be allowed to travel without testing and quarantine requirement"—and found no treatment effect on this outcome.

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

Participants were informed that selecting Yes meant we would create a tweet containing the above and post it on a public Twitter page after data collection.[2]

2. **Posting with own name:** "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."*

Participants were informed 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-).[3]
- 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.

Full instructions are in Appendix D.

[2]Some anonymous responses were posted at https://twitter.com/SurveyResponses, but our tweeting program was later blocked by Twitter due to spam-like behavior (understandably).

[3]At the time, Prolific did not explicitly forbid the collection of personal data, stating: "There may be some other cases where your study design requires the collection of personal data. In this case, please get in touch with Prolific Support to discuss which approaches might be possible." (Link, accessed 2025-05-29).

Since we study the effect of perceived social cost on opinion expression, it was essential that participants seriously considered the possibility of posts being made public.[4] However, we did not want to actually publish names due to potential negative consequences for participants. In our aim of aligning with norms of no deception we truthfully informed participants that if they consented we would attempt to obtain their names, but publication was conditional on an event with an extremely low probability, as we then explained in the debrief (see Section 2.5 below). Previous requests to Prolific for participant names have been denied, as expected.

This design allowed us to measure real behavioral responses while avoiding deception and minimizing harm. The experiments were approved by two separate ethics committees.

2.3 Experimental interventions

We implemented three interventions designed to vary social pressure, or the cost of "speaking up" against a prevailing social norm. The **Prime** treatment primed participants on the risks of online backlash. The **Publication Awareness** treatment informed them that their responses could be published, increasing the saliency of social backlash. The **High Peer Participation (HiPeer)** treatment (and LoPeer as an active control) indicated whether a high/low percentage of previous participants agreed to publish. Below, we describe each treatment together with a brief discussion of our pre-registered predictions. Further details are available in the pre-registration.

2.3.1 Prime treatment

This intervention primed participants on the potential risk of social disapproval by displaying the following text and image:

[4]The Twitter page was new and unlikely to be seen by participants' followers, meaning any observed effect likely underestimates the true impact of online backlash, especially for participants with many followers. Consistent with this, we find heterogeneity in responses by social media use, with the strongest effect among Independents (Appendix Figure A6).

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.**

Figure 1: Prime treatment

The text, adapted from a New York Times article on cancel culture, highlights the risks of expressing personal opinions online, which is expected to reduce participants' willingness to publish. Experiment 2 included a weaker variant (WeakPrime, Figure 3) to test whether any effect is driven by the term "cancel culture" or the accompanying image.[5] We included an attention check, asking: "what does the text say cancel culture can result in?" Participants had to select "losing a job" before proceeding.

For the control text, participants saw information about University College Dublin (UCD) or the University of Turin (UniTo), with the university logo replacing the cancel culture image. Control texts were randomized, each with a corresponding attention check.

[5]Details in Appendix Section A.1. The effects of Strong (original) and Weak primes are not statistically different, so we pool them in the analysis. While there may be concerns that these two priming interventions could induce experimenter demand effect, it is unclear ex-ante which direction it would go and may depend on political affiliation. De Quidt, Haushofer and Roth (2018) show that typical demand effects are probably small. Similarly, Mummolo and Peterson (2019) find little evidence for demand effects in online survey experimental settings like ours. While our priming treatment cannot replicate real online backlash, it underscores the perceived threat. For ethical reasons, we do not expose participants to actual backlash.

We also vary the timing of the prime, as shown in Figure 2.[6]

2.3.2 Publication Awareness treatment

In a second treatment dimension of Experiment 1, we varied participants' awareness of the potential publication of their views—revealing opinions to the public (T4-T5) instead of just the experimenter (T1-T3). Publication awareness creates an additional channel of social pressure, as participants perceive public exposure of their views as carrying higher social costs, which in turn is expected to reduce their willingness to publish.

In Treatments 1–3 (*Publication Unaware*), attitudes and willingness to publish were elicited in separate stages. In Treatments 4–5 (*Publication Aware*), both were elicited within the same stage (on separate pages), allowing participants to revise their stated opinion after indicating their willingness to publish it. Additionally, participants in T4–T5 were already aware of the possibility of publication when responding to the *second* opinion statement. This design allows us to test whether awareness of potential public exposure affects the content of expressed views.

To examine interaction effects, T4 included the online backlash prime before attitude elicitation, while T5 used a control text instead.

2.3.3 High/Low Peer Participation treatment

In a second intervention of Experiment 2, we varied the information given about how likely previous participants were to share their opinions. Higher peer participation may signal lower social sanctions.[7] Additionally, in a collective action setting, expressing an opinion can be costly, leading to free-riding (strategic substitutes) or coordination benefits (strategic complements). The net effect is theoretically ambiguous (Cantoni et al., 2019; Hager et al., 2023).

Before asking about willingness to publish, participants saw:

> 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 was randomized into either the "LoPeer" (13%/7% for gender/race) or "HiPeer" (60%/67% for gender/race) treatment. These values were drawn from previous surveys

[6] **Treatment 1** shows the prime *before* attitude elicitation. **Treatment 2** shows the prime *after* attitude elicitation, to investigate whether individuals who revealed their attitudes under no prime were more sensitive to publishing ex-post (though no significant difference is found relative to Treatment 1).

[7] In our experimental design, we do not specify *what* views others are expressing, only the proportion of individuals willing to publish—so an implicit assumption is that the increase in public expression is broad. We elicit participants' beliefs about the views others express after the intervention.

in US states with similar high/low willingness to publish, making LoPeer an active control for HiPeer. Participants were debriefed on this at the end of the experiment.[8]

2.4 Additional questions

Value of publishing: We also measured participants' willingness to pay to publish, with their name, their view on one randomly chosen statement (Gender or Race). Participants received 10 tickets for a USD 100 lottery (they were informed of their odds, which depended on the wave and number of participants). If they agreed to publish, they were asked if they would give up 10, 5, or 1 ticket to not post, sequentially. If they declined all, WTP was coded as 0. Those who initially refused to publish were offered 1, 5, 20, or 50 additional tickets to reconsider. If they still declined, WTP was coded as -100.

Topic importance and knowledge: Participants also rated the importance of each of the two issues (1 = Not important, 5 = Extremely important). In Experiment 2, they also rated their topic knowledge (1 = Little to no knowledge, 5 = Expert).

Perceived norms: To measure perceived norms and perceived social distortions in public opinion, we asked about the average opinion of (i) all participants and (ii) those willing to publish without payment, for one randomly chosen statement. In Experiment 2, we instead elicited the *majority* opinion (Krupka and Weber, 2013). For instance, if participants perceive the overall opinion as more right-leaning than that of those willing to publish, they perceive that left-wing views are more publicly accepted. Correct answers earned 5 extra lottery tickets.[9]

Concerns about political correctness: We also asked participants five questions regarding their views on the political climate regarding online freedom of speech and social pressure. For example: "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) and "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).

[8] LoPeer: [13%/7%] of [Maryland/Indiana] participants; HiPeer: [60%/67%] of [Arizona/Oregon] participants.

[9] Results for average and majority opinion are similar (Appendix Table A20). Participants were unaware of alternative treatment arms.

Demographic questionnaire: In both experiments, we collected data on demographics, risk attitude, political preferences, and social media use, eliciting the latter two pre-treatment due to potential heterogeneity (Montgomery, Nyhan and Torres, 2018). Political preference was measured via self-placement on a left-right scale and party affiliation, while social media use was assessed by daily time spent on various platforms Facebook, Twitter, Instagram and other social media platforms.

In Experiment 2, we added the Hong psychological reactance scale (Hong and Faedda, 1996), how common participants think their first and last name are (7-point scale), and the likelihood of participants' opinions truly being posted on social media (7-point scale).[10] Participants also provided open-ended reasons for deciding whether to publishing or not, and could give general feedback.

2.5 Debrief

We conclude by debriefing participants on the purpose of the study and informing them that any opinions approved for anonymous publication would be posted on a public Twitter page. Regarding publication with names, we inform participants that: "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. 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." The full survey is in Appendix D.

2.6 Implementation

For Experiment 1, the first wave of data collection (Wave 1) took place in August 2021 via Prolific, recruiting 900 participants: 300 each from self-identified Democrats, Republicans, and Independents (including those with affiliation "Other" or "None").[11]

Wave 1 unintentionally oversampled young females due to Prolific-specific sampling issues.[12] To address this, Wave 2 (Nov 2021) recruited a US nationally representative

[10]Only 13% of participants believed that it was "extremely unlikely" that responses would end up being posted. This belief did not differ significantly between Democrats and Republicans. Results are not qualitatively different when excluding these participants, see Appendix Tables A11 and A15.

[11]We allocated 27.5% of participants to T1-2 and 15% to T3-5 to maximize power for our main tests. Using one-sided tests with 5% significance, we had 80% power to detect and effect size $\geq$ 0.20sd for our main test of H1 and and effect size $\geq$ 0.24sd for H2.

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

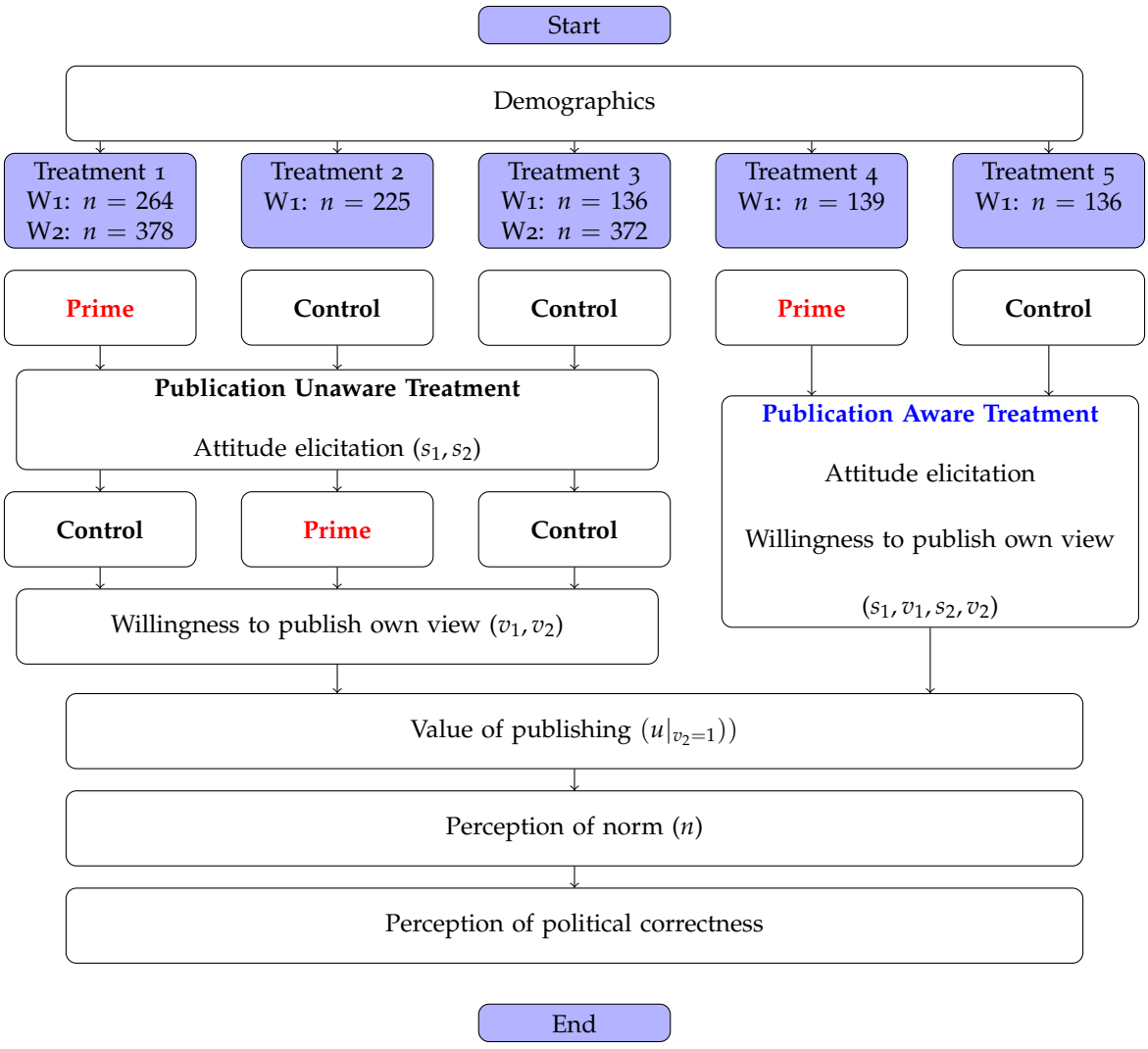


Figure 2: Experiment 1 timeline

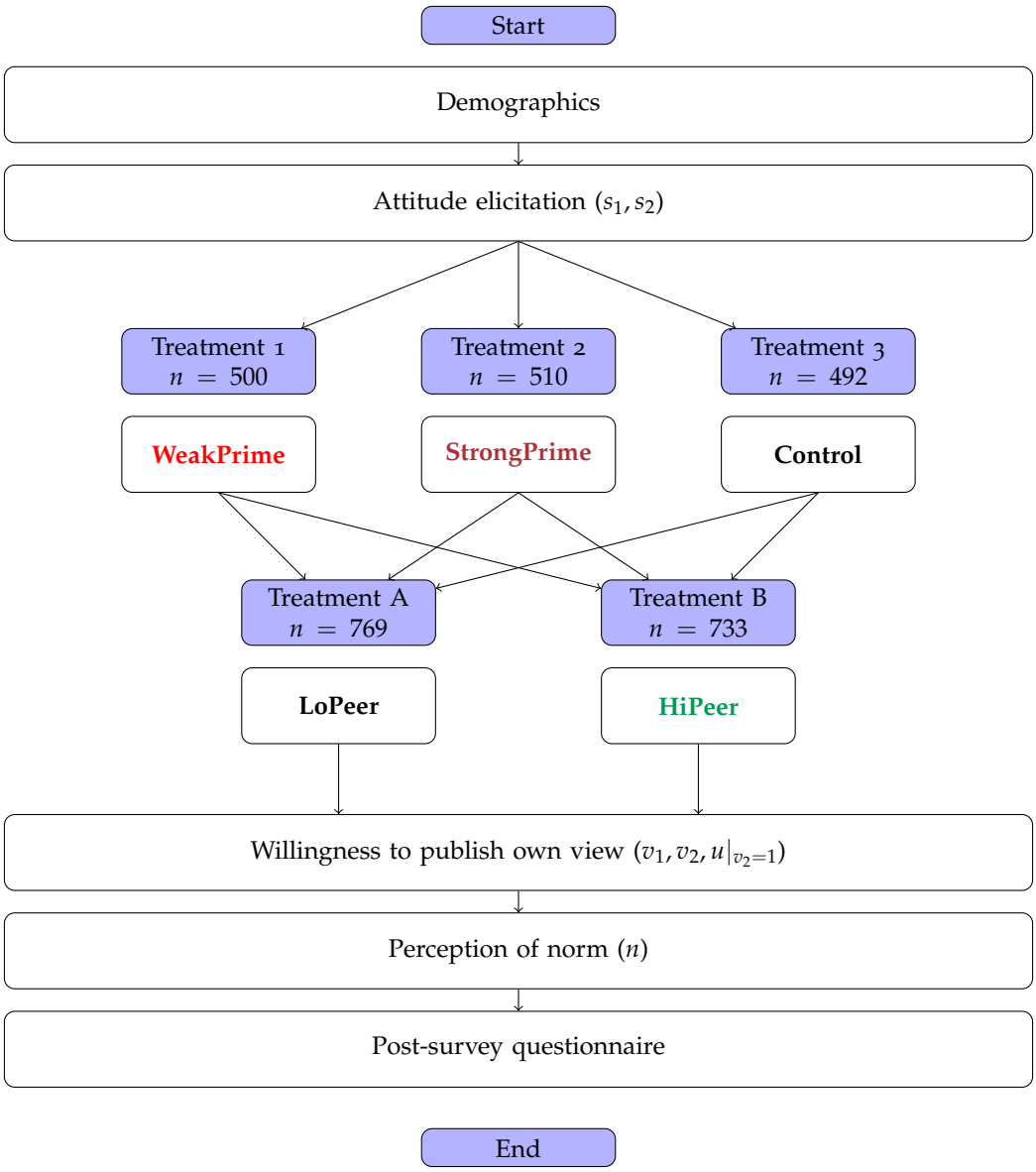


Figure 3: Experiment 2 timeline

sample in terms of age, gender and ethnicity (750 participants), with the following modifications: (i) running only Treatments 1 (primed) and 3 (control) with equal numbers due to budget constraints, (ii) using only the gender question as it resulted in more distinct norms across partisan groups, and (iii) adding the question on perceived likelihood of actual publication.

For Experiment 2, to study the effect of peer participation, Wave 3 (March 2023) recruited 1502 participants: 400 Democrats, 700 Independents (including unaligned), and 400 Republicans—oversampling Independents to examine a potential backlash effect found in Experiment 1 (but which was not confirmed by the second experiment). Participants were randomized evenly into a 3x2 design (StrongPrime/WeakPrime/Control x HiPeer/LoPeer).[13]

Key descriptive statistics are in Appendix Table A1.[14] As stated above, Wave 1 participants skew younger, more female, and more white than the nationally representative Wave 2. They are also more active on social media (80% spending $\geq$ 1 hour daily vs 55% in Wave 2). Waves 3 and 4 more closely resemble Wave 2 demographically, despite not being nationally representative.

Unless stated otherwise, we pool Waves 1–3 in all analyses. Although this decision was not pre-registered (we only pre-registered pooling Waves 1–2 of Experiment 1), it streamlines the presentation of results. Importantly, the main findings are similar when Experiments 1 and 2 are analyzed separately, with relevant tables cited below.

3 Results

3.1 Descriptive evidence

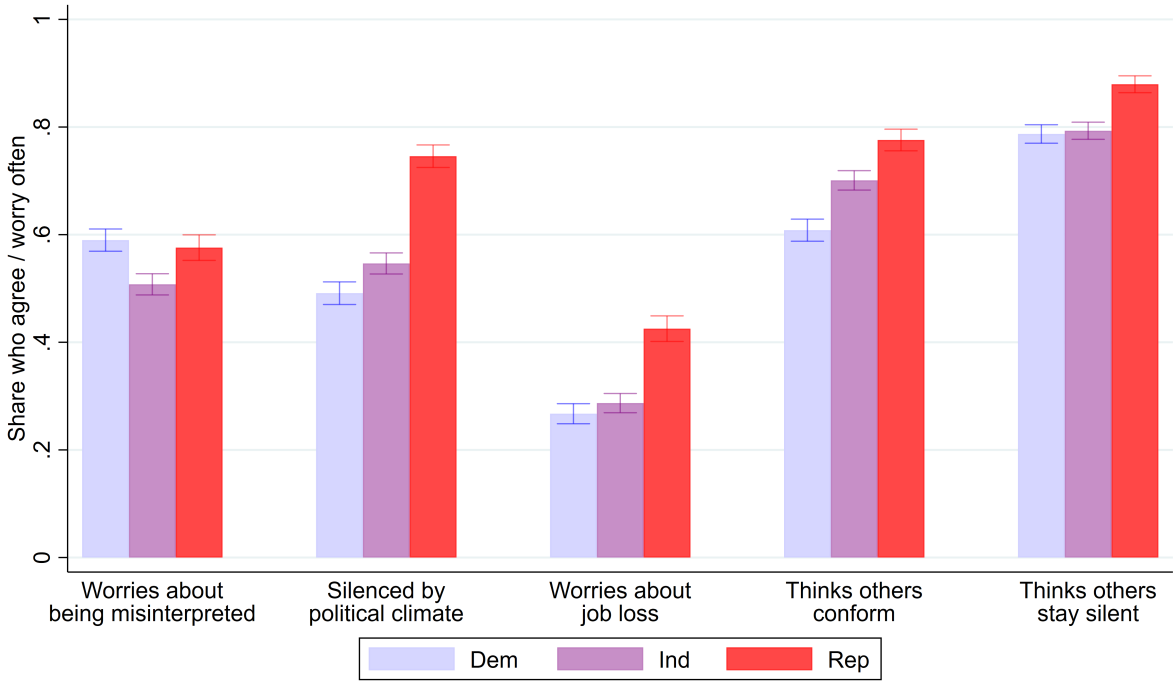
Perceptions of political correctness in our survey align with the broader narratives on popular media at the time: Republicans reported the highest concern about cancel culture and censorship. They were most likely to agree that the political climate prevents them from expressing their beliefs, feared job loss due to political opinions, and believed social pressure leads people to misrepresent their views or remain silent (Figure 4).[15]

[13]To investigate whether the above absence of backlash effect was due to the addition of the High/Low Peer Participation statement, we conducted Wave 4 as a robustness check which replicated Wave 3 with an additional control treatment where participants were uninformed about others' participation (as in Experiment 1). The results confirm those found in Wave 3. Additionally, the HiPeer treatment also increases speaking up relative to the control by 4.5pp though this is not statistically significant.

[14]Balance tables across treatments are in Appendix Tables A2-A4.

[15]Further descriptive evidence about perceptions of political correctness is presented in Appendix Section A.2.

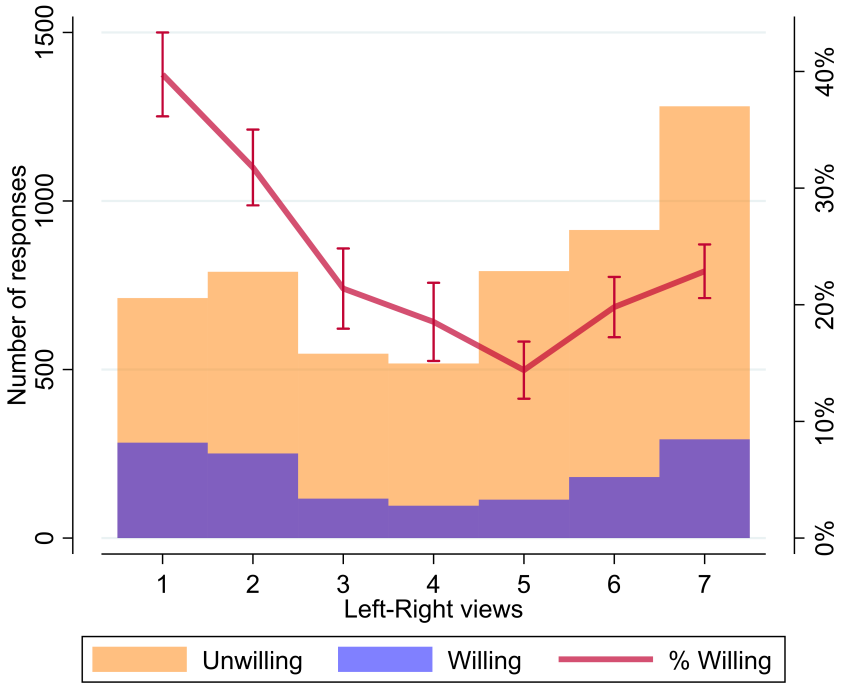
Figure 4: Perceptions of political climate and freedom of speech online



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?".

We next present descriptive evidence on attitudes toward gender and race topics and individuals' willingness to share their views online. Figure 5 shows the distribution of attitudes (s) on a 1-7 scale, where the gender topic was reverse-coded so that 1 represents more progressive views, split by willingness to publish opinions with names. These topics are very divisive, as revealed by a bimodal distribution of reported views.[16]

Figure 5: Public expression and reported attitudes



Notes: Agreement to statement in Experiments 1 and 2 (all participants 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. Bars represent 95% confidence intervals.

Only 24% of participants were willing to publish their views with their names. Despite a greater number of respondents holding conservative views on the gender and race issues elicited, these individuals were more likely to self-censor; potentially consistent with the presence of biased norms or pluralistic ignorance (Bursztyn and Yang, 2022; Bursztyn, González and Yanagizawa-Drott, 2020). In contrast, those strongly endorsing liberal positions were most willing to publish (about 40%). This pattern suggests that

[16] Appendix Figure A1 presents distributions by party, Appendix Figure A2 by topic, and Appendix Figure A3 by Hi/LoPeer treatment. Our pooled sample is not nationally representative and oversamples Independents, among whom conservative views are more prevalent.

social norms align with liberal views, with social pressure increasing for those expressing conservative opinions. Moreover, moderates were least willing to publish, consistent with research showing that extreme views are more likely to be expressed in social media (Bor and Petersen, 2022; Barberá and Rivero, 2015; Bail, 2022; Robertson, Del Rosario and Van Bavel, 2024).

Table 1 presents demographic correlates of willingness to speak up. Conservative views predict lower willingness to publish (row 1), even after controlling for covariates and party identification, suggesting this pattern is not solely driven by party-level differences in public expression. However, the relationship is not significant among Republicans, implying weaker social pressure to conform to liberal norms within this group.

Willingness to publish is positively associated with age, employment, risk tolerance, and social media use, but negatively associated with higher education. Women were less willing to publish than men. Compared to White individuals, Black participants were more willing to publish, while Asian participants were less willing. No significant differences were found across experimental waves (coefficients not shown).[17]

Result 1. *Conservative attitudes were more common in our sample, but those who held them were more likely to self-censor. In contrast, participants who reported left-leaning attitudes were more willing to publish their opinions—consistent with the presence of liberal social norms.*

3.2 Treatment effects on reported attitudes

We first analyze whether participants change their stated views (s) in response to the Prime and Awareness treatments in Experiment 1. We define s_{iq} as respondent i's reported stance (scaled from 0 to 1) for question q and estimate:

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

where $Prime_i$ equals 1 if subject i sees the online backlash prime before responding (T1 and T4), $Awareness_i$ equals 1 if the response is elicited when subjects are aware of potential publication (T4 and T5), and ε_{iq} is an individual-question specific error term. We include topic fixed effects δ_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

[17] Appendix Table A5 includes questions regarding fear of social backlash as additional (bad) controls. These explain 24 and 34 percent of the variation in willingness to speak up across left-right attitudes, for Democrats and Independents respectively (but none for Republicans). These findings are consistent with the idea that social pressure enforces liberal norms among the former groups. Similar patterns hold using our incentivized measure of willingness to pay to publish (Appendix Table A6).

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

	(1) All	(2) Dem	(3) Ind	(4) Rep
Left-Right opinion scale	-0.167***	-0.218***	-0.167***	-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)
Female	-0.049***	-0.085***	-0.016	-0.047
	(0.015)	(0.026)	(0.025)	(0.029)
Asian or Pacific Islander	-0.062**	-0.072	-0.059	-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	-0.009	0.097
	(0.031)	(0.055)	(0.047)	(0.070)
Other race	0.004	-0.128	-0.003	0.166
	(0.053)	(0.088)	(0.066)	(0.153)
College degree	-0.051***	-0.036	-0.088***	0.002
	(0.016)	(0.030)	(0.024)	(0.028)
Employed	0.045***	-0.027	0.056**	0.108***
	(0.016)	(0.031)	(0.025)	(0.028)
Risk attitude	0.075***	0.083***	0.083***	0.056***
	(0.008)	(0.013)	(0.013)	(0.013)
Active SM users	0.048***	0.030	0.045*	0.071**
	(0.016)	(0.028)	(0.025)	(0.029)
Democrat	0.018			
	(0.017)			
Republican	0.007			
	(0.018)			
Constant	0.255***	0.472***	0.232***	0.005
	(0.034)	(0.062)	(0.052)	(0.067)
N	5554	1802	2233	1519
R-sq	0.066	0.077	0.072	0.080

Notes: OLS regressions of willingness to publish with name (0/1) as outcome. *Left-Right opinion scale* is agreement to Gender or Race statement, coded on a 0-1 scale. *Risk attitude* is response to "Please tell us, in general, how willing or unwilling you are to take risks." (standardised). *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. Fixed effects for survey waves $\times$ topic included (not shown). Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

estimates (but are by construction uncorrelated to our randomized treatment).[18] Robust standard errors are clustered at the individual level (since we have two observations per subject).

Table 2 presents the results. In columns 1-2, we begin by comparing only individuals who were not exposed to the Prime condition (T2, T3, and T5). The results show that publication awareness shifts responses 0.051 points (out of 1) toward the left, suggesting that awareness of potential publication leads participants to conform to liberal norms.[19] The effect is driven by question 2 (columns 5-6), likely because respondents are explicitly informed of potential publication *after* answering question 1 (for which we do not observe a significant effect, columns 3-4).[20]

Table 2: Conformity in stated views (Experiment 1)

	All		Q1		Q2		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Awareness	-0.041	-0.051**	-0.019	-0.030	-0.061*	-0.073**	-0.041	-0.051**
	(0.030)	(0.023)	(0.035)	(0.027)	(0.035)	(0.031)	(0.030)	(0.022)
Prime							-0.025	-0.028
							(0.025)	(0.018)
Awareness x Prime							0.082*	0.084***
							(0.043)	(0.032)
Mean in control	0.575	0.575	0.576	0.576	0.579	0.579	0.575	0.575
N	994	994	497	497	497	497	1800	1800
R-sq	0.004	0.367	0.005	0.408	0.006	0.343	0.004	0.360
Topic FE	X	X	X	X	X	X	X	X
Controls		X		X		X		X

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

We then examine the effect of the Prime on stated opinions. We find weak effects: primed participants' responses shift slightly toward the liberal norm but not significantly

[18] All controls were pre-registered except for wave FEs (as we did not anticipate running multiple waves when we first pre-registered).

[19] Our effect size corresponds to 0.145 sd, which is in line with those reported in the meta-analysis of social recognition interventions by Goette and Tripodi (2024), specifically for studies with similar sample sizes.

[20] Only six subjects used the "Back" button to revise their initial response to the first question, suggesting that participants may worry about being perceived as misrepresenting their opinion or not care enough about their opinion being published to lose a few seconds to change their answer (since most subjects choose not to publish anyway); but that when faced with the second question and now aware of the possibility of publication, they move closer to the liberal position.

(row 2, columns 7-8).[21] While we expected the effect of the prime to be more negative when individuals are asked about their stated opinion with publication in mind ($\theta_3 < 0$), the interaction coefficient is instead significant in the opposite direction: responses move further to the right. One possible explanation is that increased salience of cancel culture, for those who are already in the mindset of publishing and thus anticipate public scrutiny, increases sensitivity to free speech concerns. This interaction may lead participants to resist self-censorship. The overall effect of Prime and Publication Awareness combined ($\theta_1 + \theta_3$, rows 1 and 3) is not statistically different from zero.

Result 2. *Publication awareness leads participants to report less conservative views, while the online backlash prime has insignificant effects.*

3.3 Treatment effects on public expression

We next study individuals' willingness to publish their opinion with their name (v). We pool our studies and first estimate:

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

where v_{iq} is a binary variable equal to 1 if subject i is willing to publish their opinion on question q with their name (without additional lottery tickets as incentive). The parameters θ_1, θ_2 and θ_3 capture the average treatment effect of our interventions across all subjects. Since Publication Awareness was only part of Experiment 1, and HiPeer was only part of Experiment 2, we impute a value of 0 for observations with missing treatments and include study wave fixed effects in all specifications. We also include topic fixed effects δ_q. In some specifications we include additional demographic controls.

The results (Table 3) show that, on average, neither the online backlash Prime nor the Publication Awareness treatment significantly affect willingness to publish.[22] However, the HiPeer treatment—in which respondents learn that a high share of previous participants published their opinions—has a statistically significant positive effect, increasing willingness to speak up by about 5 percentage points relative to those in the LoPeer group (and shown in Appendix Figure A3 by reported attitude).[23]

[21] Results focusing on the main effect of the prime (without interaction) are provided in Appendix Table A7, similarly showing it has no significant treatment effect.

[22] Results for Experiments 1 and 2 separately are shown in Appendix Tables A8 and A9.

[23] This effect is smaller than in similar interventions, such as Bursztyn et al. (2023), where social cover increased public expression by 11 percentage points among progressives. Additionally, we pre-registered the WTP to publish as an alternative outcome variable. The results, displayed in Appendix Table A10, are very similar to the main results presented above, though less statistically precise.

Table 3: Willingness to publish and experimental interventions

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	-0.022	-0.015	-0.032	-0.028	-0.013	-0.010	-0.020	-0.011
	(0.016)	(0.015)	(0.028)	(0.027)	(0.025)	(0.024)	(0.029)	(0.028)
HiPeer	0.054***	0.043**	0.058	0.061*	0.038	0.031	0.077**	0.057
	(0.020)	(0.020)	(0.039)	(0.037)	(0.030)	(0.030)	(0.038)	(0.038)
Awareness	0.002	0.007	0.005	-0.005	0.040	0.043	-0.029	-0.001
	(0.030)	(0.028)	(0.054)	(0.054)	(0.050)	(0.047)	(0.048)	(0.044)
Mean in control	0.235	0.235	0.280	0.280	0.219	0.219	0.194	0.194
N	5554	5554	1802	1802	2233	2233	1519	1519
R-sq	0.004	0.063	0.011	0.084	0.003	0.083	0.007	0.084
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 leaning and its square, 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 explore three pre-registered empirical extensions. First, we interact the main Prime treatment with HiPeer to examine whether the effect of high peer participation found above varies with social pressure. Second, we interact Prime with Awareness to test whether the Prime's effect is attenuated when opinions may adjust due to potential publication. Third, we test whether social pressure's negative effect on public expression intensifies with greater ideological distance from the norm by interacting Prime with the Left-Right opinion scale.

The extended model is:

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

where s_{iq} is the Left-Right opinion of individual i for question q, defined as agreement to each of the Gender (reverse-coded) and Race statements, and scaled to 0-1.

The results are shown in Table 4. While the HiPeer treatment generally increases willingness to publish (row 2, $\theta_2 > 0$), this effect is 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.

Table 4: Willingness to publish and experimental interventions

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	0.018	0.017	-0.060	-0.049	0.084	0.092*	0.111	0.045
	(0.032)	(0.031)	(0.046)	(0.045)	(0.051)	(0.049)	(0.092)	(0.087)
HiPeer	0.091***	0.087***	0.048	0.054	0.099*	0.105**	0.138**	0.119**
	(0.034)	(0.033)	(0.060)	(0.057)	(0.052)	(0.052)	(0.064)	(0.060)
Awareness	-0.016	-0.005	-0.024	-0.028	0.028	0.046	-0.031	0.000
	(0.040)	(0.039)	(0.076)	(0.075)	(0.064)	(0.061)	(0.068)	(0.062)
Prime x HiPeer	-0.056	-0.065*	0.009	0.007	-0.080	-0.105*	-0.094	-0.096
	(0.038)	(0.037)	(0.068)	(0.065)	(0.058)	(0.058)	(0.074)	(0.070)
Prime x Awareness	0.044	0.029	0.047	0.039	0.054	0.026	0.017	0.006
	(0.052)	(0.050)	(0.097)	(0.097)	(0.088)	(0.085)	(0.085)	(0.079)
Prime x LR scale	-0.053	-0.031	0.048	0.041	-0.138**	-0.132*	-0.137	-0.041
	(0.042)	(0.041)	(0.075)	(0.073)	(0.069)	(0.067)	(0.101)	(0.096)
N	5554	5554	1802	1802	2233	2233	1519	1519
R-sq	0.021	0.077	0.027	0.102	0.024	0.094	0.014	0.090
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 leaning and its square, social media use, and wave FE $\times$ topic FE. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We hypothesized that the online backlash Prime would have a stronger negative effect when individuals are unaware of potential publication, as those in the Publication Awareness treatment could adjust s to conform. Thus, the Prime's effect should be weaker in the latter ($\theta_5 > 0$). While the coefficient is positive (row 5), the effect is not statistically significant.

We also hypothesized that the negative effect of the Prime on public expression would increase with ideological distance from the norm ($\theta_6 < 0$). The coefficient is in the expected direction (row 6) but is not statistically significant. However, among Independent voters, the interaction between the Left-right scale (s) and the Prime is statistically significant and negative (columns 5 and 6).[24] Note that there is more variation in s for Independent voters (see Appendix Figure A1), while they may perceive higher costs from expressing dissenting views. Additionally, we observe a weak but positive main effect

[24]Political subgroup analyses were pre-registered as exploratory.

of the Prime for Independents at the left-most position, i.e. $s_i = 0$, perhaps consistent with increased perceived importance of speaking up. These findings suggest that social pressure may be particularly relevant for Independent voters, who appear to conform to left-wing speech norms.[25] The findings also support the idea that Independents may be especially concerned with social image (Klar and Krupnikov, 2016).

While our experimental design does not pinpoint the exact mechanism behind the HiPeer treatment's effect on willingness to publish, we explore potential explanations in Section 4.1 below.

Result 3. *Both public awareness and the online backlash prime have insignificant effects on public expression. High peer participation reduces self-censorship, but this effect is dampened by the online backlash prime (potentially as higher perceived costs discourage dissent).*

3.4 Heterogeneous treatment effects

We explore whether the effect of our main Prime treatment was heterogeneous along various dimensions with specifications of the following form:[26]

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

where Var_i indicates the dimension of interest. In Figure 6 we report the coefficients θ_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 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).

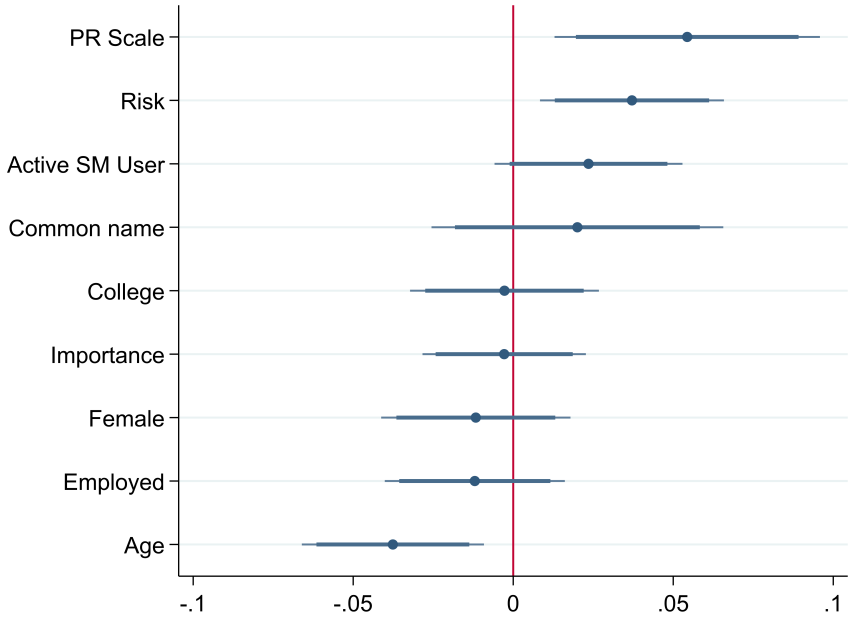
We find statistically significant heterogeneity along reactance, risk attitudes and age. Participants with higher reactance are more likely to backlash against the Prime treatment, by increasing their willingness to speak up, while those with lower reactance are more likely to self-censor as a result. 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. That these

[25] We replicate this analysis for willingness to pay in Appendix Table A14. To address potential endogeneity in reported attitudes (s), we exclude treatments where attitudes are measured post-treatment. The results are very similar and shown in Appendix Table A16.

[26] Heterogeneity analyses for the Publication Awareness and HiPeer treatments are shown in Appendix Figures A4 and A5 respectively.

types of less/more inhibited individuals respond differentially to the prime treatment is perhaps not surprising, but suggests that public views may be disproportionately shaped by vocal minorities who are less constrained by social repercussions.[27]

Figure 6: Heterogeneity of Prime treatment



Notes: The figure shows the coefficient θ_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:* 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 leaning and its square, social media use, and wave FE $\times$ topic FE. Robust standard errors are clustered at the individual level.

[27] In Appendix Figure A6 we report the results separately by political party.

4 Empirical Extensions

4.1 Mechanisms in HiPeer treatment

The positive effect of peer participation (captured in our HiPeer treatment) is likely to arise from a shift in the perception about the strength of social pressure. Nevertheless, it is also possible that the HiPeer treatment increases willingness to speak up through other mechanisms, such as changes to the perceived norm. In this section, we explore these possibilities in our data. The analyses below were not part of our pre-registration so they should be considered exploratory.

First, the finding that the HiPeer treatment increases speaking up is concentrated amongst those not receiving the Prime, and neutralised by the Prime treatment. Above we speculated that this is driven by those far from the norm, who are more likely to follow others by speaking up when the perceived cost (relative to the LoPeer group) is low but free-ride when the perceived cost increases (when primed). As further suggestive evidence for this idea, in Appendix Table A22 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 A23). Overall, these findings support our interpretation that the effect of HiPeer is due to a perception of lower social sanctioning, for those who hold more right-wing views.

To explore whether the HiPeer treatment affects the perceived norm, we can use our data on participants' beliefs about the opinion of others. Specifically, we use our measures of n_{all} (opinion for everyone) and n_{pub} (opinion for those who publish) as described above and transform these to a 0-1 left-right attitude scale. We regress these outcomes on the HiPeer and Prime treatment dummies and their interaction and results are shown in Appendix Table A24. As shown in columns 1-4, the HiPeer treatment does not significantly affect the estimated opinion for all other fellow participants. However, knowing that many others speak up shifts the estimated view for participants who publish to the right (columns 5-8), and this perception shift is larger for Democrats. Consequently, the perception of public views being distorted, measured by the gap $n_{all} - n_{all}$, is significantly lower in the HiPeer treatment (columns 9-12).

At the same time, knowing many others speak up may also change participants'

perception of the political climate. To explore this possibility, we use responses to the post-experiment 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?". We regress standardised responses to these questions on the HiPeer and Prime treatment dummies and their interaction, results are shown in Appendix Tables A25 and A26.

While the HiPeer treatment does not significantly change perception about others conforming or staying silent, it lowers participant's perception of being silenced by the political climate, which may also contribute to increased willingness to speak up. However, this effect is neutralized when participants are primed, consistent with our baseline result that participants are not as willing to follow their peers in speaking up when social sanctions are relatively more salient.

Overall, the HiPeer treatment increases willingness to speak up mainly by lowering perceived social costs and shifting perceptions of public norms, but these effects vanish when concerns about backlash are made salient.

4.2 Open text analysis

At the end of Experiment 2, we asked participants in an open text box about their reasons for (agreeing to) posting, or not, their opinions on Twitter. We classify these responses using LLM, as outlined in Appendix Section A.3. We then analysed the classified responses, and the results are presented below. Although not pre-registered, this analysis provides additional insight into how social pressure influences public expression in our survey.

Descriptive evidence of reasons for (not) posting

Appendix Figure A7 shows the categorization of reasons given by participants for posting/not posting opinions across the ideological spectrum of the stated opinion.We observe some key empirical patterns which are consistent with our results above. First, among those not posting, fear of backlash is more prevalent on the right—consistent with the view that left-wing norms moderate expression. Second, freedom of speech is more frequently cited by those on the right as a reason for posting. Third, viewing the

issue as important is a common reason among individuals at the ends of the ideological distribution for posting their views (somewhat more on left), and conversely, indifference to the issues appears as a relatively more common reason among moderates for not posting. This final observation highlights one important driver of public expression: the degree to which individuals care about these particular issues.

Regression analysis

Appendix Table A28 shows OLS regressions of the likelihood of stating a particular reason for posting/not posting on our treatment indicators and the ideological spectrum as coded by participants' opinions. Being indifferent to social pressure is more likely to be a reason for posting for those in the HiPeer group, who were informed that many others were also willing to post. For those unwilling to post, fear of backlash is more likely to be stated as participants' views move to the right and away from the norm, consistent with our other findings. Interest in and importance of the race and gender issues are less frequent for moderates (consistent with our survey evidence on topic importance), and somewhat less salient to those in the HiPeer group.

Appendix Table A29 shows results for the other reasons. Freedom of speech is more likely mentioned for those in the centre and more right-wing participants, though the relationship becomes statistically insignificant with controls. Those on the right are less likely to cite privacy concerns but more likely to care about financial incentives. They are also less likely to cite disengagement from social media as a reason for not posting.

Keyword analysis

Appendix Figure A8 shows the proportion of respondents mentioning certain keywords related to our specific research hypotheses, and how these vary across the left-right ideology scale. Keywords related to social backlash are more common among right-leaning participants, keywords related to race or gender tend to be least common among moderate participants (and most common among left-leaning participants), and keywords related to other people increase with left-right ideology.

Treatment effects mediators

Appendix Tables A30 and A31 regress the number of statements published (0, 1 or 2) on our main treatment indicators (HiPeer or Prime) in Experiment 2, and controling in turn for each possible reason to post/not to post. This exercise includes each reason as "bad control" in a baseline regression framework to understand whether specific reasons

may mediate the treatment effect. The effect of the HiPeer treatment becomes statistically insignificant when controling for "indifference to social pressure", and when controling for "privacy concerns". The effect of the Prime treatment is attenuated when controling for "indifference to social pressure" and "freedom of speech".

Altogether, our open text analyses provide evidence consistent with our other findings. Social backlash is most salient for those on the right, however the HiPeer treatment reduces perception of social pressure, while the Prime has the opposite effect. Importantly, willingness to speak up is shaped not only by fear of backlash but also by how much individuals value the issues.

5 Conclusion

Our paper offers new insights into public expression in the context of growing social and political polarization, contributing to our understanding of how social pressure and online backlash can shape public discourse. We find that a substantial share 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 concerns about restricted freedom of speech is notably more pronounced among Republicans. Our findings revealed the presence of liberal or left-leaning 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 negative social backlash had only modest effects on individuals' willingness to speak up (possibly because cancel culture was already widely discussed in the media). However, the prime treatment increased self-censorship among Independents with conservative views, suggesting that this group may be particularly sensitive to these norms of expression.

Free speech is a key component of liberal democracies, and many participants in our experiment reported fear of repercussions as a reason not to voice their views publicly. At the same time, for better or worse, strong norms which induce self-censorship can unravel quickly (Morales, 2020; Bursztyn, Egorov and Fiorin, 2020; Álvarez-Benjumea, 2023; Albornoz, Bradley and Sonderegger, 2022; Apffelstaedt, Freundt and Oslislo, 2022). We provide evidence that informing people about others' willingness to speak up can reduce self-censorship, suggesting that simple interventions may increase peoples' willingness to share their views and shift perceptions of these norms. Interestingly, this result was present despite participants not being informed about *what* others were sharing, sug-

gesting that a norm of free speech more generally may play a role in shaping online behaviour, and not just a norm of what constitutes acceptable speech.

While our studies were conducted at a time when liberal norms were especially salient in both social and mainstream media, the pressure to conform to group norms is a general phenomenon that is likely to persist—even as the specific norms themselves shift. Moreover, understanding how social pressure and group dynamics shape online public expression, and how this in turn influences discourse on sensitive social and political issues, remains an important topic of study in a persistently polarized media environment.

Declaration

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work the authors used ChatGPT in order to improve the readability and language of some sentences in the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for truth-telling." *Econometrica*, 87(4): 1115–1153.

Albornoz, Facundo, Jake Bradley, and Silvia Sonderegger. 2022. "Updating the social norm: The case of hate crime after the Brexit referendum." Working Paper.

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

Andreoni, James, and B Douglas Bernheim. 2009. "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects." *Econometrica*, 77(5): 1607–1636.

Andreoni, James, and Ragan Petrie. 2004. "Public goods experiments without confidentiality: A glimpse into fund-raising." *Journal of Public Economics*, 88(7-8): 1605–1623.

Apffelstaedt, Arno, Jana Freundt, and Christoph Oslislo. 2022. "Social norms and elections: How elected rules can make behavior (in) appropriate." *Journal of Economic Behavior & Organization*, 196: 148–177.

Bail, Chris. 2022. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Barberá, Pablo, and Gonzalo Rivero. 2015. "Understanding the political representativeness of Twitter users." *Social Science Computer Review*, 33(6): 712–729.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review*, 96(5): 1652–1678.

Benson, Jonathan. 2024. "Democracy and the epistemic problems of political polarization." *American Political Science Review*, 118(4): 1719–1732.

Bernheim, B Douglas. 1994. "A theory of conformity." *Journal of Political Economy*, 102(5): 841–877.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature*, 489(7415): 295–298.

Bor, Alexander, and Michael Bang Petersen. 2022. "The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis." *American Political Science Review*, 116(1): 1–18.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro. 2024. "Cross-country trends in affective polarization." *Review of Economics and Statistics*, 106(2): 557–565.

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.

Braghieri, Luca. 2024. "Political correctness, social image, and information transmission." *American Economic Review*, 114(12): 3877–3904.

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 David Y Yang. 2022. "Misperceptions about others." *Annual Review of Economics*, 14(1): 425–452.

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): 20563051231177899.

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

DellaVigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao. 2016. "Voting to tell others." *The Review of Economic Studies*, 84(1): 143–181.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bounding experimenter demand." *American Economic Review*, 108(11): 3266–3302.

Dias, Nicholas C, James N Druckman, and Matthew S Levendusky. 2025. "Unraveling a "Cancel Culture" Dynamic: When, Why, and Which Americans Sanction Offensive Speech." *The Journal of Politics*, 87(2): 588–600.

Dimant, Eugen. 2024. "Hate trumps love: The impact of political polarization on social preferences." *Management Science*, 70(1): 1–31.

Dinas, Elias, Sergi Martínez, and Vicente Valentim. 2024. "Social norm change, political symbols, and expression of stigmatized preferences." *The Journal of Politics*, 86(2): 488–506.

Draca, Mirko, and Carlo Schwarz. 2024. "How polarised are citizens? Measuring ideology from the ground up." *The Economic Journal*, 134(661): 1950–1984.

Forestal, Jennifer. 2024. "Social Media, Social Control, and the Politics of Public Shaming." *American Political Science Review*, 118(4): 1704–1718.

Friedrichsen, Jana, and Dirk Engelmann. 2018. "Who cares about social image?" *European Economic Review*, 110: 61–77.

Gibson, James L, and Joseph L Sutherland. 2023. "Keeping your mouth shut: Spiraling self-censorship in the United States." *Political Science Quarterly*, 138(3): 361–376.

Goette, Lorenz, and Egon Tripodi. 2024. "The limits of social recognition: Experimental evidence from blood donors." *Journal of Public Economics*, 231: 105069.

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.

Holm, Hakan J, and Margaret Samahita. 2018. "Curating social image: Experimental evidence on the value of actions and selfies." *Journal of Economic Behavior & Organization*, 148: 83–104.

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

Ho, Yuen, and Yihong Huang. 2024. "Breaking the spiral of silence." Working Paper.

Isler, Ozan, and Simon Gächter. 2022. "Conforming with peers in honesty and cooperation." *Journal of Economic Behavior & Organization*, 195: 75–86.

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.

Michaeli, Moti, and Daniel Spiro. 2017. "From peer pressure to biased norms." *American Economic Journal: Microeconomics*, 9(1): 152–216.

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.

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review*, 113(2): 517–529.

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

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

Norris, Pippa. 2025. "Cancel culture: Heterodox self-censorship or the curious case of the dog which didn't bark." *International Political Science Review*, 46(3): 422–441.

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.

Rege, Mari, and Kjetil Telle. 2004. "The impact of social approval and framing on cooperation in public good situations." *Journal of Public Economics*, 88(7-8): 1625–1644.

Robertson, Claire E, Kareena S Del Rosario, and Jay J Van Bavel. 2024. "Inside the funhouse mirror factory: How social media distorts perceptions of norms." *Current Opinion in Psychology*, 60: 101918.

Valentim, Vicente. 2024. "Political stigma and preference falsification: theory and observational evidence." *The Journal of Politics*, 86(4): 1382–1402.

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.

von Siemens, Ferdinand A. 2020. "I care what you think: Social image concerns and the strategic revelation of past pro-social behavior." *Journal of the Economic Science Association*, 6: 43–56.

A Appendix (For Online Publication)

A.1 Weak and Strong Prime treatments

In Experiment 2 (see Figure 3), 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 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.

A.2 Concerns and perceptions about political correctness

In this section, we provide additional results on participants' concerns and perceptions about political correctness. Using responses to the five questions on political correctness, we construct an index of concern for freedom of speech using the first principal component. Concern is higher among conservatives (on Left-Right scale), Republicans, active social media users, college graduates, individuals of Asian ethnicity, and those more willing to take risks (Appendix Table A17). In Appendix Table A18, we regress this index on our experimental interventions. Our Prime treatment reduces concern among Democrats (perhaps due to perceptions that the issue is overstated) but increases concern among Republicans.

We also assess perceptions of political correctness through participants' beliefs about others' views. We define the *perceived gap* as $n_{all} - n_{pub}$, where n_{all} represents a participant's belief about the average (or majority, in Experiment 2) view of all participants in

the wave, and n_{pub} reflects their belief about the views of those willing to publish. On average, participants accurately estimate the overall norm (n_{all} 4.390 vs. the true value of 4.396), with Independents being closest (4.475). Democrats perceive views as more left-leaning (4.130), while Republicans perceive them as more right-leaning (4.604).[28]

Participants correctly recognize that publicly expressed views are more progressive than the true norm, with Independents perceiving the largest gap. This supports our finding of left-leaning norms on these topics. Appendix Table A21 presents correlations between the perceived gap and demographic factors. Perceived censorship increases for participants with more conservative views. However, conditional on views, Republicans perceive the smallest censorship gap despite expressing the highest concerns about free speech and the political climate.

A.3 LLM classification methodology

In turn, we provided ChatGPT (version 4.0) with the sample of responses to the two open text questions, "Why did you decide to let us post one or more of your opinions on social media?" and "Why did you decide not to let us post any of your opinions on social media?", and asked it to create a broad categorization of rationales provided by the participants. Based on our study and our own reading of the reasons provided by participants, we fine-tuned and aggregated the categorization into five main types of reason for each action (post or not post), including "Other". Then, we used the OpenAI API to classify each response using the prompts below:

> In a study we conducted, we asked participants whether they would allow us to post their opinion and name on potentially controversial topics (about gender and race) on social media. If they said yes, we asked them for a reason. Categorize the following reason for saying yes –*text*– into one of these: 1. Indifference to Social Pressure: Lack of concern for public reaction or social backlash, not afraid or ashamed of sharing their opinion. 2. Importance of Issues: Reflects strong beliefs about these social issues (gender and race), regardless of stance. 3. Freedom of Speech: Valuing the right and importance of free speech. 4. Financial Incentives: Rewards and incentives from being part of a study and potential financial gain. 5. Other. In your output, write only the number corresponding to the category, no text.
>
> In a study we conducted [...] If they said no, we asked them for a reason. Categorize the following reason for saying no –*text*– into one of these: 1. Privacy

[28] See Appendix Tables A19 and A20 for pooled results and breakdowns by wave/topic, respectively.

and Security Concerns: Concern about personal information and privacy and a desire for minimal digital presence. 2. Fear of Social or Professional Backlash: Worry about negative consequences from opinion being public or being misrepresented. 3. Disinterest or Lack of Confidence in the Issues: Reflects indifference or lack of knowledge about social issues (gender and race). 4. Disengagement from Social Media: Reflects low engagement or activity on social platforms due to personal preference and lack of interest. 5. Other. In your output, write only the number corresponding to the category, no text.

We used both GPT 4.0 and GPT 3.5-turbo to classify each of the statements using these prompts. Examples of participants' classified responses are shown in Appendix Table A27.

A.4 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.[29] 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:

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

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

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

either negative or positive feedback, and $negativeFeedback_u$ indicates whether user u received positive or negative feedback.[30] We also include user (γ_u) and time (γ_t) fixed effects.

Our coefficients of interest β'_w are normalized relative to the week before the comments and reported in Appendix Figure A11. 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.

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

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 leaning	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 leaning	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 leaning* 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: Balance table for Awareness treatment

	Unaware		Aware		Difference		
Variable	Mean	SD	Mean	SD	Diff	SE	N
Age	28.707	(9.764)	27.956	(8.690)	-0.751	(0.684)	900
Female	0.734	(0.442)	0.760	(0.428)	0.026	(0.032)	900
Asian or Pacific Islander	0.035	(0.184)	0.062	(0.241)	0.027*	(0.015)	900
Black or African American	0.094	(0.293)	0.073	(0.260)	-0.022	(0.020)	900
Hispanic or Latino	0.083	(0.276)	0.076	(0.266)	-0.007	(0.020)	900
White	0.765	(0.424)	0.771	(0.421)	0.006	(0.031)	900
College degree	0.630	(0.483)	0.695	(0.461)	0.064*	(0.034)	900
Employed	0.715	(0.452)	0.735	(0.442)	0.019	(0.032)	900
Risk attitude	6.315	(1.872)	6.415	(1.750)	0.099	(0.133)	900
Political leaning	4.445	(2.883)	4.636	(2.917)	0.192	(0.209)	900
Active SM users	0.797	(0.403)	0.822	(0.383)	0.025	(0.029)	900
Observations	625		275		900		

Notes: The Awareness treatment was only implemented in Wave 1. *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 A3: Balance table for Prime treatment

	No prime		Prime		Difference		
Variable	Mean	SD	Mean	SD	Diff	SE	N
Age	36.575	(14.006)	36.979	(13.496)	0.404	(0.508)	3,152
Female	0.548	(0.498)	0.569	(0.495)	0.021	(0.018)	3,152
Asian or Pacific Islander	0.068	(0.251)	0.055	(0.228)	-0.013	(0.009)	3,152
Black or African American	0.100	(0.301)	0.084	(0.278)	-0.016	(0.011)	3,152
Hispanic or Latino	0.061	(0.239)	0.064	(0.245)	0.003	(0.009)	3,152
White	0.755	(0.430)	0.773	(0.419)	0.018	(0.016)	3,152
College degree	0.630	(0.483)	0.654	(0.476)	0.023	(0.018)	3,152
Employed	0.739	(0.440)	0.742	(0.438)	0.004	(0.016)	3,152
Risk attitude	5.671	(2.286)	5.571	(2.256)	-0.099	(0.084)	3,152
Political leaning	4.248	(2.928)	4.347	(2.917)	0.098	(0.108)	3,152
Active SM users	0.596	(0.491)	0.578	(0.494)	-0.018	(0.018)	3,152
Observations	1,136		2,016		3,152		

Notes: *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 A4: Balance table for HiPeer treatment

	LoPeer		HiPeer		Difference		
Variable	Mean	SD	Mean	SD	Diff	SE	N
Age	42.215	(13.353)	41.932	(13.490)	-0.283	(0.693)	1,502
Female	0.507	(0.500)	0.464	(0.499)	-0.043*	(0.026)	1,502
Asian or Pacific Islander	0.062	(0.242)	0.059	(0.235)	-0.004	(0.012)	1,502
Black or African American	0.085	(0.278)	0.056	(0.230)	-0.029**	(0.013)	1,502
Hispanic or Latino	0.052	(0.222)	0.059	(0.235)	0.007	(0.012)	1,502
White	0.775	(0.418)	0.801	(0.400)	0.026	(0.021)	1,502
College degree	0.632	(0.483)	0.621	(0.486)	-0.011	(0.025)	1,502
Employed	0.743	(0.438)	0.752	(0.432)	0.009	(0.022)	1,502
Risk attitude	4.957	(2.422)	5.265	(2.432)	0.308**	(0.125)	1,502
Political leaning	4.515	(2.952)	4.415	(2.932)	-0.100	(0.152)	1,502
Active SM users	0.471	(0.499)	0.468	(0.499)	-0.003	(0.026)	1,502
Observations	769		733		1,502		

Notes: The HiPeer treatment was only implemented in Wave 3. *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 A5: Willingness to publish and fear of social backlash

	Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)
Left-Right opinion scale	-0.218***	-0.165***	-0.172***	-0.113***	-0.069	-0.071
	(0.035)	(0.036)	(0.033)	(0.035)	(0.045)	(0.045)
Misinterpret		-0.027***		-0.019***		-0.024***
		(0.008)		(0.006)		(0.007)
Silent		-0.025***		-0.043***		-0.041***
		(0.008)		(0.008)		(0.010)
Job loss		-0.005		-0.003		0.016
		(0.015)		(0.013)		(0.013)
P. conform		0.030**		0.015		0.009
		(0.012)		(0.012)		(0.014)
P. silence		-0.020		0.004		-0.007
		(0.013)		(0.012)		(0.015)
N	1802	1802	2233	2233	1519	1519
R-sq	0.077	0.109	0.072	0.121	0.081	0.129
Controls	X	X	X	X	X	X

Notes: OLS regressions of willingness to publish (0/1). *Misinterpret:* "How often do you worry that things you post on social media can be misinterpreted?" *Silent:* "The political climate these days prevents me from saying things I believe because others might find them offensive." *Job loss:* "Are you worried about losing your job or missing out on job opportunities if your political opinions become known?" *P. conform:* "How often do you think social pressure causes people to misrepresent or lie about their political opinions on social media?" *P. silence:* "How often do you think social pressure causes people to refrain or abstain from expressing political opinions on social media?" Controls for active social media use, demographic characteristics, wave x topic FE are included in the specifications (See Table 1). Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Correlates of willingness to pay for online publication

	(1) All	(2) Dem	(3) Ind	(4) Rep
Left-Right opinion scale	-21.718***	-23.566***	-22.593***	-13.120**
	(2.684)	(4.591)	(3.993)	(6.250)
Age	-0.049	-0.210	-0.035	0.091
	(0.070)	(0.130)	(0.109)	(0.127)
Female	-6.093***	-8.478***	-2.696	-7.432**
	(1.769)	(3.017)	(2.783)	(3.601)
Asian or Pacific Islander	-6.352*	-5.042	-7.620	-13.210*
	(3.303)	(5.235)	(5.076)	(7.321)
Black or African American	4.387	3.077	-1.288	21.534***
	(2.988)	(4.489)	(4.581)	(8.260)
Hispanic or Latino	-0.922	-1.520	-1.673	-1.589
	(3.513)	(5.950)	(5.467)	(7.816)
Other race	8.122	-8.234	10.172	23.051
	(5.912)	(11.519)	(7.425)	(15.337)
College degree	-7.659***	-7.836**	-10.069***	-2.695
	(1.797)	(3.382)	(2.757)	(3.467)
Employed	5.356***	3.679	3.286	10.255***
	(1.911)	(3.502)	(2.958)	(3.675)
Risk attitude	8.842***	8.774***	10.432***	6.828***
	(0.871)	(1.537)	(1.410)	(1.622)
Active SM users	7.587***	6.942**	7.001**	8.801**
	(1.832)	(3.237)	(2.870)	(3.572)
Democrat	5.158**			
	(2.016)			
Republican	3.844*			
	(2.171)			
Constant	-55.942***	-39.931***	-58.037***	-68.461***
	(4.262)	(7.449)	(6.380)	(9.335)
N	3152	1085	1237	830
R-sq	0.092	0.094	0.105	0.091

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. 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 A7: Conformity in stated views (Experiment 1)

	Unaware (T1-3)		All			
	(1)	(2)	(3)	(4)	(5)	(6)
Prime	-0.025	-0.027	0.000	-0.003	0.000	-0.002
	(0.025)	(0.018)	(0.020)	(0.015)	(0.020)	(0.015)
Awareness					-0.001	-0.010
					(0.022)	(0.016)
N	1250	1250	1800	1800	1800	1800
R-sq	0.002	0.379	0.001	0.356	0.001	0.357
Topic FE	X	X	X	X	X	X
Controls		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 leaning and its square, 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 A8: Willingness to publish and experimental interventions (Experiment 1)

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	0.022	0.027	-0.010	-0.009	0.069*	0.085**	0.014	0.029
	(0.022)	(0.021)	(0.036)	(0.036)	(0.037)	(0.034)	(0.042)	(0.039)
Awareness	0.014	0.019	0.011	-0.003	0.069	0.073	-0.020	0.013
	(0.030)	(0.029)	(0.054)	(0.055)	(0.051)	(0.045)	(0.049)	(0.045)
N	2550	2550	1002	1002	831	831	717	717
R-sq	0.001	0.078	0.006	0.077	0.009	0.146	0.001	0.124
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 leaning and its square, 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 (Experiment 2)

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	-0.058***	-0.048**	-0.059	-0.057	-0.059*	-0.057*	-0.050	-0.034
	(0.022)	(0.022)	(0.042)	(0.041)	(0.033)	(0.033)	(0.041)	(0.040)
HiPeer	0.054***	0.044**	0.057	0.069*	0.040	0.036	0.075*	0.063*
	(0.020)	(0.020)	(0.039)	(0.037)	(0.030)	(0.030)	(0.038)	(0.038)
N	3004	3004	800	800	1402	1402	802	802
R-sq	0.009	0.062	0.009	0.105	0.006	0.081	0.015	0.071
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 leaning and its square, 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: Willingness to pay and experimental interventions

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	-2.963	-2.037	-1.371	-1.148	-4.128	-3.196	-2.670	-1.738
	(1.828)	(1.773)	(3.124)	(3.031)	(2.909)	(2.817)	(3.545)	(3.444)
HiPeer	4.093*	2.685	3.466	3.184	2.380	1.495	7.848*	5.066
	(2.436)	(2.362)	(4.789)	(4.647)	(3.533)	(3.403)	(4.762)	(4.695)
Awareness	-1.161	-0.710	-0.951	-3.211	7.276	8.064	-8.852	-5.527
	(3.574)	(3.446)	(6.228)	(6.175)	(6.134)	(5.819)	(5.961)	(5.623)
N	3152	3152	1085	1085	1237	1237	830	830
R-sq	0.007	0.084	0.017	0.101	0.008	0.120	0.009	0.090
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 leaning and its square, 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 A11: Willingness to publish and experimental interventions

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	-0.024	-0.015	-0.036	-0.033	-0.018	-0.009	-0.016	-0.006
	(0.017)	(0.016)	(0.029)	(0.028)	(0.027)	(0.026)	(0.031)	(0.030)
HiPeer	0.060***	0.048**	0.053	0.051	0.055*	0.044	0.076*	0.055
	(0.022)	(0.021)	(0.041)	(0.040)	(0.033)	(0.032)	(0.041)	(0.040)
Awareness	0.001	0.007	0.004	-0.005	0.038	0.045	-0.028	-0.001
	(0.030)	(0.028)	(0.054)	(0.054)	(0.050)	(0.047)	(0.048)	(0.045)
N	5086	5086	1673	1673	2006	2006	1407	1407
R-sq	0.004	0.062	0.011	0.079	0.004	0.091	0.006	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), *excluding those who believed it was extremely unlikely that posts would be published.* Controls include age, gender, race, education, employment, risk attitude, political leaning and its square, 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: Willingness to publish and experimental interventions (Experiment 1)

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	0.027	0.025	-0.052	-0.034	0.174**	0.170***	0.025	0.013
	(0.041)	(0.041)	(0.058)	(0.058)	(0.071)	(0.064)	(0.127)	(0.115)
Awareness	0.000	0.013	-0.005	-0.009	0.077	0.094	-0.037	0.008
	(0.041)	(0.040)	(0.078)	(0.077)	(0.066)	(0.061)	(0.073)	(0.067)
Prime x Awareness	0.023	0.008	0.022	0.010	-0.003	-0.034	0.023	-0.000
	(0.054)	(0.052)	(0.099)	(0.100)	(0.092)	(0.085)	(0.091)	(0.084)
Prime x LR scale	-0.023	-0.000	0.100	0.065	-0.191*	-0.143	-0.029	0.017
	(0.057)	(0.056)	(0.104)	(0.102)	(0.104)	(0.096)	(0.138)	(0.130)
N	2550	2550	1002	1002	831	831	717	717
R-sq	0.028	0.103	0.017	0.097	0.048	0.159	0.019	0.138
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 leaning and its square, 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 A13: Willingness to publish and experimental interventions (Experiment 2)

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	0.003	0.011	-0.099	-0.104	0.022	0.043	0.237**	0.164
	(0.049)	(0.048)	(0.080)	(0.079)	(0.072)	(0.070)	(0.119)	(0.124)
HiPeer	0.072*	0.074**	0.013	0.030	0.067	0.078	0.151**	0.139**
	(0.037)	(0.036)	(0.069)	(0.067)	(0.056)	(0.056)	(0.069)	(0.067)
Prime x HiPeer	-0.027	-0.044	0.063	0.054	-0.033	-0.060	-0.105	-0.106
	(0.044)	(0.043)	(0.083)	(0.082)	(0.066)	(0.065)	(0.083)	(0.082)
Prime x LR scale	-0.077	-0.062	0.015	0.050	-0.114	-0.127	-0.282**	-0.174
	(0.061)	(0.060)	(0.110)	(0.106)	(0.092)	(0.092)	(0.131)	(0.137)
N	3004	3004	800	800	1402	1402	802	802
R-sq	0.018	0.068	0.034	0.122	0.017	0.088	0.023	0.076
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 leaning and its square, 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 A14: Willingness to pay and experimental interventions

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	-1.487	-1.738	-7.646	-7.996	2.930	4.914	12.337	3.890
	(3.713)	(3.597)	(5.326)	(5.202)	(6.165)	(5.881)	(12.216)	(11.981)
HiPeer	7.346*	6.764*	2.282	2.429	13.070**	14.012**	5.790	2.595
	(3.897)	(3.821)	(6.863)	(6.516)	(5.959)	(6.001)	(7.692)	(7.486)
Awareness	0.676	1.614	0.322	-1.952	5.672	9.359	-1.055	1.151
	(4.972)	(4.878)	(8.911)	(9.117)	(8.021)	(7.585)	(9.063)	(8.712)
Prime x HiPeer	-5.117	-6.164	1.440	1.159	-14.413**	-17.904***	2.562	3.467
	(4.410)	(4.308)	(7.815)	(7.468)	(6.702)	(6.658)	(8.803)	(8.586)
Prime x Awareness	-3.850	-5.113	-4.667	-4.723	6.775	1.095	-14.949	-13.688
	(6.229)	(6.074)	(11.124)	(11.248)	(10.781)	(10.378)	(10.559)	(10.058)
Prime x LR scale	-0.065	2.554	15.570*	17.945**	-6.625	-6.384	-17.690	-6.241
	(5.120)	(5.003)	(9.337)	(9.026)	(8.582)	(8.307)	(13.734)	(13.582)
N	3152	3152	1085	1085	1237	1237	830	830
R-sq	0.031	0.101	0.038	0.120	0.039	0.134	0.024	0.098
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 leaning and its square, 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 A15: Willingness to publish and experimental interventions

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	0.015	0.014	-0.078	-0.066	0.090*	0.100**	0.134	0.070
	(0.033)	(0.032)	(0.048)	(0.048)	(0.054)	(0.051)	(0.094)	(0.089)
HiPeer	0.085**	0.078**	0.020	0.019	0.109**	0.110**	0.121*	0.106*
	(0.035)	(0.034)	(0.063)	(0.060)	(0.055)	(0.055)	(0.065)	(0.061)
Awareness	-0.023	-0.008	-0.034	-0.037	0.023	0.049	-0.032	0.001
	(0.040)	(0.039)	(0.077)	(0.076)	(0.064)	(0.062)	(0.068)	(0.064)
Prime x HiPeer	-0.037	-0.048	0.036	0.034	-0.069	-0.095	-0.069	-0.079
	(0.040)	(0.039)	(0.072)	(0.069)	(0.062)	(0.061)	(0.076)	(0.073)
Prime x Awareness	0.053	0.035	0.060	0.053	0.061	0.025	0.019	0.004
	(0.052)	(0.051)	(0.097)	(0.098)	(0.089)	(0.085)	(0.086)	(0.080)
Prime x LR scale	-0.064	-0.038	0.066	0.059	-0.167**	-0.150**	-0.170	-0.070
	(0.044)	(0.044)	(0.079)	(0.078)	(0.074)	(0.072)	(0.103)	(0.099)
N	5086	5086	1673	1673	2006	2006	1407	1407
R-sq	0.022	0.076	0.028	0.099	0.027	0.101	0.014	0.087
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), *excluding those who believed it was extremely unlikely that posts would be published.* Controls include age, gender, race, education, employment, risk attitude, political leaning and its square, 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 A16: 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	0.015	-0.031	-0.032	0.062	0.070	0.077	-0.001
	(0.038)	(0.037)	(0.060)	(0.058)	(0.058)	(0.056)	(0.110)	(0.106)
HiPeer	0.090**	0.087**	0.059	0.064	0.087	0.096*	0.128*	0.111*
	(0.035)	(0.034)	(0.065)	(0.062)	(0.053)	(0.053)	(0.066)	(0.062)
Prime x HiPeer	-0.054	-0.064	-0.007	-0.004	-0.064	-0.088	-0.081	-0.076
	(0.041)	(0.040)	(0.076)	(0.073)	(0.062)	(0.061)	(0.077)	(0.074)
Prime x LR scale	-0.055	-0.032	0.019	0.019	-0.129*	-0.126*	-0.111	-0.007
	(0.047)	(0.047)	(0.083)	(0.082)	(0.077)	(0.075)	(0.120)	(0.117)
N	4098	4098	1234	1234	1749	1749	1115	1115
R-sq	0.020	0.076	0.030	0.118	0.018	0.093	0.016	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). The sample excludes treatment branches in which LR-scale (s) is endogenous, namely: Treatments 1, 4 and 5 in Experiment 1 (see Figure 2). Controls include age, gender, race, education, employment, risk attitude, political leaning and its square, 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 A17: Correlates of concern about political climate and freedom of speech online

	(1) All	(2) Dem	(3) Ind	(4) Rep
Left-Right opinion scale	0.780***	0.909***	0.897***	0.160
	(0.079)	(0.124)	(0.131)	(0.152)
Willing to publish with name	-0.517***	-0.418***	-0.529***	-0.592***
	(0.057)	(0.091)	(0.098)	(0.104)
Age	-0.007***	0.003	-0.008**	-0.014***
	(0.002)	(0.004)	(0.004)	(0.004)
Non-binary	0.207	-0.526	0.562**	-0.687***
	(0.229)	(0.498)	(0.252)	(0.181)
Female	0.078	-0.026	0.152	0.104
	(0.057)	(0.102)	(0.095)	(0.109)
Asian or Pacific Islander	0.238**	0.359**	0.370**	-0.353
	(0.111)	(0.167)	(0.163)	(0.347)
Black or African American	0.093	0.310**	-0.021	-0.202
	(0.100)	(0.156)	(0.168)	(0.197)
Hispanic or Latino	0.045	-0.140	0.224	0.046
	(0.120)	(0.207)	(0.186)	(0.222)
Other race	0.152	0.306	0.097	0.581
	(0.179)	(0.409)	(0.230)	(0.365)
College degree	0.221***	-0.042	0.222**	0.423***
	(0.057)	(0.102)	(0.090)	(0.106)
Employed	0.037	-0.043	0.085	0.062
	(0.063)	(0.107)	(0.102)	(0.124)
Risk attitude	0.107***	0.149***	0.120**	0.036
	(0.030)	(0.054)	(0.050)	(0.050)
Pol. leaning	0.262***	0.186**	0.181**	0.482***
	(0.032)	(0.092)	(0.074)	(0.156)
Pol. leaning sq	-0.079***	-0.129*	-0.121**	-0.143*
	(0.026)	(0.069)	(0.060)	(0.079)
Active SM users	0.333***	0.394***	0.233**	0.417***
	(0.059)	(0.102)	(0.093)	(0.111)
Constant	-0.204	-0.327	-0.267	0.242
	(0.129)	(0.233)	(0.204)	(0.251)
N	5554	1802	2233	1519
R-sq	0.165	0.150	0.153	0.143

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. Fixed effects for survey waves $\times$ 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 A18: Concerns about freedom of speech and experimental interventions

	All		Dem		Ind		Rep	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	-0.105	-0.098	-0.338**	-0.269*	0.027	0.047	0.633**	0.446
	(0.103)	(0.100)	(0.139)	(0.137)	(0.170)	(0.168)	(0.308)	(0.300)
HiPeer	-0.102	-0.096	-0.115	-0.111	-0.140	-0.092	-0.038	-0.078
	(0.121)	(0.119)	(0.218)	(0.214)	(0.185)	(0.184)	(0.234)	(0.228)
Awareness	-0.112	-0.175	-0.418	-0.459*	0.126	0.068	-0.079	-0.109
	(0.137)	(0.136)	(0.261)	(0.258)	(0.217)	(0.225)	(0.238)	(0.223)
Prime x HiPeer	0.119	0.115	0.112	0.065	0.194	0.171	0.049	0.093
	(0.138)	(0.136)	(0.246)	(0.242)	(0.215)	(0.213)	(0.264)	(0.257)
Prime x Awareness	-0.016	0.005	0.539*	0.561*	-0.368	-0.310	-0.244	-0.245
	(0.176)	(0.176)	(0.326)	(0.315)	(0.288)	(0.301)	(0.308)	(0.301)
Prime x LR scale	0.046	0.049	0.199	0.178	-0.249	-0.277	-0.575*	-0.370
	(0.141)	(0.138)	(0.243)	(0.235)	(0.235)	(0.232)	(0.346)	(0.338)
N	5554	5554	1802	1802	2233	2233	1519	1519
R-sq	0.099	0.147	0.092	0.141	0.090	0.137	0.050	0.125
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 leaning and its square, 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 A19: 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
True value	4.396	3.915	0.481

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. The gender question is reverse-coded: 1 = progressive, 7 = conservative.

Table A20: 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. The gender question is reverse-coded: 1 = progressive, 7 = conservative.

Table A21: Correlates of perceived censorship

	(1) All	(2) Dem	(3) Ind	(4) Rep
Left-Right opinion scale	0.452***	0.319	0.538***	0.680***
	(0.123)	(0.217)	(0.194)	(0.248)
Age	-0.002	-0.013**	0.003	0.003
	(0.003)	(0.005)	(0.005)	(0.005)
Non-binary	0.179	0.187	0.110	-0.596**
	(0.304)	(0.497)	(0.371)	(0.250)
Female	0.124*	0.165	0.172	-0.073
	(0.074)	(0.124)	(0.122)	(0.146)
Asian or Pacific Islander	-0.058	0.092	-0.445**	0.263
	(0.143)	(0.216)	(0.222)	(0.351)
Black or African American	-0.183	-0.405**	0.165	-0.394
	(0.120)	(0.185)	(0.190)	(0.287)
Hispanic or Latino	-0.105	-0.083	-0.278	0.256
	(0.143)	(0.211)	(0.232)	(0.359)
Other race	0.144	0.916	-0.048	-0.465
	(0.253)	(0.581)	(0.322)	(0.489)
College degree	0.004	0.023	0.070	-0.027
	(0.076)	(0.134)	(0.119)	(0.150)
Employed	-0.028	-0.016	-0.122	0.105
	(0.082)	(0.142)	(0.127)	(0.165)
Risk attitude	-0.063*	-0.052	-0.068	-0.059
	(0.037)	(0.063)	(0.062)	(0.073)
Pol. leaning	-0.089**	-0.153	-0.085	0.009
	(0.042)	(0.099)	(0.085)	(0.199)
Pol. leaning sq	0.043	-0.084	0.166**	0.025
	(0.033)	(0.086)	(0.066)	(0.096)
Active SM users	-0.160**	-0.115	-0.190	-0.170
	(0.077)	(0.133)	(0.124)	(0.151)
Constant	0.039	0.432	-0.067	-0.513
	(0.184)	(0.323)	(0.296)	(0.354)
N	3152	1085	1237	830
R-sq	0.045	0.053	0.049	0.072

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. 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 A22: 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***	-0.180***	-0.165***	-0.194***	-0.142**	-0.141**
	(0.027)	(0.029)	(0.046)	(0.047)	(0.038)	(0.043)	(0.063)	(0.069)
HiPeer	0.021	0.001	-0.034	-0.066	-0.036	-0.038	-0.098	-0.118
	(0.043)	(0.043)	(0.067)	(0.062)	(0.062)	(0.062)	(0.093)	(0.091)
HiPeer x Left-Right scale	0.040	0.045	0.118	0.132	0.178**	0.184**	0.286**	0.332***
	(0.057)	(0.056)	(0.091)	(0.085)	(0.085)	(0.084)	(0.135)	(0.127)
N	3654	3654	1478	1478	1900	1900	755	755
R-sq	0.026	0.089	0.035	0.128	0.019	0.074	0.024	0.101
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 leaning and its square, 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 A23: Willingness to publish and experimental interventions, by states

	Dem states				Rep states			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prime	-0.032	-0.025	0.030	0.031	-0.015	-0.008	0.007	0.002
	(0.026)	(0.025)	(0.049)	(0.047)	(0.020)	(0.019)	(0.041)	(0.040)
HiPeer	0.051	0.041	0.137**	0.116**	0.056**	0.044*	0.061	0.060
	(0.033)	(0.033)	(0.057)	(0.054)	(0.025)	(0.025)	(0.041)	(0.041)
Awareness	-0.058	-0.035	-0.069	-0.045	0.037	0.042	0.018	0.025
	(0.047)	(0.045)	(0.064)	(0.062)	(0.038)	(0.036)	(0.051)	(0.049)
Prime x HiPeer			-0.123*	-0.108*			-0.012	-0.026
			(0.063)	(0.061)			(0.047)	(0.047)
Prime x Awareness			0.029	0.029			0.053	0.036
			(0.082)	(0.078)			(0.067)	(0.065)
Prime x LR scale			-0.060	-0.054			-0.043	-0.014
			(0.066)	(0.066)			(0.054)	(0.053)
N	2084	2084	2084	2084	3470	3470	3470	3470
R-sq	0.007	0.081	0.023	0.101	0.004	0.064	0.023	0.074
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 leaning and its square, social media use, and wave FE $\times$ topic FE. Sample is split by states based on 2016 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 A24: Perceived norms and HiPeer treatment

	Perceived norm of all participants				Perceived norm of those who publish				Perceived gap, $n_{all} - n_{pub}$			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
HiPeer	-0.031	-0.035	-0.007	-0.058	0.060**	0.105**	0.054	0.015	-0.091***	-0.140***	-0.061	-0.073
	(0.020)	(0.038)	(0.030)	(0.037)	(0.024)	(0.043)	(0.037)	(0.047)	(0.025)	(0.044)	(0.041)	(0.048)
Prime	-0.006	-0.032**	0.022	-0.016	0.011	0.006	0.025	-0.001	-0.018	-0.039*	-0.004	-0.015
	(0.010)	(0.015)	(0.015)	(0.020)	(0.013)	(0.021)	(0.021)	(0.025)	(0.014)	(0.023)	(0.023)	(0.029)
HiPeer x Prime	0.022	0.054	-0.014	0.039	-0.008	-0.045	-0.003	0.024	0.029	0.099**	-0.011	0.015
	(0.022)	(0.042)	(0.034)	(0.041)	(0.027)	(0.049)	(0.042)	(0.052)	(0.028)	(0.047)	(0.045)	(0.054)
N	3152	1085	1237	830	3152	1085	1237	830	3152	1085	1237	830
R-sq	0.083	0.086	0.064	0.126	0.033	0.037	0.040	0.037	0.047	0.060	0.050	0.069
Topic + Wave FE	X	X	X	X	X	X	X	X	X	X	X	X
Controls	X	X	X	X	X	X	X	X	X	X	X	X
Sample	All	Dem	Ind	Rep	All	Dem	Ind	Rep	All	Dem	Ind	Rep

Notes: OLS regressions of n_{all}, n_{pub}, and $n_{all} - n_{pub}$ on a Left-Right scale (reverse-coded for Gender statement) and scaled to 0-1. Controls include age, gender, race, education, employment, risk attitude, political leaning and its square, 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 A25: Perceived political climate and HiPeer treatment

	Worries about being misinterpreted				Silenced by political climate				Worries about job loss			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
HiPeer	0.036	-0.074	0.122	-0.010	-0.145*	-0.297**	-0.017	-0.194	-0.027	0.085	-0.132	0.006
	(0.078)	(0.139)	(0.123)	(0.152)	(0.078)	(0.147)	(0.117)	(0.151)	(0.079)	(0.141)	(0.117)	(0.154)
Prime	-0.083**	-0.079	-0.168**	0.027	-0.095**	-0.172***	-0.081	0.006	-0.007	0.062	-0.101	0.048
	(0.040)	(0.063)	(0.068)	(0.081)	(0.040)	(0.066)	(0.065)	(0.077)	(0.041)	(0.064)	(0.066)	(0.087)
HiPeer x Prime	0.038	0.011	0.078	0.026	0.168*	0.263	0.048	0.260	0.041	-0.174	0.218	0.024
	(0.088)	(0.153)	(0.139)	(0.172)	(0.087)	(0.163)	(0.133)	(0.166)	(0.089)	(0.153)	(0.134)	(0.177)
N	6304	2170	2474	1660	6304	2170	2474	1660	6304	2170	2474	1660
R-sq	0.066	0.061	0.086	0.083	0.131	0.117	0.119	0.068	0.092	0.102	0.093	0.095
Topic + Wave FE	X	X	X	X	X	X	X	X	X	X	X	X
Controls	X	X	X	X	X	X	X	X	X	X	X	X
Sample	All	Dem	Ind	Rep	All	Dem	Ind	Rep	All	Dem	Ind	Rep

Notes: OLS regressions of standardized responses 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." Controls include age, gender, race, education, employment, risk attitude, political leaning and its square, 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 A26: Perceived political climate and HiPeer treatment

	Thinks others conform				Thinks others stay silent			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
HiPeer	-0.044	0.025	-0.066	-0.080	-0.036	0.038	-0.066	0.005
	(0.079)	(0.143)	(0.125)	(0.143)	(0.083)	(0.140)	(0.135)	(0.159)
Prime	0.013	-0.043	0.036	0.045	0.019	-0.037	0.019	0.102
	(0.041)	(0.069)	(0.064)	(0.080)	(0.042)	(0.068)	(0.072)	(0.078)
HiPeer x Prime	0.021	-0.023	0.069	-0.005	0.004	-0.003	-0.012	0.016
	(0.088)	(0.163)	(0.141)	(0.156)	(0.092)	(0.160)	(0.150)	(0.168)
N	6304	2170	2474	1660	6304	2170	2474	1660
R-sq	0.082	0.083	0.062	0.078	0.045	0.037	0.048	0.067
Topic + Wave FE	X	X	X	X	X	X	X	X
Controls	X	X	X	X	X	X	X	X
Sample	All	Dem	Ind	Rep	All	Dem	Ind	Rep

Notes: OLS regressions of standardized responses to "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 leaning and its square, 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 A27: Examples of participants' reasons for posting/not posting

Reasons for posting	**Reasons for not posting**
Indifference to social pressure	Fear of backlash
• I'm not too concerned with what people think about me or my opinions • I don't see it as a big deal to post my opinion. • I don't care who knows my opinions	• it has the potential to cause a lot of trouble for me • It would have severe implications for my job should my superiors saw it and might result in my termination. • I would never post controversial opinions on social media because it could cause backlash that I don't want to deal with
Importance of issues	Disinterest in issues
• I feel the issue to me is important to comment on. • This is an opinion I feel strongly about, and I have posted similar statements on social media before. • Because it is an important issue that needs to be addressed and not undermined.	• The matter is not important to me. • This is not an issue that I'm passionate about. • i do not want to have this conversation with anyone because i do not have a strong opinion either way
Freedom of speech	Privacy concerns
• Because sometimes it just feels good to let your voice be heard. • i like to be heard • I am free to express my opinion and i think it will go a long way to enlighten the youths	• These are my personal opinions and mine only, I do not want to share it even if my personal information is not involved. • I did not want to post my name or any info on Twitter. • I do not approve of the dissemination of my personal information on Twitter or any other social media.
Financial incentives	Social media disengagement
• I need money and I don't care what strangers on the internet think about my opinions. • To try to win more lottery tickets to win the money- • because I thought i might be more likely to win the lottery	• I don't post things on Twitter. I wasn't incentivized enough either. • I dont use social media and prefer to keep it that way. • I never post on social media, and don't want to start now.
Other	Other
• I wanted to help with this study as much as possible. • I was just curious • It sounded interesting to do.	• I am sure others will post. • I can post my own opinion on my own if I want to • Lottery tickets are not worth the hassle, a different story is 100% sure cash or gift card

Notes: Examples of responses to "Why did you decide to let us post one or more of your opinions on social media?" (left) and "Why did you decide not to let us post any of your opinions on social media?" (right), as classified by GPT 4.0. These questions are only asked in Experiment 2 to the relevant group (respectively, those who chose to publish for at least one of the two issues, $n = 407$, or those who chose not to publish for both issues, $n = 1095$).

Table A28: Reasons for posting and not posting across left-right scale

	No social pressure		Fear of backlash		Importance of issues		Disinterest in issues	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LR-scale	-0.105*	-0.039	0.280***	0.244***	0.022	-0.063	-0.003	0.003
	(0.061)	(0.088)	(0.040)	(0.060)	(0.062)	(0.087)	(0.015)	(0.025)
LR-scale (sq. dev)	-0.069	-0.070	-0.031	-0.010	0.241***	0.243***	-0.038***	-0.035***
	(0.056)	(0.061)	(0.036)	(0.038)	(0.055)	(0.057)	(0.012)	(0.013)
Prime	-0.025	-0.024	0.011	0.022	0.056	0.049	0.006	0.005
	(0.044)	(0.044)	(0.027)	(0.027)	(0.043)	(0.042)	(0.011)	(0.011)
HiPeer	0.071*	0.073*	0.025	0.025	-0.076*	-0.082**	-0.005	-0.006
	(0.043)	(0.042)	(0.025)	(0.024)	(0.042)	(0.041)	(0.010)	(0.010)
N	407	407	1095	1095	407	407	1095	1095
R-sq	0.019	0.099	0.036	0.068	0.057	0.171	0.007	0.018
Controls		X		X		X		X
Sample	Post	Post	NoPost	NoPost	Post	Post	NoPost	NoPost

Notes: OLS regressions of stating a particular reason for posting/not posting, average of DV as categorized by GPT 4.0 and GPT 3.5-turbo using responses to "Why did you decide to let us post one or more of your opinions on social media?" or "Why did you decide not to let us post any of your opinions on social media?". These questions are only asked in Experiment 2 to the relevant group (respectively, those who chose to publish for at least one of the two issues, $n = 407$, or those who chose not to publish for both issues, $n = 1095$). LR scale is derived from agreement to each of the topic statements, reverse-coded for the Gender topic, and scaled to 0-1. For those who posted one of the two statements, we used the response to the relevant statement. For those who posted neither or both of the two statements, we used the average response of the two statements. Sq. dev is squared deviation of LR scale from the middle position of 0.5, as a measure of non-linearity. Controls include age, gender, race, education, employment, risk attitude, political leaning and its square and social media use. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A29: Reasons for posting and not posting across left-right scale

	Freedom of speech		Privacy concerns		Financial incentives		SM disengagement	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LR-scale	0.060*	0.023	-0.234***	-0.170***	0.057**	0.098***	-0.025	-0.059
	(0.032)	(0.053)	(0.049)	(0.065)	(0.023)	(0.037)	(0.037)	(0.049)
LR-scale (sq. dev)	-0.066**	-0.042	0.071*	0.051	-0.033	-0.039	-0.009	-0.006
	(0.031)	(0.031)	(0.040)	(0.043)	(0.020)	(0.026)	(0.028)	(0.030)
Prime	-0.033	-0.036	-0.033	-0.033	0.000	0.006	0.014	0.006
	(0.027)	(0.028)	(0.029)	(0.029)	(0.021)	(0.020)	(0.021)	(0.022)
HiPeer	0.015	0.016	-0.039	-0.033	-0.011	-0.007	0.008	0.005
	(0.025)	(0.026)	(0.026)	(0.026)	(0.020)	(0.020)	(0.020)	(0.020)
N	407	407	1095	1095	407	407	1095	1095
R-sq	0.021	0.036	0.024	0.037	0.015	0.062	0.001	0.030
Controls		X		X		X		X
Sample	Post	Post	NoPost	NoPost	Post	Post	NoPost	NoPost

Notes: OLS regressions of stating a particular reason for posting/not posting, average of DV as categorized by GPT 4.0 and GPT 3.5-turbo using responses to "Why did you decide to let us post one or more of your opinions on social media?" or "Why did you decide not to let us post any of your opinions on social media?". These questions are only asked in Experiment 2 to the relevant group (respectively, those who chose to publish for at least one of the two issues, $n = 407$, or those who chose not to publish for both issues, $n = 1095$). LR scale is derived from agreement to each of the topic statements, reverse-coded for the Gender topic, and scaled to 0-1. For those who posted one of the two statements, we used the response to the relevant statement. For those who posted neither or both of the two statements, we used the average response of the two statements. Sq. dev is squared deviation of LR scale from the middle position of 0.5, as a measure of non-linearity. Controls include age, gender, race, education, employment, risk attitude, political leaning and its square and social media use. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A30: Number of statements published and reason given

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
HiPeer	0.088**	0.027	0.094***	0.072*	0.090**	0.091**	0.086**	0.056	0.087**
	(0.040)	(0.031)	(0.036)	(0.037)	(0.039)	(0.037)	(0.039)	(0.036)	(0.039)
No social pressure		1.705***							
		(0.051)							
Importance of issues			1.287***						
			(0.058)						
Freedom of speech				1.922***					
				(0.107)					
Financial incentives					1.109***				
					(0.135)				
Fear of backlash						-0.690***			
						(0.032)			
Disinterest in issues							-0.579***		
							(0.057)		
Privacy concerns								-0.753***	
								(0.032)	
SM disengagement									-0.608***
									(0.033)
N	1502	1502	1502	1502	1502	1502	1502	1502	1502
R-sq	0.068	0.436	0.255	0.187	0.089	0.174	0.078	0.223	0.118
Controls	X	X	X	X	X	X	X	X	X

Notes: OLS regressions of number of statements published. Each of the reasons is the average of DV as categorized by GPT 4.0 and GPT 3.5-turbo using responses to "Why did you decide to let us post one or more of your opinions on social media?" or "Why did you decide not to let us post any of your opinions on social media?". These questions are only asked in Experiment 2 to the relevant group (respectively, those who chose to publish for at least one of the two issues, $n = 407$, or those who chose not to publish for both issues, $n = 1095$). Thus, the DV equals 0 for the group not asked. Controls include age, gender, race, education, employment, risk attitude, political leaning and its square and social media use. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

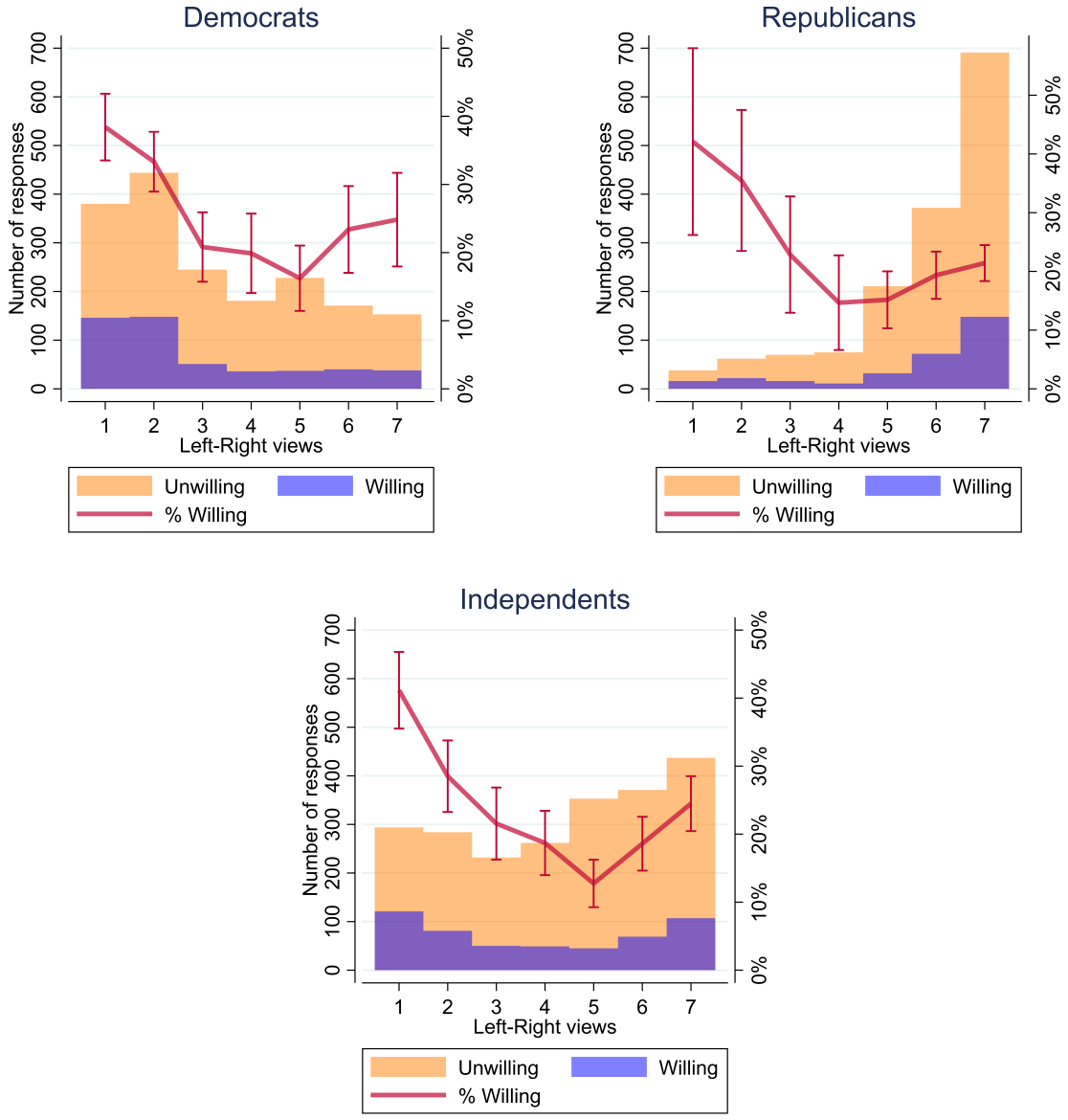
Table A31: Number of statements published and reason given

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Prime	-0.095**	-0.042	-0.087**	-0.062	-0.091**	-0.073*	-0.092**	-0.092**	-0.084**
	(0.043)	(0.034)	(0.039)	(0.040)	(0.043)	(0.041)	(0.043)	(0.039)	(0.042)
No social pressure		1.704***							
		(0.051)							
Importance of issues			1.283***						
			(0.059)						
Freedom of speech				1.918***					
				(0.106)					
Financial incentives					1.098***				
					(0.133)				
Fear of backlash						-0.685***			
						(0.032)			
Disinterest in issues							-0.579***		
							(0.056)		
Privacy concerns								-0.756***	
								(0.032)	
SM disengagement									-0.604***
									(0.033)
N	1502	1502	1502	1502	1502	1502	1502	1502	1502
R-sq	0.068	0.436	0.254	0.186	0.089	0.172	0.078	0.225	0.118
Controls	X	X	X	X	X	X	X	X	X

Notes: OLS regressions of number of statements published. Each of the reasons is the average of DV as categorized by GPT 4.0 and GPT 3.5-turbo using responses to "Why did you decide to let us post one or more of your opinions on social media?" or "Why did you decide not to let us post any of your opinions on social media?". These questions are only asked in Experiment 2 to the relevant group (respectively, those who chose to publish for at least one of the two issues, $n = 407$, or those who chose not to publish for both issues, $n = 1095$). Thus, the DV equals 0 for the group not asked. Controls include age, gender, race, education, employment, risk attitude, political leaning and its square and social media use. Robust standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

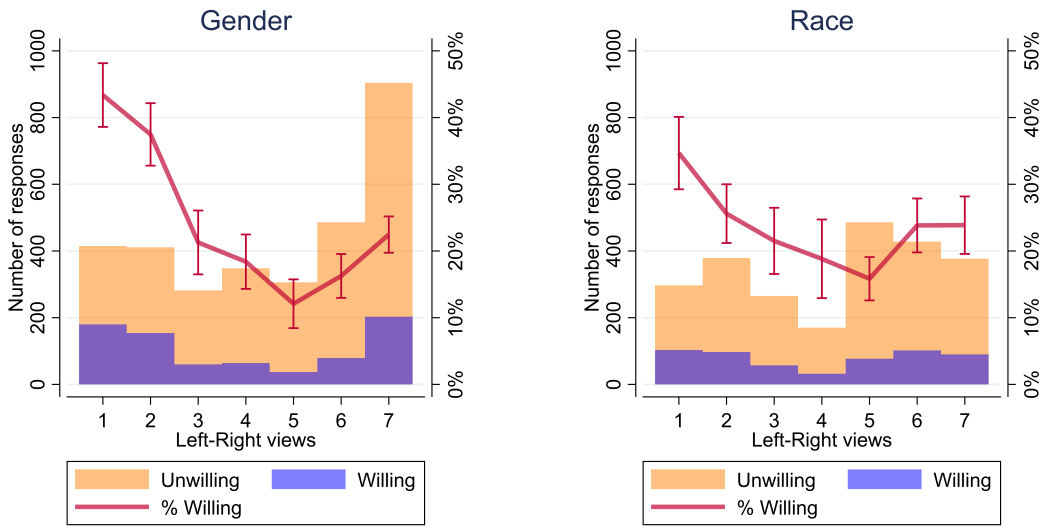
C Appendix Figures

Figure A1: Public expression and attitudes across parties



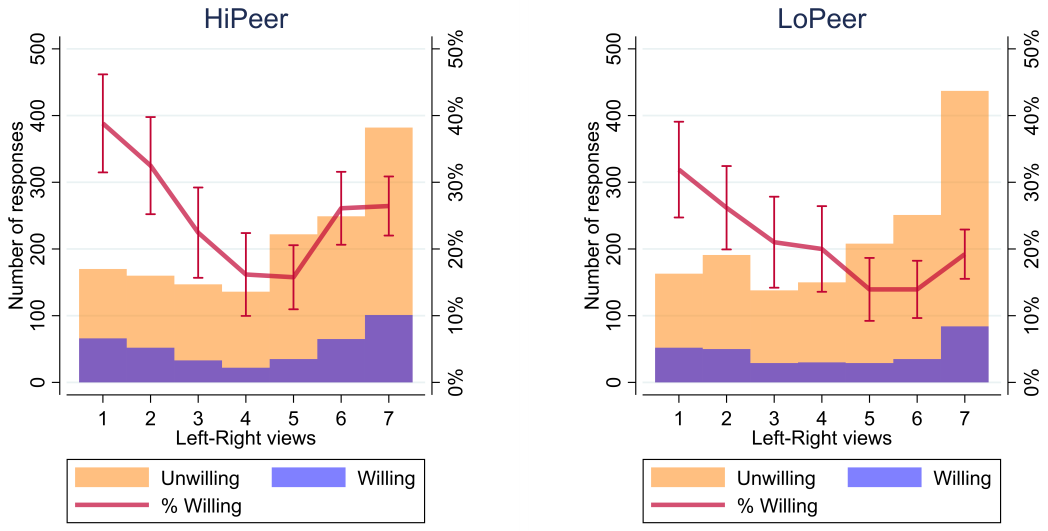
Notes: Agreement to statement in Experiments 1 and 2 (pooled) by willingness to publish and proportion willing to publish. Responses to Gender statement are reverse-coded. Figures show attitudes and public expression separately by party: Democrats (left), Independents (middle) and Republicans (right). Bars represent 95% confidence intervals.

Figure A2: Public expression and attitudes across topics



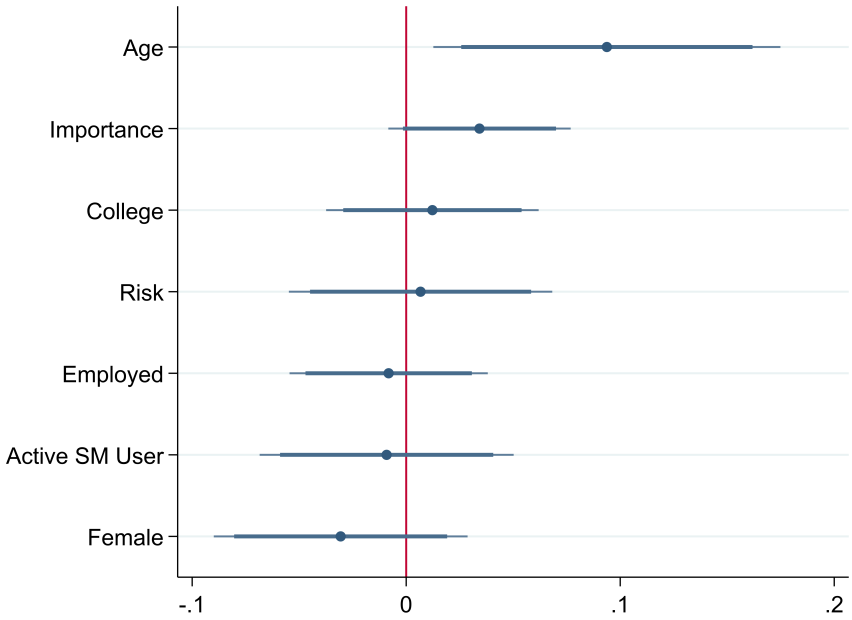
Notes: Agreement to statement in Experiments 1 and 2 (pooled) by willingness to publish and proportion willing to publish. Responses to Gender statement are reverse-coded. Figures show attitudes and public expression separately by topic: Gender (left) and Race (right). Bars represent 95% confidence intervals.

Figure A3: Public expression and attitudes across HiPeer and LoPeer treatments



Notes: Agreement to statement in Experiment 2 by willingness to publish and proportion willing to publish. Responses to Gender statement are reverse-coded. Figures show attitudes and public expression separately by peer treatment: HiPeer (left) and LoPeer (right). Bars represent 95% confidence intervals.

Figure A4: Heterogeneity of Publication Awareness treatment

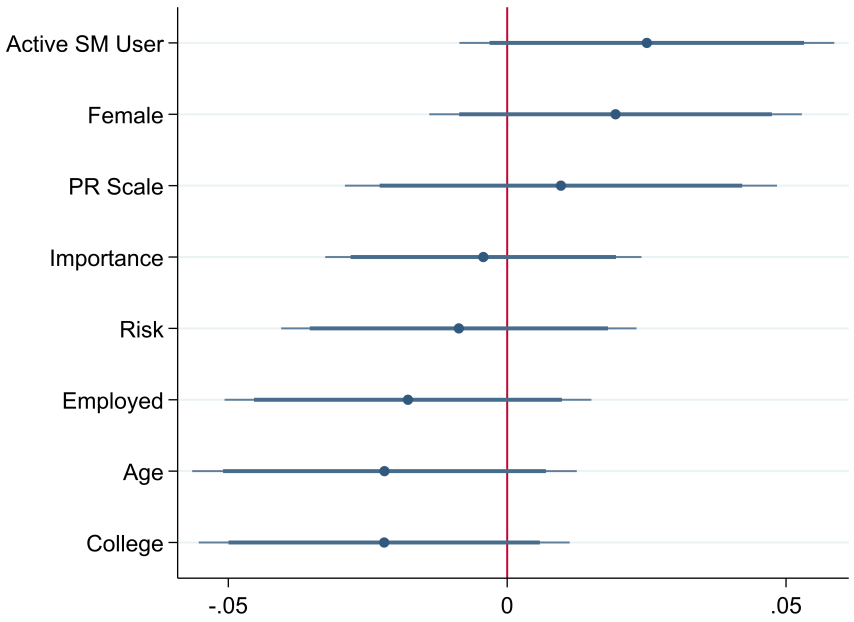


Notes: The figure shows the coefficient θ_5 from the following model:

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

where Var_i indicates the dimension of interest in exploring heterogeneous effects. *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:* 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 leaning and its square, social media use, and wave FE $\times$ topic FE. Robust standard errors are clustered at the individual level.

Figure A5: Heterogeneity of HiPeer treatment

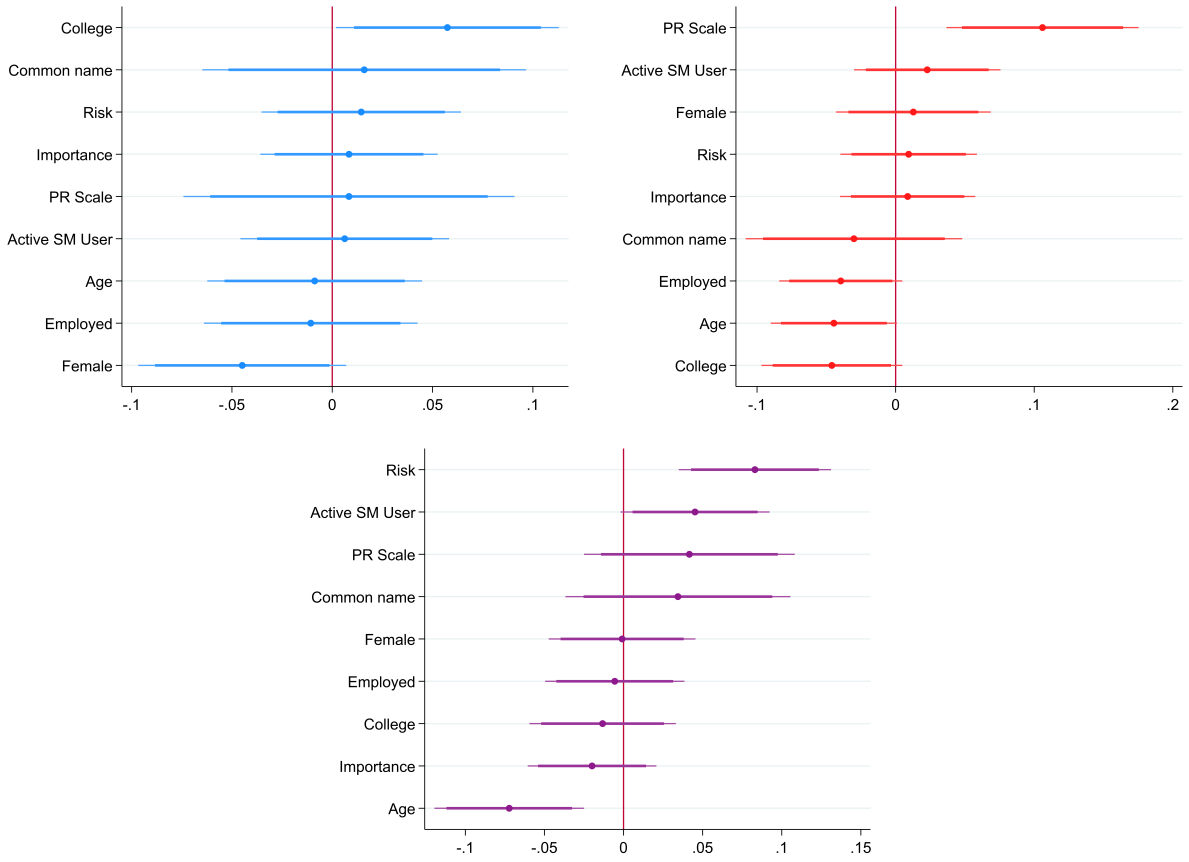


Notes: The figure shows the coefficient θ_5 from the following model:

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

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:* 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 leaning and its square, social media use, and wave FE $\times$ topic FE. Robust standard errors are clustered at the individual level.

Figure A6: Heterogeneity of Prime treatment by party

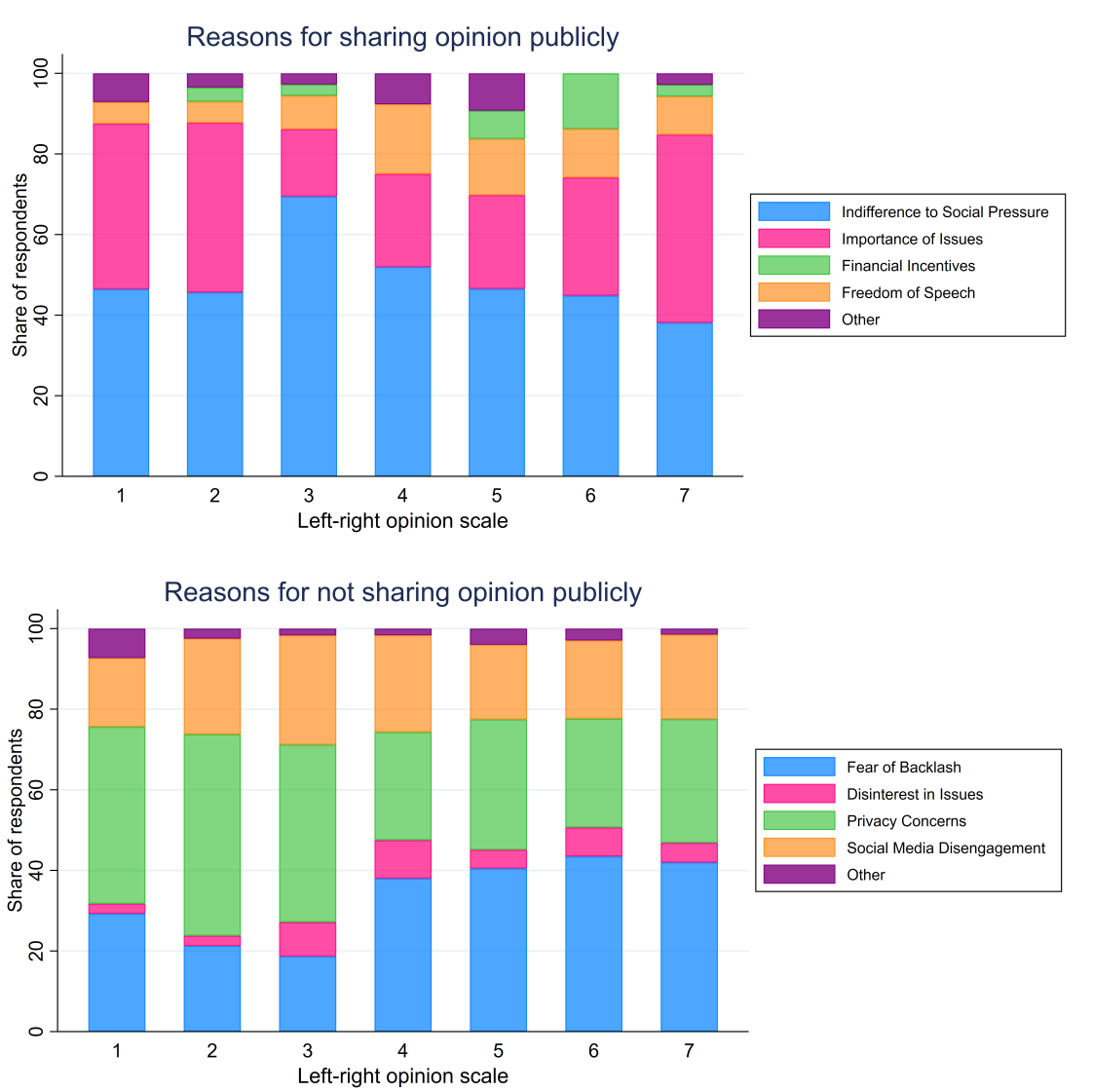


Notes: The figure shows the coefficient θ_5 from the following model:

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

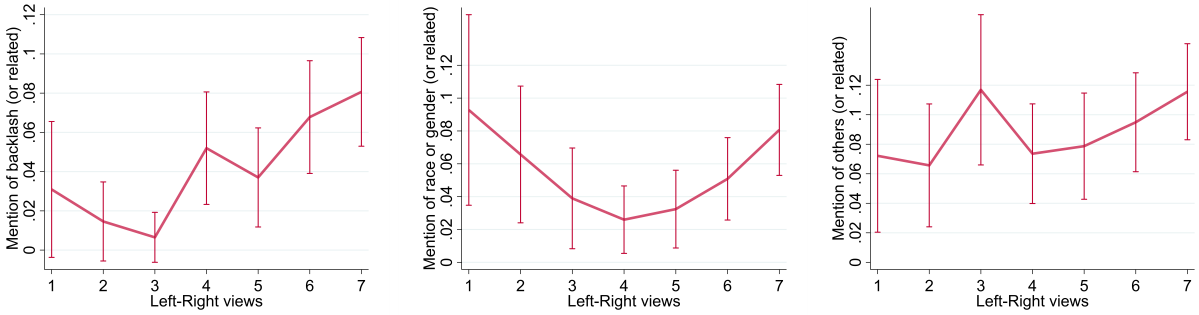
where Var_i indicates the dimension of interest in exploring heterogeneous effects, estimated separately for each political party identification. *PR Scale:* responses to the Hong psychological reactance scale (Hong and Faedda, 1996) (Experiment 2 only). *Risk:* response to "Please tell us, in general, 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:* 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 leaning and its square, social media use, and wave FE $\times$ topic FE. Robust standard errors are clustered at the individual level.

Figure A7: Text analysis of participants' open text-box responses



Notes: GPT 4.0 categorization of responses to "Why did you decide to let us post one or more of your opinions on social media?" (top) and "Why did you decide not to let us post any of your opinions on social media?" (bottom). These questions are only asked in Experiment 2 to the relevant group (respectively, those who chose to publish for at least one of the two issues, $n = 407$, or those who chose not to publish for both issues, $n = 1095$). Left-Right views are derived from agreement to each of the topic statements, reverse-coded for the Gender topic. For those who posted one of the two statements, we used the response to the relevant statement. For those who posted neither or both of the two statements, we used the average response of the two statements.

Figure A8: Keyword mentions across left-right scale



Notes: These figures show the proportion of respondents mentioning certain keywords in their responses to "Why did you decide to let us post one or more of your opinions on social media?" or "Why did you decide not to let us post any of your opinions on social media?". These questions are only asked in Experiment 2 to the relevant group (respectively, those who chose to publish for at least one of the two issues, $n = 407$, or those who chose not to publish for both issues, $n = 1095$). The left panel shows the proportion of respondents mentioning "pressure", "cancel", "online mob" or "backlash". The middle panel shows the proportion of respondents mentioning "trans", "race", "gender", "sport" or "lgbt". The right panel shows the proportion of respondents mentioning "others", "other people" or "public". Left-Right views are derived from agreement to each of the topic statements, reverse-coded for the Gender topic. For those who posted one of the two statements, we used the response to the relevant statement. For those who posted neither or both of the two statements, we used the average response of the two statements. Bars represent 95% confidence intervals.

Figure A9: Attitudes in Pilot 1

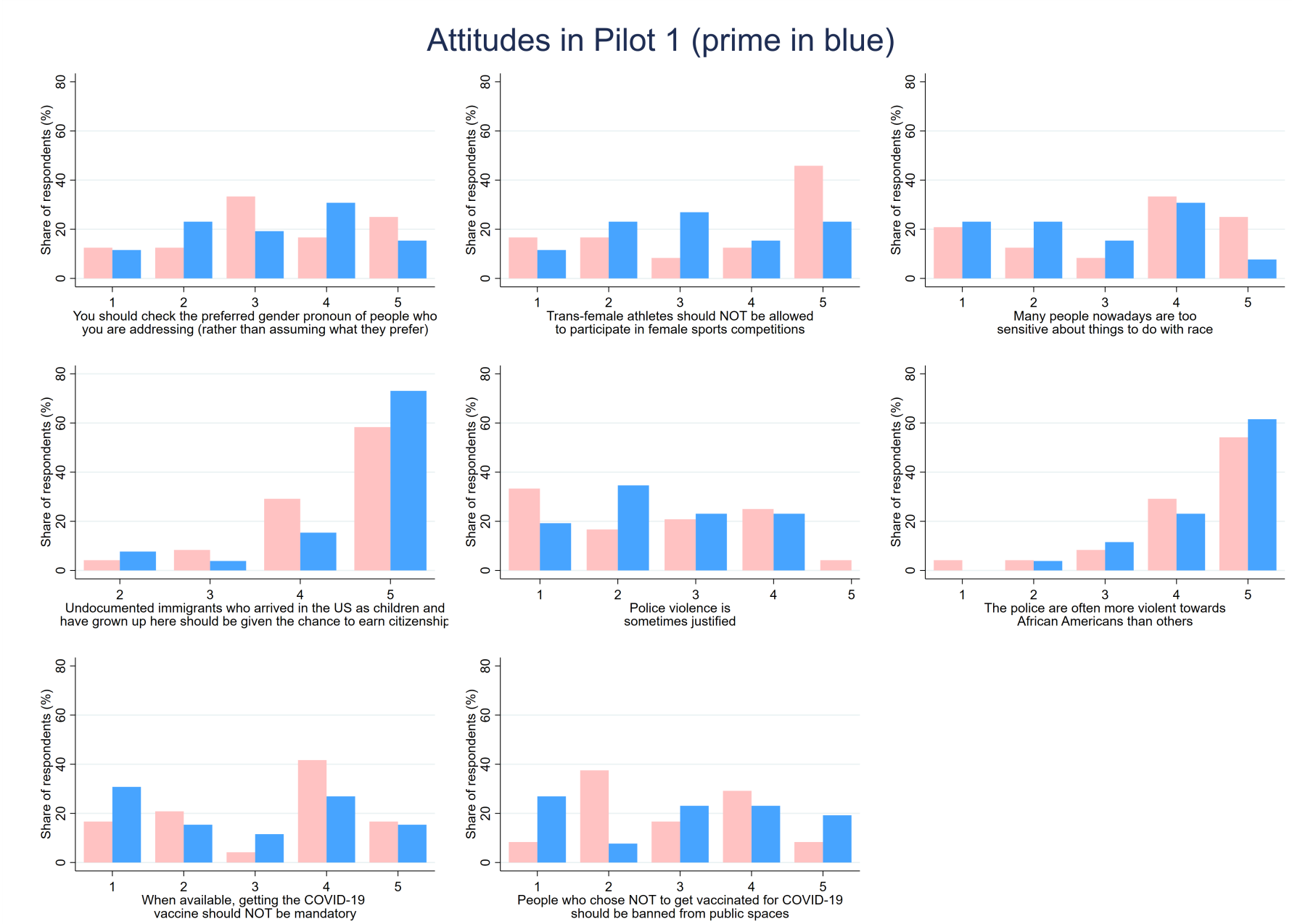


Figure A10: Attitudes in Pilot 2

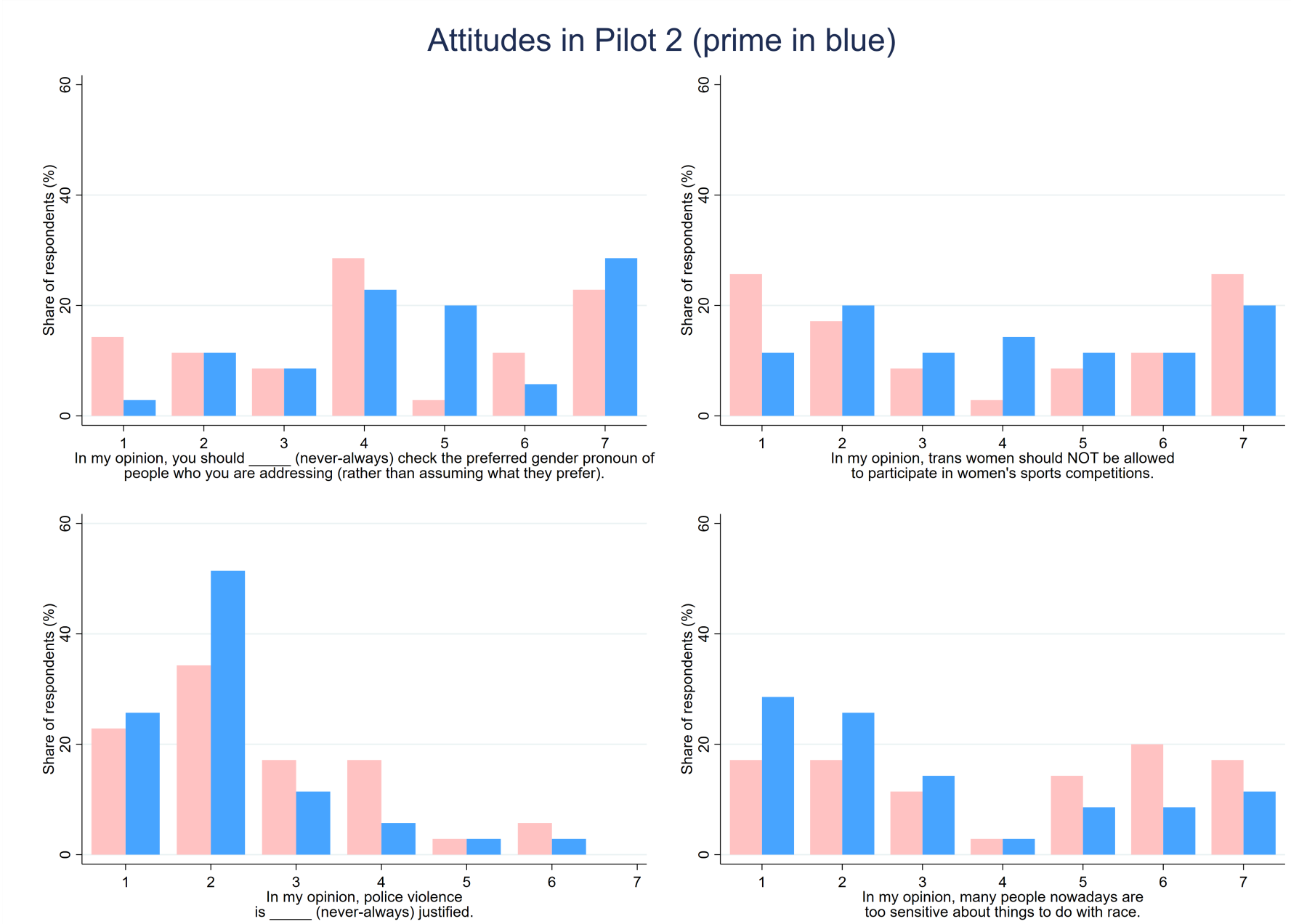
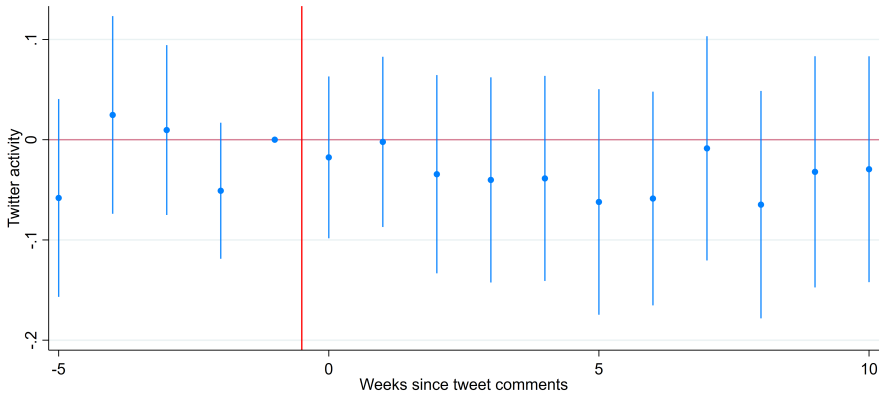


Figure A11: Changes in Twitter activity after negative comments



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.

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 [redacted]

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
[redacted]

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.

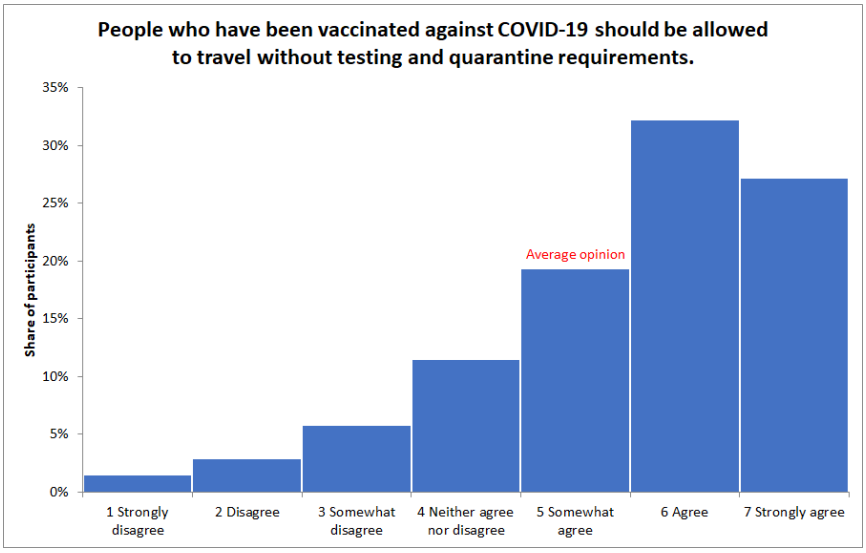
You will now see **2 questions**. You will earn **5 additional lottery tickets** for each question you answer correctly, in addition to your existing tickets.

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 [redacted]