Political Correctness, Social Image, and Information Transmission

Luca Braghieri*

January 9, 2021

Abstract

A prominent argument in the political-correctness debate is that people may feel pressure to publicly espouse socio-political views that they may not privately hold, and that such misrepresentations may render public discourse less vibrant and informative. This paper formalizes the argument in terms of social image and evaluates it experimentally in the context of college campuses. The results show that: i) social image concerns drive a wedge between the sensitive socio-political attitudes that college students report in private and in public; ii) public utterances are indeed less informative than private utterances; iii) information loss is exacerbated by (partial) audience naivete.

*Stanford University. lucabrag@stanford.edu. I thank Matthew Gentzkow and B. Douglas Bernheim for invaluable advising. I thank Hunt Allcott, Sandro Ambuehl, Ned Augenblick, Abhijit Banerjee, Roland Bénabou, Leonardo Bursztyn, Davide Cantoni, Ruben Durante, Sarah Eichmeyer, Marcel Fafchamps, Stefano Fiorin, Thomas Ginn, Jacob Goldin, Muriel Niederle, Kirby Nielsen, Salvatore Nunnari, Collin Raymond, Alvin Roth, Frank Schilbach, Klaus Schmidt, Colin Sullivan, Stefanie Stantcheva, Dmitry Taubinsky, Stephanie Wang, David Yanagizawa-Drott, David Yang and seminar participants at LMU Munich and HSE Moscow for helpful comments. I thank Aman Khinvasara and Dhruv Jatkar for excellent research assistance. I gratefully acknowledge financial support from the George P. Shultz Dissertation Support Fund Fellowship, the B.F. Haley and E.S. Shaw Fellowship for Economics, the Stanford Graduate Research Opportunity Fund, the Stanford Center for American Democracy, and the Stanford Institute for Research in the Social Sciences (IRiSS). The study was approved by the Institutional Review Boards at Stanford (eProtocol #52865). The experiment was registered in the American Economic Association Registry for randomized control trials under trial number AEARCTR-0005063 (https://doi.org/10.1257/rct.5063-1.0). Survey instruments are included in the submission documents. They are also available from https://sites.google.com/view/lucabraghieri/research. Disclosures: I have no relevant or material disclosures.

1 Introduction

Political correctness has been at the center of a heated debate in the U.S. and abroad for more than 30 years (Lea, 2009; Roth, 2019). The terms of the debate are in constant evolution, but a prominent and recurrent argument is that people may feel pressure to publicly espouse views on a set of sensitive socio-political topics that they may not privately hold, and that such misrepresentations may render public discourse less vibrant and informative (Ackerman et al., 2020; Lea, 2009; Loury, 1994; Morris, 2001; Schwarz, 2020; Zimmer, 2016). The debate is particularly heated on college campuses, where the quality of public discourse and the diffusion of information, both among students and between students and faculty, is a matter of first-order importance (Roth, 2019; Zimmer, 2016). Despite the wealth of articles and books making some version of the argument above, there is little direct evidence assessing its merits.[1]

Collecting such direct evidence requires overcoming a crucial challenge: the degree of informativeness of statements made in public is determined not by the literal meaning of the statements, but rather by the meaning that the statements acquire in equilibrium. Specifically, in environments in which public statements are a distorted version of private opinions, a rational listener would of course not take public statements at face value; rather, she would interpret them in light of her understanding of the equilibrium distortions.[2] The presence and extent of information loss after such process of interpretation is not immediately obvious. For instance, if public statements are simply an inflated version of private opinions, a rational listener may be able to invert the equilibrium mapping and recover full information.

This paper develops a tightly connected theoretical and experimental apparatus to study: i) whether and the extent to which, in the context of the political correctness debate, individuals publicly state opinions that they do not privately hold, and ii) the implications of such distortions for the equilibrium informativeness of public discourse. In line with other work in economics, the theoretical framework adopts a social-image interpretation of the phenomenon of political correctness (Loury, 1994; Morris, 2001). The framework shows that social image concerns may distort the sensitive socio-political attitudes that individuals portray in public compared to the ones they hold in private, but that such distortions need not necessarily imply a loss of information in equilibrium. In light of the theoretical indeterminacy, the paper proceeds to study the question empirically by means of an experiment carried out in the context of a university campus. The results of the experiment show that, compared to private statements, students' public statements about topics related

[1] A version of the argument was recently made by Nobel Prize winning economist James Heckman (Schwarz, 2020), by the president of the University of Chicago Robert Zimmer (Zimmer, 2016), and by many prominent academics and writers - including Noam Chomsky, Francis Fukuyama, Jonathan Haidt, Deirdre McCloskey, and Steven Pinker - who signed an open letter published in Harper's Magazine titled "A Letter on Justice and Open Debate" (Ackerman et al., 2020).

[2] As a familiar example, consider letters of recommendation: in the context of a letter of recommendation, a statement whose literal meaning is positive may be understood in equilibrium as actually conveying a negative signal.

to the political correctness debate are on average skewed in the direction that is perceived to be more socially acceptable on campus. Furthermore, the experimental results show that such distortions have two implications for information loss: first, even after taking into account equilibrium considerations, the students' public statements are less informative than the private statements according to a host of measures of informativeness suggested by the theoretical model.[3] Second, the distortions generate an additional degree of information loss, because the natural audience in the environment, namely other college students, do not fully appreciate the ways in which social image distorts their peers' public statements and, as a result, make systematically erroneous inferences.

The theoretical framework underlying the experiment consists in a signaling model with lying costs along the lines of Kartik (2009). The model features an agent who trades off the cost of misrepresenting her private information against the benefit of garnering social esteem by means of the misrepresentation. The equilibrium of the model determines the extent to which statements are informative of the agent's private attitudes, where informativeness is defined according to the canonical Blackwell order (1951, 1953).[4]

The baseline model shows that the distortions induced by social image are a necessary but not sufficient condition for information loss; furthermore, the model highlights the role of heterogeneity in social image concerns in driving information loss. When social image concerns are homogeneous, full separation of types is in principle possible even in the presence of distortions and, as a consequence, misreporting in equilibrium does not necessarily imply information loss. Intuitively, when social image concerns are homogeneous, the model admits equilibria in which public statements are an inflated version of private statements, and in which a sophisticated audience, by applying the appropriate deflator, is able to recover full information. Conversely, when social image concerns are sufficiently heterogeneous, full separation of types is impossible and any degree of misreporting in equilibrium entails a loss of information. An extension of the baseline model also shows that, if the interpretation of natural language is allowed to be heterogeneous across agents, misreporting driven by social image concerns may even lead to gain in informativeness rather than a loss.

The Social Image Experiment, which mimics the decision the agent makes in the model, consists of an online survey administered to students at the University of California Santa Barbara (UCSB) in which we manipulated the students' social image concerns by varying the perceived degree of anonymity of their answers. In one of the treatments (Private Treatment), participants were assured that their answers to the survey would remain completely anonymous; in the other treatment (Public Treatment) participants were not given such assurance of anonymity and were in fact given hints suggesting that their individual-level answers might be shared with other UCSB students

[3]The informativeness of private statements is taken as a natural benchmark against which to measure information loss in public due to social image concerns.

[4]It is worth highlighting that Blackwell informativeness is a demanding criterion of informativeness. Specifically, if an information structure is Blackwell-more-informative than another, any expected-utility maximizer, independently of her prior beliefs about the state of the world and her utility function, is weakly better off gathering information from the former information structure than from the latter.

participating in the study, together with their names. The survey asked all subjects to report, on a scale from 0 to 10, the extent to which they agreed with a set of statements related to recurrent topics in the political correctness debate. The statements covered issues such as the removal of confederate statues, the use of preferred gender pronouns and trigger warnings in classrooms, cultural appropriation, reparations for slavery, and more.

The first set of experimental results shows that social image concerns do indeed drive a wedge between the sensitive socio-political attitudes that UCSB students report in private and in public. Specifically, the answers of subjects in the Public Treatment are on average 0.24 standard deviations closer to the end of the spectrum that participants in an exploratory survey considered more socially acceptable at UCSB than the answers of subjects in the Private Treatment. Furthermore, both manual classification and text-analysis reveal that, when answering an open-ended question about whether it is acceptable to disrupt the talk of controversial speakers invited to campus to deliver lectures, students assigned to the Public Treatment are more likely to argue that the disruptions are acceptable and less likely to discuss issues related to freedom of speech.

The treatment effects differ widely depending on self-reported political ideology. Specifically, the average difference between the answers of participants in the Private and the Public Treatments is significantly larger for students who self-identify as moderate or conservative than for students who self-identify as liberal. As a result, the ideological gap between the opinions of self-identified liberals and self-identified moderates/conservatives is one-and-a-half times larger in the Private Treatment than in the Public Treatment. Therefore, if one were to naively assess the degree of ideological diversity on campus by taking public statements at face value, one would substantially underestimate the breadth of opinions held by students at UCSB regarding the topics covered in the survey.

The second set of results shows that the attitudes students report in public are less informative than ones they report in private even for an audience that does not take public statements at face value, but rather interprets them in light of equilibrium considerations. The theoretical framework underlying the experiment suggests an array of empirical exercises to compare the relative informativeness of the answers of participants in the Private and the Public Treatments. One of the exercises shows that the answers of participants in the Public Treatment are close to being a *garbling* of the answers of participants in the Private Treatment (Blackwell, 1951, 1953). In other words, the distribution of answers from the Public Treatment can be closely mimicked by taking the distribution of answers from the Private Treatment and adding noise to them in a manner consistent with the theoretical framework. Another exercise compares the relative extent to which the answers to the sensitive questions given by students assigned to the Private vs. the Public Treatment are predictive of the students' demographic characteristics and incentivized behaviors. The results show that the predictions made using the answers of students assigned to the Private Treatment are more accurate than the predictions made using the answers of students assigned

to the Public Treatment. A third exercise leverages the information-theoretic concept of *mutual information* to study the magnitude of information loss (Cover and Thomas, 2006). The results show that the mutual information between the students' statements and their demographic characteristics and incentivized behaviors is 25-to-50% lower in the Public Treatment than in the Private Treatment.

The third set of results shows that the natural audience in the environment, namely other college students, do not fully appreciate the ways in which social image distorts their peers' public statements and, as a result, make systematically erroneous inferences.[5] Specifically, a separate experiment - referred to henceforth as the Forecasting Experiment - asked students to forecast the results of the Social Image Experiment and incentivized them for accuracy.[6] Participants in the Forecasting Experiment were asked to make three sets of forecasts: first, for each of the statements from the Social Image Experiment, participants were shown the distribution of answers of subjects assigned to the Public Treatment and were asked to forecast the average answer of subjects assigned to the Private Treatment. Second, participants in the Forecasting Experiment were asked to forecast the dimensions along which the treatment effects in the Social Image Experiment were heterogeneous. Third, participants in the Forecasting Experiment were asked to forecast the answers of students with different self-reported political affiliations.

The results of the Forecasting Experiment show that students exhibit a fairly sophisticated understanding of average effects, but are rather naive about the dimensions of heterogeneity driving them. Specifically, when forecasting treatment effect heterogeneity: i) the majority of participants in the Forecasting Experiment failed to identify self-reported political ideology as the fundamental dimension of heterogeneity driving the average results; ii) a plurality of participants forecasted, contrary to the heterogeneous treatment effect estimates from the Social Image Experiment, that *whites* and *males* exhibit larger treatment effects than their complementary categories. Such incorrect beliefs about heterogeneity have implications for the accuracy of the students' inferences: students in the Forecasting Experiment perceived agreeing with certain political statements in public as being a lot more diagnostic of a speaker's underlying political ideology than it actually is.

Overall, the experimental results lend empirical support to the idea that social image concerns around topics related to political correctness may render public discourse less vibrant and informative. In the context considered, namely a college campus, impediments to information transmission

[5]The second set of results can be thought of as measuring information loss from the perspective of a sophisticated audience that perfectly understands the ways in which social image concerns distort the students' public statements. Conversely, the third set of results can be thought of as measuring information loss from the perspective of the actual audience in the environment, which *may* or *may not* fully understand the ways in which social image concerns affect their peers' public statements.

[6]The Forecasting Experiment is related to recent work that elicits forecasts of experimental results (DellaVigna and Pope, 2018a, 2018b)

arguably fall on the cost side of a welfare calculation.[7] Clearly, the welfare calculation also has a benefit side, which is not captured in this paper. As a consequence, this paper should be seen as a first step in a broader welfare analysis of the phenomenon of political correctness.

This project relates to various strands of the literature. On the theoretical side, Morris (2001) showed that, in a cheap talk game with incomplete information about the sender's type, the sender's concerns for social image may limit the amount of information that she can transmit to the uninformed receiver. This paper develops and extends the insights from Morris (2001): first, in order to fit the experimental design and to draw a sharp distinction between the literal and the equilibrium meaning of language, the framework relies on a signaling model with lying costs along the lines of Kartik (2009) rather than a cheap talk game. Second, and in line with the recent work of Ali and Bénabou (2020), the model highlights the importance of heterogeneity in social image concerns in driving information loss.[8] Specifically, beside the mechanism for information loss explored in Morris (2001), the model introduced in this paper identifies a second and distinct mechanism, related to social image heterogeneity, whereby the social image concerns related to political correctness may lead to information loss.[9] Third, an extension of the model shows that the distortions caused by social image concerns may even lead to a gain in informativeness if the interpretation of natural language is allowed to be heterogeneous across agents.

On the empirical side, the two papers that are most closely related to this project are Bursztyn, González, and Yanagizawa-Drott (2020b) and Bursztyn, Egorov and Fiorin (2020a).[10] The former paper shows that young men in Saudi Arabia underestimated the level of support for female labor force participation (FLFP) among their peers, and that an information intervention that randomly disclosed the actual popularity of FLFP increased married men's willingness to let their wives join the labor force. The latter paper shows that Donald Trump's rise in popularity and eventual victory in the 2016 presidential election increased the proportion of people willing to publicly display anti-immigration attitudes, and suggests the results are due the election precipitating common knowledge about the prevalence of anti-immigration sentiments in the population. Contrary to these two articles, which study the behavioral implications of people suddenly updating their perceptions of the attitudes held by their peers, this paper investigates the mechanisms whereby social image

[7] One can easily imagine situations in which limiting information transmission may fall on the benefit side of a welfare calculation. For instance, in ordinary times, limiting the transmission of information about how to construct rudimentary explosive devices is arguably welfare improving.

[8] Heterogeneity in social image concerns also differentiates the model from the one in Kartik (2009).

[9] Frankel and Kartik (2019) and Ball (2020) develop related theoretical insights in a different context.

[10] For additional field evidence on the effects of social image, see also Alpizar, Carlsson, and Johansson-Stenman (2008), Banerjee et al. (2020), Barker (1994), Bursztyn, Fujiwara and Pallais (2017), Bursztyn et al. (2018), Chandrasekhar, Golub and Yang (2019), DellaVigna, List and Malmendier (2012), DellaVigna et al. (2017), Enikolopov et al. (2018), Funk (2010), and Gerber, Green, and Larimer (2008), Karing (2018). For laboratory evidence, see Andreoni and Bernheim (2009), Andreoni and Petrie (2004), Ariely, Bracha, and Meier (2009), Friedrichsen, König, and Schmacker (2018), McManus and Rao (2015), Montano-Campos and Perez-Truglia (2019), and Rege and Telle (2004).

concerns may induce people to hold incorrect beliefs in the first place.[11]

Finally, this paper relates to the large literature in the social sciences about social desirability bias in survey responses (Edwards, 1957; Folsom et al. 1973; Greenberg et al. 1971; Krumpal, 2013; Raghavarao and Federer, 1979; Sniderman and Piazza 1993; Sudman and Bradburn, 1974; Tourangeau, Rips, and Rasinski 2000; Warner, 1965, 1971). This paper enriches the literature by relating social desirability bias in surveys to a formal signaling model. The theoretical framework develops insights that are novel to the literature on social desirability. Specifically, to the best of my knowledge, this paper is the first that: i) formalizes the distinction between the literal meaning of language and the equilibrium meaning of language in the context of social desirability in surveys; ii) relates survey answers to the canonical notion of Blackwell informativeness; iii) shows that the presence of social desirability bias need not necessarily imply information loss; iv) highlights the importance of heterogeneity in social image concerns in driving information loss; v) suggests various measures of informativeness that can be estimated on empirical data to compare the extent of information loss under different elicitation mechanisms; vi) studies information loss also from the perspective of the natural audience in the environment, who may not fully understand the ways in which social image concerns may distort the attitudes that speakers portay in public.

The remainder of the paper is organized as follows: Section 2 introduces the motivating framework; Section 3 provides some background information about the Social Image Experiment and the environment in which it is run; Section 4 presents the design of the Social Image Experiment and shows results about whether and the extent to which social image concerns distort the attitudes that students report in public compared to the ones they report in private; Section 5 discusses the implications of such distortions for information loss; Section 6 presents the design and results of the Forecasting Experiment; Section 7 concludes.

2 Motivating Framework

In order to formalize the notion of informativeness and to analyze the relationship between social image concerns and information loss in the context of the experiment, we develop a signaling model with lying costs along the lines of Kartik (2009). All proofs are omitted from the main manuscript and relegated to Appendix A.

[11] Other strands of the economic literature that study related but slightly more distant phenomena are the body of research that studies frictions to information diffusion (Beaman and Dillon, 2018; Cullen and Perez-Truglia, 2020; Mobius, Phan, Szeidl, 2015; Mobius and Rosenblat, 2014), the one that explores the effects of transparency on the quality of decision making in committees (Gradwohl and Feddersen, 2018; Gradwohl, 2018; Fehrler and Hughes,2018; Name-Correa and Yildirim, 2017; Renes and Visser, 2018), and the one that analyzes reputational cheap talk games, in which forecasters want to signal their forecasting ability (Meloso, Nunnari and Ottaviani, 2018; Ottaviani and Sørensen, 2006). See also Kuran (1995) and related work.

2.1 Model Setup

The backbone of the model is a reduced-form signaling game in which an agent trades off the cost of misrepresenting her private information against the benefit of garnering social esteem by means of the misrepresentation. The setup aims to capture the idea that, whenever one is asked to publicly make a statement about a sensitive topic, the desire to truthfully report one's private opinion might conflict with the desire to make a statement that may garner greater social approbation (Sudman and Bradburn, 1974).

The following example, taken from the experiment, will help build intuition for the theoretical framework. One of the questions in the experiment asks subjects to state the extent to which they agree that the university administration should mandate yearly sexual harassment training. We can think of each subject having a private and unobservable level of agreement with the proposed policy (denoted by π in the model below) and making a public and observable report about her level of agreement (denoted by x in the model below). Social image concerns may induce the subject to report a level of agreement with the proposed policy that differs from her private level of agreement. A rational audience, who observes the subject's report but not her private level of agreement, will not take the report at face value; rather, the audience will understand that the report may be a distorted version of the subject's private level of agreement and will try to infer the private level of agreement from the report.

Formally, let there be an agent who has to make some statement $x \in X$ about a potentially sensitive topic. Let $X = \{1, 2, ..., \bar{x}\}$, where $\bar{x} \in \mathbb{N}$, $\bar{x} \geq 3$. Let the agent be endowed with two-dimensional private information - her type - represented by $(\pi, \varsigma) \in \Pi \times \Sigma = \{1, 2, ..., \bar{\pi}\} \times [0, \bar{\varsigma}]$, where $\bar{\pi} = \bar{x}$ and $\bar{\varsigma} > 0$. Let π (pi) stand for the agent's *private issue-position* - the agent's private opinion about the topic - and let ς (sigma) stand for the agent's *social-image susceptibility* - the extent to which the agent is susceptible to the demands of social image relative to the misrepresentation costs. Let (π, ς) be drawn from cumulative distribution F with full support and let the marginal distribution of π be denoted by β.

In the baseline model, we think of statements $x \in X = \Pi$ as having exogenous commonly-understood meaning "my private issue-position is x".[12] As a consequence, we say that an agent is misrepresenting her private issue-position whenever she chooses a statement $x \neq \pi$. We assume the agent pays a cost whenever she misrepresents her private issue-position and we denote such misrepresentation costs by $c(x, \pi)$. $c(x, \pi)$ can be thought of as a psychological cost related to an aversion to lying or to the cognitive dissonance of making a statement that does not correspond to one's private beliefs. We assume $c(x, \pi)$ is single-troughed, symmetric around $x = \pi$, and convex.[13]

[12] We relax this assumption in Appendix B and allow the interpretation of natural language to be heterogeneous across agents.

[13] The weaker assumption of strict quasi-convexity is sufficient to prove Proposition 1 and the first half of Proposition 2.

For simplicity, we let $c(x, \pi) = 0$ when $x = \pi$.

After the agent makes statement $x \in X$, an audience forms posterior belief $\beta|x \in \Delta(\Pi)$ about the agent's private issue-position, where $\Delta(\Pi)$ denotes the set of probability distributions over Π. We assume the social image component of the agent's payoff function depends on the audience's conditional expectation of the agent's private issue-position given her statement. Denote such conditional expectation by $E_\beta(\pi|x)$.

We let the agent's utility depend positively on $E_\beta(\pi|x)$; i.e. we assume the environment is one where being perceived as having a higher issue-position garners a higher degree of social esteem. It is worth noting that there are in principle many alternative ways to model the agent's social image concerns. For instance, the agent could be modeled as caring about some functional of β other than the expectation; similarly, the agent could be modeled as caring about the extent to which the audience perceives her private issue-position as close to the average in the population; etc. The chosen formulation was selected for its plausibility in the experimental environment and for the sake of tractability. As discussed in Section 2.5, the key mechanisms for information loss that emerge from the model are quite general and do not depend on the specific formulation of social image adopted in this model.

Overall, we let the agent's payoff function be

$$u(x, \pi, \varsigma) = \varsigma \cdot s \cdot r \cdot E_\beta(\pi|x) - c(x, \pi)$$

where s and r are the two environmental parameters that are manipulated in the experiment. Specifically, $s \in [0, 1]$ denotes the extent to which, in the environment at hand, the topic the agent is asked to make a statement about is considered sensitive, and $r \in [0, 1]$ denotes the fraction of the natural audience in the environment that is expected to learn about the statement made by the agent.

For the sake of notational compactness, we let $e = s \cdot r \in [0, 1]$ and consider comparative statics directly on e. We refer to e as the extent to which the environment engages the agent's social image concerns.

The agent's payoff function, together with the prior distribution over types F, induces a signaling game. The equilibrium analysis focuses on (weak) Perfect-Bayesian Equilibria, which we refer to henceforth simply as equilibria. In this model, an equilibrium consists of a profile of statements $\psi : \Pi \times \Sigma \to \Delta(X)$ and a belief mapping $\phi : X \to \Delta(\Pi)$ satisfying the following conditions: the profile of statements ψ must be optimal given the audience's beliefs $\beta|x$, and the belief mapping ϕ must be consistent with Bayes' rule on the equilibrium path. No restrictions on beliefs are imposed off the equilibrium path.

2.2 Defining Informativeness

Equilibrium statements reveal information about the agent's type (π, ς). In the context of the debate about political correctness, we assume the audience is interested in learning about the agent's private issue-position π.[14] π is a natural object of interest whenever the audience cares directly about learning the agent's private issue-position on a particular topic or whenever the audience cares about gathering information about any random variable that correlates with π. For instance, the agent's private issue-position may be correlated with some other unobservable trait that the audience may care about. Alternatively, the agent's private issue-position π may be correlated with a signal that the agent observed, and the audience may be interested in learning about the realization of that signal so as to gather information about the state of the world. Finally, π may be correlated with the agent's private behavior, and the audience may be interested in predicting such behavior.

The example about sexual harassment training introduced in the previous section may once again help build intuition. In the example, a subject is asked to publicly report her level of agreement with a university-mandated yearly sexual harassment training. Since social image concerns may cause the subject's report (x) to differ from her private level of agreement with the proposed policy (π), a rational audience, who observes x but not π, will try to infer π from x. The audience may have a direct interest in learning about the agent's private issue-position π on the topic, or an indirect interest in π arising from the fact that π may be correlated with some other random variable that the audience cares about. In the example, π could be correlated with the agent's degree of misogyny and the audience may be interested in learning about that unobservable trait. Alternatively, the agent may have observed a signal about the effectiveness of sexual harassment training in curtailing the problem of sexual harassment, that signal may have contributed to shaping the agent's private issue-position π, and the audience may be interested in gathering information about the realization of the signal. Lastly, the audience may be interested in predicting whether the agent would attend the sexual harassment training and pay attention if such training was in fact mandated.

The extent to which the agents' statements are informative about π depends on the equilibrium distribution of $\beta|x$. From the ex-ante perspective, the distribution of $\beta|x$ is an element of $\Delta(\Delta(\Pi))$. Bayesian rationality requires that, in every equilibrium, $\beta|x$ satisfy the martingale property of beliefs, namely $E[\beta|x] = \beta$. Other than that, the distribution of $\beta|x$ depends on the nature of the equilibrium. For instance, equilibria in which agents with different private issue-positions make different statements are fully informative about π; conversely, equilibria in which agents with different private issue-positions all make the same statement are completely uninformative about π.

[14] One can in principle imagine scenarios where an audience could be interested in learning about the agent's social-image susceptibility ς.

We formalize the notion of relative informativeness by appealing to the canonical partial order of Blackwell (Blackwell, 1951, 1953). Specifically, we say that equilibrium $\mathcal{E}$ is more informative about π than equilibrium $\mathcal{E}'$ if the distribution of $\beta|x$ under equilibrium $\mathcal{E}$ is Blackwell more informative than the distribution of $\beta|x$ under equilibrium $\mathcal{E}'$. As discussed, Blackwell informativeness is a demanding criterion of informativeness. Specifically, if an information structure is Blackwell-more-informative than another, any expected-utility maximizer, independently of her prior beliefs about the state of the world and her utility function, is weakly better off gathering information from the former information structure than from the latter.

We conclude this section by stating a simple result that we will leverage in the experiment.[15] Specifically, in the experiment we will not be able to study the extent to which statements (x) are informative about private issue-positions (π), because the subjects' private issue-positions are private information and, as such, they are not directly observable. Instead, we will study the extent to which statements (x) are informative about observable characteristics and private behaviors (τ) that we expect to be correlated with the subjects' private issue-positions (π). The result below links information loss about π to information loss about τ.

Claim. $\forall$ equilibria $\mathcal{E}$ and $\mathcal{E}'$ with $\mathcal{E}$ more informative about π than $\mathcal{E}'$ and $\forall$ random variables $\tau : \Omega \to \mathbb{R}$ with finite support correlated with π and, conditional on π, independent of ς, $\mathcal{E}$ is more informative about τ than $\mathcal{E}'$.[16,17]

2.3 Equilibrium Analysis

Before proceeding with the equilibrium analysis, we make an assumption about the misrepresentation costs $c(x, \pi)$:

Assumption 1. $\forall x \in X, \pi \in \Pi$, *with* $x \neq \pi$, *we assume* $\frac{1}{2}|x - \pi|\bar{\varsigma} < c(x, \pi) < |x - \pi|\bar{\varsigma}$.

The upper bound on the misrepresentation cost schedule makes the model non-trivial. If the upper bound was violated, the agent would trivially report her private issue-position truthfully independently on the extent $e \in [0, 1]$ to which the environment engages her social image concerns. The lower bound on the misrepresentation cost schedule helps ensure that types who misrepresent their private information in equilibrium do so in the direction corresponding to private issue-positions that garner higher social esteem.[18]

[15] In order to conserve space, the proof of the claim below is omitted. It is available upon request.

[16] Informativeness about τ is defined in a similar way as informativeness about π. Specifically, for any random variable $\tau : \Omega \to \mathbb{R}$, with distribution $\gamma \in \Delta(\Omega)$ and finite support, we say that equilibrium $\mathcal{E}$ is more informative about τ than equilibrium $\mathcal{E}'$ if the distribution of $\gamma|x$ under equilibrium $\mathcal{E}$ is Blackwell more informative than the distribution of $\gamma|x$ under equilibrium $\mathcal{E}'$.

[17] The requirement that, conditional on π, τ be independent of ς is to rule out the possibility that the audience may gather information about τ by using equilibrium statements to learn about ς.

[18] When $\bar{\pi} > 3$, there exist distributions $F \in \Delta(\Pi \times [0, \bar{\varsigma}])$, environments $e \in [0, 1]$, and misrepresentation cost schedules $c(\pi, \varsigma)$ that, in the absence of the lower bound from Assumption 1, give rise to equilibria in which some

The equilibrium analysis shows that an equilibrium exists and that all equilibria share a set of features that is typical of signaling models with lying costs similar to the one in Kartik (2009): in equilibrium, all types either report their private issue-positions truthfully or make statements that correspond to issue-positions that garner higher social esteem than the issue-position they actually hold. We refer to this phenomenon as "misreporting in the socially acceptable direction". Finally, we show that the nature of the equilibrium depends crucially on e, the parameter summarizing the extent to which the environment engages the agent's social image concerns. Intuitively, environments that do not substantially engage the agent's social image concerns (e small) do not lead to misreporting in equilibrium; conversely, environments that do engage the agent's social image concerns (e large) do lead to misreporting in equilibrium.

Proposition 1. $\forall e \in [0,1]$, *an equilibrium exists. In all equilibria, if any misreporting occurs, it occurs in the socially acceptable direction. Furthermore,* $\exists e^* \in (0,1)$ *s.t.* $\forall e < e^*$, *no misreporting occurs in equilibrium and* $\forall e > e^*$ *the equilibrium involves some misreporting.*

2.4 Implications for Information Loss

Proposition 1 suggests that the degree of informativeness of equilibrium statements depends on the extent to which the environment engages the agent's social image concerns. Trivially, when the environment does not substantially engage the agent's social image concerns ($e < e^*$), the equilibrium reveals full information about π simply because, in equilibrium, all types of the agent truthfully report their private issue-positions. Formally, we say an equilibrium is fully informative about π when types with different private issue-positions make different equilibrium statements.

Corollary 1. $\forall e < e^*$, *all equilibria are fully informative about* π.

When the environment does engage the agent's social image concerns ($e > e^*$), we know from Proposition 1 that some misreporting occurs in equilibrium. Importantly, equilibrium misreporting does not necessarily imply information loss: depending on the primitives of the model, misreporting can lead to anything ranging from no information loss to complete information loss.

In order to shed light on the determinants of information loss and to separately identify two important mechanisms driving it, it is worth comparing two cases. The first case corresponds to the version of the model considered thus far, where social-image susceptibility ς is heterogeneous and the agent's type is two-dimensional. The second case corresponds to a more standard signaling game, where social-image susceptibility is homogeneous and the agent's type is one-dimensional. Formally, we consider:

types of the agent unintuitively misrepresent their private information in the direction corresponding to private issue-positions that would normally garner lower - rather than higher - social esteem. The intuition for such behavior is that agents with a high private issue-position and high social image susceptibility may take advantage of areas of low density in the distribution of types F to separate from agents with slightly lower private-issue positions and lower social image susceptibility.

Case 1. F is such that the marginal distribution of ς has full support on $[0, \bar{\varsigma}]$.

Case 2. F is such that the marginal distribution of ς is degenerate and puts probability mass equal to unity on $\bar{\varsigma}$.

As shown in the next proposition, the relationship between misreporting and information loss is quite different depending on the degree of heterogeneity in social-image susceptibility. When social-image susceptibility is sufficiently heterogeneous (Case 1), full separation of types along the private issue-position dimension is impossible and any degree of misreporting in equilibrium implies information loss. Conversely, when social-image susceptibility is homogeneous (Case 2), full separation of types along the private issue-position dimension is in principle possible even in the presence of distortions and, as a consequence, misreporting in equilibrium does not necessarily imply information loss.

Before stating the next proposition, we impose an additional assumption on F:

Assumption 2. *F puts zero probability mass on $\bar{\pi}$; i.e. $\beta(\bar{\pi}) = 0$.*

Assumption 2 effectively creates some slack in the message space so that language can become inflated in equilibrium without necessarily generating pooling at the upper boundary of the message space.

Proposition 2. *In Case 1, no equilibrium in which a positive measure of types misreport their private issue-positions is fully informative about π. Conversely, in Case 2, $\exists e \in [0, 1]$ sustaining equilibria that involve a positive measure of types misreporting their private issue-positions, but that, nonetheless, are fully informative about π.*

The intuition for why the relationship between misreporting and information loss is different in the two cases is as follows. In Case 1, heterogeneity in social-image susceptibility makes it impossible, in equilibrium, to tell apart truthful statements made by types with low social-image susceptibility and misreports made by types with high social-image susceptibility. In Case 2, homogeneity in social-image susceptibility leads to homogeneity in language inflation, thus making it in principle possible to recover full information about π by inverting the equilibrium mapping between private issue-positions and statements.[19]

The comparison between Case 1 and Case 2 highlights the existence of two separate mechanisms whereby social image concerns can lead to information loss. The first mechanism relates to the well-known phenomenon of *pooling* in signaling games: social image concerns may induce types with different private issue-positions to pool on statements that garner high social esteem, thus

[19]It is worth noting that, in Case 2, the separating equilibrium involving misreporting does not survive the D1 criterion of Banks and Sobel (1987). The equilibrium can be shown to survive the D1 criterion if a certain fraction of the audience is credulous and takes statements at face value. Letting a certain fraction of the audience be credulous does not qualitatively affect the conclusions from Case 1. See Section 2.5 for further details.

making it impossible to distinguish such types. Intuitively, pooling relates to the idea of conformity; specifically, in an equilibrium involving pooling, there exists a particular view that many individuals publicly conform to. The second mechanism, which we refer to as *scrambling*, is related to the idea that, when social-image susceptibility is sufficiently heterogeneous, types with high social-image susceptibility who misrepresent their private issue-positions cannot be distinguished from types with low social-image susceptibility who truthfully report their private issue-positions.[20] Intuitively, scrambling does not refer to the existence of a particular view that many individuals conform to when in public; rather, it refers to the statistical noise generated by the fact that, for virtually all possible views on an issue, multiple types portray that view in public. The two mechanisms are distinct: in the context of the model, pooling arises independently of whether social image concerns are homogeneous or heterogeneous and is weakly order-preserving in the sense that agents with higher private issue-positions make weakly higher equilibrium statements. Conversely, scrambling arises only when social image concerns are heterogeneous and is not order-preserving.

Thus far, we assumed the interpretation of natural language - i.e., the mapping between x and π in the absence of distortions - is homogeneous across agents. Appendix B relaxes the assumption and shows that, if the mapping is allowed to be heterogeneous across agents, the distortions caused by social image concerns may in fact lead to an information gain rather than a loss. The intuition behind the result is simple: heterogeneity in the interpretation of natural language may generate an inadvertent degree of scrambling and pooling even in the absence of distortions. In some such cases, the distortions caused by social image concerns may in fact counteract the inadvertent degree of scrambling and pooling and lead to an information gain.

2.5 Discussion

Overall, the theoretical framework shows that environments that sufficiently engage the agents' social image concerns distort the statements agents make in equilibrium, but that the presence of equilibrium distortions, while necessary, is not a sufficient condition for information loss. Furthermore, the extension of the model developed in Appendix B shows that, if the interpretation of natural language is allowed to be heterogeneous across agents, the distortions caused by social image concerns may even lead to an increase in informativeness. Therefore, the theoretical framework suggests that, in the context of the political correctness debate, the presence and extent of information loss due to social image concerns is ultimately an empirical question.

It is worth noting that the theoretical framework introduced in this section is tailored to the experimental design, but that the intuitions are quite general. In a variety of models in which social image concerns are assumed to be homogeneous, the equilibrium distortions caused by social image concerns need not necessarily lead to information loss. Specifically, under suitable assumptions,

[20] The idea of scrambling is closely related to mechanisms for information loss in Frankel and Kartik (2019) and Ali and Bénabou (2020).

separating equilibria exist in models similar to the one in this paper, but that feature uncountably infinite type spaces, whether bounded or unbounded (Bernheim, 1994; Kartik, Ottaviani, and Squintani, 2007). Furthermore, as shown in Bernheim (1994), such equilibria can survive the D1 criterion of Banks and Sobel (1987). Similarly, information loss due to pooling is not an artifact of the particular formalization of social image adopted in this paper whereby agents like to be perceived as having a high private issue-position; in related models, pooling may occur at the boundary of the type space, or in the interior of the type space (Bernheim, 1994; Kartik, 2009). Lastly, the intuition that heterogeneity in social image concerns is an important driver of information loss holds true even in models with uncountably infinite and unbounded type spaces (Ali and Bénabou, 2020).

3 Background Information about the Social Image Experiment

The experiment was run at the University of California Santa Barbara (UCSB). According to Niche, a company that gathers information about colleges to generate college-rankings, UCSB falls in the 16^{th} percentile in an ordered list of colleges from most liberal to most conservative (Niche, 2020).[21] Among colleges that are at least as selective as UCSB in terms of admission rate, UCSB does not seem to be an outlier in terms of the political leaning of its students. Appendix Table A4 shows a list of selected universities that, together with UCSB, are found in the top quintile of the liberal-conservative distribution according to the Niche ranking.

Two of the main desiderata for designing the experiment were: first, to find a set of topics related to the political correctness debate that students at UCSB would consider sensitive; second, for each sensitive topic, to determine the opinions that are considered to be socially acceptable at UCSB and the opinions that are not.[22] An Exploratory Survey, described in detail in Appendix C, helped us achieve both goals.[23] Specifically, the Exploratory Survey allowed us to select 10 statements related to recurrent topics in the political correctness debate that students at UCSB would likely consider sensitive (*sensitive statements*) and 5 statements that they would likely not consider sensitive (*placebo statements*). Furthermore, for each of the sensitive and the placebo statements, the Exploratory Survey provided information as to whether, at UCSB, agreeing with the statement is perceived to be more socially acceptable than disagreeing with it, or vice-versa (*direction of social acceptability*).

[21] The ranking is calculated by surveying a sample of students from each college and asking them both about their personal political leaning and about their beliefs about the political leanings of the other students at the college (Niche, 2020). In the ranking, the 1^{st} percentile corresponds to the most liberal colleges and the 100^{th} percentile to the most conservative.

[22] In the context of the experiment, we think of an issue as being "sensitive" if taking a certain stand on it is perceived to be significantly more socially acceptable than taking the opposite stand.

[23] The text of the Exploratory Survey is included in the submission documents. It is also available from https://sites.google.com/view/lucabraghieri/research.

4 Effects of Social Image on the Attitudes Reported by College Students

In this section, we describe the design of the Social Image Experiment and we present the results showing that social image concerns drive a wedge between the sensitive socio-political attitudes that UCSB students report in private and in public. In Section 5, we study implications of such distortions for information loss.

4.1 Experimental Design

In late November 2019, we recruited subjects through the Experimental and Behavioral Economics Laboratory (EBEL) portal at the University of California Santa Barbara (UCSB) to complete a short online survey.[24] The recruitment email did not mention the topic of the study. Before being directed to the consent form, subjects were asked to complete a brief pre-screen questionnaire that contained a few demographic questions and a short news-knowledge quiz. The aim of the pre-screen questionnaire was to screen out ineligible and low-quality survey-takers. A subject was considered eligible to participate in the study if she reported being an undergraduate student at UCSB and had not taken the Exploratory Survey; the subject was considered a high-quality survey-taker if she spent at least 30 seconds on the news-knowledge quiz.

After completing the pre-screen survey and consenting to participate in the experiment, subjects were informed that they would be entered into a lottery. The lottery would randomly select a participant to receive $100 and play a dictator game with the American Association of University Women (AAUW), a national non-for-profit organization that, among other activities, helps women who experienced sexual harassment in higher education connect with legal resources and afford legal fees. All subjects were asked how much of the $100 they would be willing to anonymously share with the AAUW if they were randomly selected and were informed that, if they were indeed selected, their decision would be automatically implemented.

Next, subjects were randomized in equal proportions into one of two treatments. In both treatments, subjects were informed that the results of the subsequent section of the survey would be shared with approximately 200 UCSB students scheduled to participate in the next phase of the study.[25] Importantly, in one of the treatments (Private Treatment), subjects were informed that only aggregate-level answers would be shared with participants in the next phase of the study and were given a guarantee of anonymity. Specifically, subjects assigned to the Private Treatment were shown a screen that read:

[24] The text of the Main Survey is included in the submission documents. It is also available from https://sites.google.com/view/lucabraghieri/research.

[25] The answers of participants in the Social Image Experiment were indeed shared with approximately 200 UCSB students who participated in the Forecasting Experiment.

> **IMPORTANT:** We will share **aggregate-level answers** to the questionnaire on the next screen with approximately 200 UCSB students who are scheduled to participate in the next phase of the study but no-one, not even the research team, will match your individual answers to any information that may identify you. **Your answers to this survey are thus completely anonymous.**

In contrast, subjects assigned to the other treatment (Public Treatment) were told that their individual-level answers would be shared with participants in the next phase of the study and were not given a guarantee of anonymity. In fact, as in Bursztyn, Egorov and Fiorin (2020a), subjects in the Public Treatment were shown instructions designed to lead them to doubt that their answers would be kept confidential. Specifically, subjects assigned to the Public Treatment were shown a screen that read:

> **IMPORTANT:** We will share **your individual answers** to the questionnaire on the next screen, as well as the individual-level answers of the other participants in this phase of the study, with approximately 200 UCSB students who are scheduled to participate in the next phase of the study. **There is no need to provide your first and last name here; your information is already in the Experimental and Behavioral Economics Laboratory (EBEL) system.**

In practice, even in the treatment we refer to as the Public Treatment, the subjects' answers were fully anonymized before being shared with the participants in the subsequent phase of the study. Importantly, however, the subjects' answers were not reported in aggregate form: instead, each subject's individual-level answers, together with the subject's (anonymous) participant ID, were shared with participants in the subsequent phase of the study.[26]

The experimental manipulation described above aimed to vary the extent to which subjects were concerned about social image when answering the subsequent set of questions. Subjects in the Private Treatment were not expected to be very concerned with seeking social approbation when answering the questions because of the guaranteed anonymity of their responses. Conversely, doubts about the confidentiality of their responses were expected to induce subjects in the Public Treatment to skew their answers in the direction that they thought would induce the UCSB students scheduled to participate in the next phase of the study to perceive them more favorably.[27] Since

[26] Notice the statements in both conditions are factually true: the Experimental and Behavioral Economics Laboratory system at UCSB does contain the first and last name of each student that participated in the experiment, and the answers of participants in the Private Treatment were in fact anonymous because each participant's answers to the sensitive questions were identified only by an arbitrary participant number. The manipulation is the same as in Bursztyn, Egorov and Fiorin (2020a).

[27] Notice it is not necessary for subjects in the Private Treatment to believe with probability one that their answers will be kept anonymous. For the experimental manipulation to have an effect, it is only necessary that subjects in the Public Treatment be sufficiently more concerned about social image when answering the questions than subjects in the Private Treatment. Furthermore, notice that virtually the entire analysis does not require that students assigned to the Private Treatment truthfully report their private level of agreement with each of the statements.

the experimental manipulation is arguably fairly subtle, finding large effects would be especially remarkable and indicative of the importance of social image in this context.

After being assigned to the Private or the Public Treatment and reading the corresponding instructions, participants were asked to report, on sliders from 0 to 10, where 0 stood for "strongly disagree" and 10 stood for "strongly agree", the extent to which they agreed or disagreed with 15 statements. The order of the statements was randomized at the participant level. Of the 15 statements, 10 were pre-registered as sensitive (and thus likely affected by social image) and 5 were pre-registered as placebos (and thus likely not affected by social image). The sensitive and placebo statements were selected based on the answers of participants in the Exploratory Survey – as described in Appendix C – and are presented in Table 1.

After reporting their levels of agreement with each of the 15 statements, participants were asked an open-ended question about whether it is acceptable to disrupt the talks of controversial speakers invited to campus to deliver lectures.

Upon completing the survey, subjects received a $5 electronic gift-card for their participation.

Around two weeks after the end of the experiment, the students who reached the randomization stage received an email asking them whether they wanted to anonymously support a petition requesting that the UCSB administration mandate yearly sexual harassment training for everybody who works or studies at UCSB.[28] The students were informed that the results of the petition would be sent to the Office of the Chancellor at UCSB.[29]

4.1.1 Discussion of the Experimental Design

The experimental design aims to elicit two sets of variables from the students: the students' reported levels of agreement with each of the statements and some demographic characteristics and incentivized behaviors.

Asking students to report on a 0-10 scale the extent to which they agree with a certain statement mimics the decision the agent makes in the model. Each student can be thought of as having a private level of agreement with the statement, denoted by π in the model, and reporting a certain level of agreement in the experiment, denoted by x in the model. The treatment manipulations vary two dimensions that are expected to affect the extent to which the environment engages the students' social image concerns and, as a consequence, the students' reported levels of agreement. First, treatment assignment manipulates the perceived degree of anonymity of the students' answers. Specifically, the Private Treatment should induce students to believe that the expected fraction of the natural audience in the environment, namely other college students at UCSB, who will be able to observe their answers is relatively small; conversely, the Public Treatment should

[28] A screenshot of the email is included in the submission documents. It is also available from https://sites.google.com/view/lucabraghieri/research.

[29] The results of the petition were sent to the Office of the Chancellor at UCSB in January 2020.

induce students to believe that the expected fraction of the natural audience in the environment who will be able to observe their answers is relatively large. Second, variation in the extent to which the statements are considered sensitive at UCSB should affect the degree to which the students' social image is at stake when answering the questions. Therefore, for the sensitive but not the placebo statements, being assigned to the Public Treatment should induce students to give answers that are skewed in the direction that, according to the answers of participants in the Exploratory Survey, is generally perceived to be more socially acceptable at UCSB.

As argued in the theoretical framework, equilibrium distortions induced by social image concerns are a necessary but not sufficient condition for information loss. In fact, as shown in the extension of the model in Appendix B, such distortion may even lead to an information gain. In order to address the question of whether the distortions found in the experiment, if any, are such as to lead to information loss, the experimental design elicits a host of observable demographic and behavioral measures that are expected to be correlated with the students' unobservable private levels of agreement with the sensitive statements. Such observable characteristics and incentivized behaviors, denoted by τ in model, are: self-reported political ideology, donation rates to the AAUW, and signatures to the petition requiring that UCSB mandate yearly sexual harassment training.[30] One of the main exercises in the information loss section of the analysis will compare the relative extent to which the answers (x) of students in the Private and the Public Treatments are informative of the demographic characteristics and incentivized behaviors (τ).

4.2 Outcome Variables

The outcome variables in the main study are the *donation rates* to the AAUW, the *levels of agreement* with each of the sensitive and placebo statements, the *responses to the open-ended question* about disrupting the talks of controversial speakers, and whether a participant *supported the anonymous petition* to require mandatory yearly sexual harassment training at UCSB.

The levels of agreement with each of the sensitive and placebo statements are re-oriented in such a way that more positive values always indicate views that participants in the Exploratory Survey perceived to be more socially acceptable at UCSB. We summarize the participants' attitudes towards the sensitive (placebo) statements in an equally-weighted index by taking a simple average of the participants' re-oriented levels of agreement with each of the sensitive (placebo) statements.

The responses to the open-ended questions were read and classified by two independent Amazon Mechanical Turk workers who were blind to treatment status.[31] For some of the analysis, we collapse

[30] Notice the demographic characteristics and the two behavioral measures should not be affected by treatment status. The demographic characteristics and the donation rates to the AAUW are elicited before the randomization stage; the petition is anonymous for all participants and is sent out two weeks after the end of the experiment. Appendix Table A10 shows that, indeed, the three variables are not affected by treatment status.

[31] The mTurk workers were instructed to classify an open-ended answer into one of three categories: a first category if the answer supported the view that it is acceptable to disrupt the talk of a controversial speaker that is invited to

each open-ended answer into a binary variable that takes value one if both mTurk reviewers agreed that the student reported finding it acceptable to disrupt the talk of controversial speakers and zero otherwise.

4.3 Descriptive Statistics

Table 2 lists the number of subjects that participated in the Social Image Experiment. Appendix Table A5 quantifies the representativeness of the experimental sample on observables, by comparing the demographics of participants in the Social Image Experiment to the demographics of the UCSB population. Comparing columns 1 and 2, we see that our sample is relatively more female, white, composed of Juniors or Seniors, and composed of humanities and social science majors.[32] Appendix Table A8 shows that the Private and Public treatments are balanced on observables. Appendix Table A9 shows that attrition after randomization was modest – around 12% – and not differential by treatment.

Figure 1 shows, for each sensitive statement, the extent to which participants in the Exploratory Survey agree about the socially acceptable position at UCSB. Overall, Figure 1 suggests that, as far as the sensitive statements are concerned, there exists a widely shared understanding at UCSB of the opinions that are socially acceptable on campus and the ones that are not.

Figure 2 shows, for each sensitive and placebo statement, the fraction of subjects who answered that agreeing and disagreeing with the statement are about the same in terms of social acceptability. The figure shows that such fraction is much larger for the placebo statements than for the sensitive statements, thus supporting the idea that the sensitive and the placebo statements are qualitatively different.

Since self-reported political ideology plays an important role in the heterogeneous treatment effect analysis, Appendix Figure A1 shows a histogram of self-reported political ideology among participants in the Social Image Experiment. Before being assigned to treatment, participants were asked how they self-identified on a 7-point political ideology spectrum ranging from extremely liberal to extremely conservative.[33] Throughout the paper, unless otherwise specified, we binarize the political ideology variable along a median split based on the participants' answers. Specifically, we group together participants who self-identify as "extremely liberal" or "liberal" and we refer to them as *liberals*; we group everyone else together and we refer to them as *moderates or conservatives*.

Lastly, Appendix Table A7 shows that, for each of the sensitive statements, the end of the agreeing-disagreeing spectrum that, according to the answers of students in the Exploratory Survey,

campus to deliver a lecture, a second category if the answer opposed that view, and a third category if it was hard to determine whether the student supported or opposed the view based on the student's answer. The classifications of the two mTurk workers agree 86% of the times.

[32] Re-weighting the sample for representativeness on the characteristics described in Appendix Table A5 does not meaningfully alter the results of the experiment.

[33] The wording of the question is from the American National Election Study (ANES, 2016).

is generally considered to be more socially acceptable at UCSB always coincides with the end of the spectrum that is closer to the average position of students in the Private Treatment of the Social Image Experiment who self-identify as liberals.

4.4 Pre-Analysis Plan & Empirical Strategy

The pre-analysis plan for the Social Image Experiment specified four main elements of the analysis: the outcome variables and construction thereof, the moderators used when testing for heterogeneous treatment effects, the estimation sample, and the nature of the statistical analysis.[34] The construction of the outcome variables outlined in the pre-analysis plans follows the description in Section 4.2. The pre-analysis plan specified which statements were considered sensitive and which placebo, and the set of statements for which the participants' answers would need to be re-oriented so that higher numbers correspond to views that are perceived to be more socially acceptable at UCSB. Furthermore, the pre-analysis plan specified the indices as the main outcome variables, rather than the answers to the individual statements. The pre-specified moderators used when testing for heterogeneous treatment effects are: gender, age, race/ethnicity, political affiliation, year in school, and whether the subject's major is in the humanities or in the sciences. The pre-analysis plans specified that the estimation sample would exclude the bottom 5% of participants in terms of survey duration and that the sample size would be around 300.[35] The pre-specified regression specification is shown below. Any deviations from the pre-analysis plan will be pointed out along the way.

Most of the empirical analysis in this section relies on a simple regression framework. Let Y_i denote an outcome variable. Let $T_i \in \{0, 1\}$ be an indicator for being assigned to the Public Treatment, $\boldsymbol{X}_i$ a vector of pre-specified controls, and ε_i an idiosyncratic error term. The pre-specified controls include gender, age, race/ethnicity, political affiliation, religiosity, year in school, and whether the subject's major is in the humanities or in the sciences. We estimate the following regression equation:

$$Y_i = \alpha + \beta T_i + \boldsymbol{\gamma} \boldsymbol{X}_i + \varepsilon_i \tag{1}$$

The coefficient of interest is β, which measures the average treatment effect of being assigned to the Public Treatment compared to being assigned to the Private Treatment. We use robust standard errors in all regressions.

[34] The Social Image Experiment was registered in the American Economic Association Registry for randomized control trials under trial number AEARCTR-0005063.

[35] Including the bottom 5% of participants in terms of survey duration adds some noise, but it does not meaningfully affect the results of the experiment.

4.5 Main Results

In order to study whether social image concerns drive a wedge between the sensitive socio-political attitudes that UCSB students report in private and in public, we estimate the average treatment effect of being assigned to the Public Treatment compared to being assigned to the Private Treatment on the attitudes reported by participants in the Social Image Experiment. Let Y_i in Equation 1 denote participant i's level of agreement with one of the statements or one of the indices. When estimating Equation 1, we normalize each outcome variable Y_i so that the standard deviation of answers in the Private Treatment equals one and the mean equals zero.

Figure 3 shows the average treatment effects and 95% confidence intervals from estimates of Equation 1 for each statement and for the indices of the sensitive and placebo statements. Appendix Table A11 presents regression results for the indices. Appendix Table A12 provides regression results for all the sensitive statements in both normalized (standard deviation) units and un-normalized (original) units. The table also provides unadjusted p-values and "sharpened" False Discovery Rate (FDR)-adjusted q-values following the procedure of Benjamini, Krieger, and Yekutieli (2006), as outlined by Anderson (2008).[36]

The pattern of results paints a consistent picture: for all but one of the sensitive statements, the average treatment effect of being assigned to the Public Treatment is positive; i.e., the results suggest that being assigned to the Public Treatment induces students to skew their answers to the sensitive questions in the direction that is generally perceived to be more socially acceptable among their peers. As robustness checks, Appendix Table A14 shows that the results on the index of sensitive attitudes are not driven by any single statement and Appendix Table A15 shows that they are robust to relaxing the assumption that the Likert scale can be interpreted as linear.

As shown in Appendix Table A11, the average treatment effect on the main outcome variable, namely the index of sensitive attitudes, has a standardized effect size of 0.24 and is significant at the 1% level. In original units, the effect size corresponds to approximately half a point on the $0-10$ Likert scale used by participants to report their levels of agreement with each of the statements. To put such magnitude in perspective, it may be useful to benchmark it against the absolute difference, on the same scale, between the average answers of students assigned to the Private Treatment who are classified as liberals and who are classified as moderates/conservatives according to the median split in terms of self-reported political ideology described in Section 4.3. Such absolute difference amounts to approximately 2 points on the 11-point Likert scale. Therefore, the magnitude of the average treatment effect on the index of sensitive attitudes corresponds to

[36]The unadjusted p-values are appropriate for readers with a priori interest in one specific outcome. The FDR-adjusted q-values for the individual outcomes limit the expected proportion of false rejections of null hypotheses across all sensitive statements. The sharpened FDR-adjusted q-values are less conservative than the unadjusted p-values for p-values greater than about 0.15, and more conservative for unadjusted p-values less than that. To alleviate concerns with multiple hypothesis testing, the pre-analysis plan pre-specified the indices as the main outcome variables, rather than the individual statements.

approximately one quarter of the difference between the average private ideological positions of students who are classified as liberals and students who are classified as moderates/conservatives.

The topics that seem to drive the largest wedge between the answers of participants in the Private and Public treatments involve whether the UCSB administration should allow students to wear Halloween costumes that may be perceived as culturally insensitive and the importance of racial microaggressions at UCSB. The other topics with relatively large treatment effects are whether the government should provide reparations for slavery, whether sexual harassment training should be mandatory for everybody who works or studies at UCSB, and whether the UCSB administration should require professors to use trigger warnings in their classes.

The placebo statements do not exhibit a similar pattern of results; in fact, for each of the placebo statements and for the corresponding index, the average level of agreement of students assigned to the Public Treatment and the average level of agreement of students assigned to the Private Treatment do not significantly differ.

One may worry that, due to the nature of the statements, the perceived socially acceptable direction assigned to the placebo statements is close to arbitrary. In that case, the index for the placebo statements may underestimate the effects on the placebo questions by averaging positive and negative effects on the individual questions. As a robustness check, we ignore the directions of the treatment effects entirely and consider the absolute values of the treatment effects on each statement. If the placebo statements engage the students' social image concerns less than the sensitive statements, the magnitude of the treatment effects on the placebo attitudes should be smaller than that on the sensitive attitudes. In line with the hypothesis, a Mann-Whitney U test reveals that the distribution of absolute treatment effects for the sensitive and placebo statements differ significantly (p-value 0.04), with the sensitive statements generally exhibiting larger absolute treatment effects.

Table 3 shows the results on the open-ended question about whether it is acceptable to disrupt the talks of controversial speakers invited to campus to deliver lectures. The table presents the estimates of a logit model where the dependent variable equals one if both mTurk reviewers agreed that the student reported finding it acceptable to disrupt the talk of controversial speakers and zero otherwise. The independent variable is a treatment indicator and, contingent on the specification, controls may be included. As shown in Table 3, the average marginal effect of being assigned to the Public Treatment is around 10 percentage points on a baseline of 30 percent. Therefore, everything else equal, we would expect around a 10 percentage-point increase in the proportion of respondents who report finding it acceptable to disrupt the lecture of a controversial speaker if we changed treatment status from Private to Public.

Appendix Table A13 shows the results of a simple text analysis on the answers to the open-ended question. Specifically, the left part of the table lists the ten trigrams that, according to a naive Bayes classifier, are most diagnostic of being in the Private Treatment compared to the Public

Treatment; conversely, the right part of the table lists the ten trigrams that are most diagnostic of being in the Public Treatment compared to the Private Treatment (Russell and Norvig, 1995). The trigram that is most diagnostic of being in the Private Treatment compared to the Public Treatment is "freedom of speech". In fact, an open response that contains the trigram "freedom of speech" is five times more likely to be from the Private Treatment than from the Public Treatment. Other trigrams that are relatively more likely to appear in the Private Treatment than in the Public Treatment are "able to voice", "right to say", and "speech is important". Conversely, one of the trigrams that is relatively more likely to appear in the Public Treatment than in the Private Treatment is "perfectly acceptable to", which, in the context of the question, likely refers to a student finding it perfectly acceptable to disrupt the talk of a controversial speaker invited to campus to deliver a lecture.[37]

The results of the text analysis dovetail with the results that rely on the classification by the mTurk workers: students assigned to the Public Treatment are more likely to argue that it is acceptable to disrupt the talks of controversial speakers and less likely to discuss issues related to freedom of speech.

Overall, the results of the Social Image Experiment show that, when it comes to sensitive topics related to the political correctness debate on college campuses, social image concerns do indeed drive a wedge between the attitudes that students report in private and the attitudes they report in public. Appendix Figure A2 provides additional support for the hypothesis by showing a positive relationship between the size of the treatment effects and the extent to which participants in the Exploratory Survey perceived each statement as being sensitive.

Figure 4 presents a histogram of the index of sensitive attitudes for participants in the Private and Public Treatments. The figure suggests that most of the misreporting occurs at the bottom of the distribution of the index of sensitive attitudes. Specifically, participants who, in private, would have given answers that are far from the end of the spectrum that is generally perceived to be more socially acceptable at UCSB appear to engage in substantial misreporting. Conversely, participants who, in private, would have given answers that are close to the end of the spectrum that is generally perceived to be more socially acceptable at UCSB appear to engage in little misreporting. Consistent with the treatment effects not being uniform across the distribution of the index of sensitive attitudes, a Kolmogorov-Smirnov test fails to detect a first-order stochastic dominance shift when comparing the distribution of the index in the Public Treatment and in the Private Treatment (p-value of K-S test: 0.31).

The heterogeneous treatment effects analysis in the next section confirms the intuition above by showing that participants who self-identify as moderates or conservative tend to both answer low numbers in the Private Treatment and drive the treatment effects.

[37]The statement of the question is: "If some student group at UCSB invited [name of controversial speaker] on campus to give a talk, would it be acceptable for other students to disrupt the talk and prevent [name of controversial speaker] from delivering [his/her] lecture or would it not be acceptable?"

4.6 Heterogeneous Treatment Effects Analysis

The pre-analysis plan specified the following moderators for the heterogeneous treatment effect analysis: gender, race/ethnicity, political affiliation, year in school, and whether the subjects' major is in the humanities or in the sciences. Figure 5 presents estimates of the interaction term between being assigned to the Public Treatment and each moderator.

As shown in Figure 5, the treatment effects differ widely depending on the participants' self-reported political ideologies. Specifically, subjects who, according to the median split on the political ideology scale, self-identify as moderate or conservative exhibit significantly larger treatment effects than subjects who self-identify as liberal. No other significant dimension of heterogeneity emerges from the heterogeneous treatment effect analysis.

As robustness checks, Appendix Table A16 shows that the heterogeneous treatment effects on the self-reported ideology dimension are robust to extending the definition of liberals to include participants who self-identify as "slightly liberal", and Appendix Table A17 shows that they are robust to relaxing the assumption that the Likert scale can be interpreted as linear. Finally, Appendix D.5.1 presents a robustness check suggesting that the heterogeneous treatment effects on political ideology are not driven by ceiling effects.[38]

The heterogeneous treatment effects on political ideology, beside having important implications for information loss as discussed below, may help build intuition about the overall magnitude of the treatment effects in the experiment and may inform the debate about the perceived lack of ideological diversity on college campuses (Inbar and Lammers, 2012).[39] As mentioned in the descriptive statistics section, for each of the sensitive statements, the end of the agreeing-disagreeing spectrum that participants in the Exploratory Survey considered to be more socially acceptable at UCSB always corresponds to the end of the spectrum that is closer to the average position of self-identified liberals in the Private Treatment of the Social Image Experiment. Therefore, assuming the Likert scale can be interpreted as linear, we can measure the ideological gap between self-reported liberals and self-reported moderates/conservatives by calculating the absolute difference between the average value of the index of sensitive attitudes among students who self-identify as liberal and among students who self-identify as moderate or conservative. Doing so separately for the Private and the Public Treatment shows that the ideological gap in the Private Treatment is around one-and-a-half times larger than the ideological gap in the Public Treatment. Therefore, if one were to naively assess the degree of ideological diversity on college campus by taking public statements at face value, one would underestimate the degree of ideological diversity on campus by a factor of one third.

In order to dig deeper into the heterogeneous treatment effects on self-reported political ideology

[38] One can imagine a world in which the treatment effects on liberals and conservatives are in principle identical, but in which ceiling effects mechanically constrain the treatment effects on liberals and not on conservatives.

[39] See also the work of non-profit advocacy group Heterodox Academy (www.heterodoxacademy.org).

and to begin exploring the implications for information loss, Figure 6 shows the value of the index of sensitive attitudes, averaged across all participants in a certain category of self-reported political ideology, for different categories of self-reported ideology and for both the Private and the Public Treatments. Comparing the averages from the Private and the Public Treatments shows that the treatment effects of being assigned to the Public Treatment become steadily larger the closer one identifies with the conservative end of the political spectrum.

Figure 6 also provides some indication that the statements of participants in the Public Treatment may be less informative than the statements of participants in the Private Treatment. If treatment effects were homogeneous across different categories of self-reported political ideology, one could in principle predict a student's political ideology equally accurately in the Private and the Public Treatments by leveraging the student's index of sensitive attitudes. As shown in Figure 6, however, heterogeneity in treatment effects causes the relationship between the index and the students' self-reported political ideology to flatten in the Public Treatment compared to the Private Treatment. The flatter relationship, together with a lack of change in variances, should make it harder to predict the self-reported political ideology of a participant in the Public Treatment than that of a participant in the Private Treatment.

The following section formalizes and extends the intuition above by showing that the answers of students in the Public Treatment are less informative than those of students in the Private Treatment.

5 Implications of the Distortions for Information Loss

Do the equilibrium distortions caused by social image lead to information loss? In this section, we show that they do and that the magnitude of information loss can be substantial.[40]

In order to study the degree of informativeness of the answers given by participants in the Social Image Experiment, it is useful to adopt a measure of informativeness that satisfies at least two desiderata: first, the measure has to be theoretically grounded in the sense of being related to the theoretical framework from Section 2. Second, the measure has to be cardinal so as to be able to capture the magnitude of information loss.

The canonical information-theoretic notion of *mutual information* satisfies both desiderata.[41,42]

[40] The analysis in this section was not specified in the pre-analysis plan, because, at the time in which the pre-analysis plan was submitted, the connection between Blackwell informativeness and mutual information had not yet been drawn. The entire analysis that was proposed in the pre-analysis plan appears, together with additional results, in Appendix E. The results from this section and the ones from Appendix E paint a highly consistent picture.

[41] Mutual information is defined as the Kullback-Leibler divergence between the joint distribution of two random variables and the product of their marginal distributions. It measures the degree of dependency of two random variables; specifically, it measures the reduction in entropy about a certain random variable that comes from observing the realization of another (Cover and Thomas, 2006)

[42] Other plausible measures of information, such as Shannon Entropy, fail at least one of the desiderata. Shannon Entropy, for instance, does not satisfy the first condition. Specifically, the model suggests that, depending on the initial

First, it follows naturally from the model in Section 2 because it is implied by Blackwell informativeness via the data-processing inequality (Cover and Thomas, 2006). Therefore, if social image concerns induce distortions that diminish the informativeness of equilibrium statements according to the Blackwell partial order, they also diminish the equilibrium statements according to the mutual information order. Second, mutual information has cardinal value.

Since the students' private issue-positions - denoted by π in the model - are not directly observable, we study the mutual information between the students' statements x and random variables τ that we would expect to be correlated with the students' private issue-positions.[43] Specifically, we consider the relationship between a participant's index of sensitive attitudes and whether the participant self-identified as a liberal or as a moderate/conservative. Furthermore, we consider the relationship between a participant's level of agreement with the statement about sexual harassment training - "Sexual harassment training should be mandatory for everybody who works or studies at UCSB" - with the amount the participant chose to donate to the American Association of University Women (AAUW), and with whether the participant chose to support the anonymous petition demanding that UCSB mandate yearly sexual harassment training.

As a preliminary step, Appendix Table A10 shows that the target random variables - self-identifying as liberal or as moderate/conservative, the amount donated to the AAUW, and the fraction of participants supporting the petition requesting that UCSB mandate yearly sexual harassment training - are not affected by treatment status. Therefore, information loss about the target random variables is unlikely to be due to those variables varying with treatment, but rather to the levels of agreement with the sensitive statements varying with treatment.

Figure 7 shows the mutual information between: i) a participant's index of sensitive attitudes and whether the participant self-identified as liberal or as moderate/conservative, ii) a participant's reported level of agreement with the statement about sexual harassment and the quartile of the participant's donation to the AAUW, and, iii) a participant's reported level of agreement with the statement about sexual harassment and whether the participant supported the anonymous petition requesting that UCSB mandate yearly sexual harassment training. The results show that, across the board, mutual information is larger in the Private Treatment than in the Public Treatment.

distribution of types, the distortions generated by social image may reduce informativeness of equilibrium statements according to the Blackwell partial order and, at the same time, increase the Shannon Entropy of the responses. Only when social image concerns are sufficiently strong, does the model imply that the distortions decrease the entropy of equilibrium statements. Empirically, it is the case that the entropy of the answers of participants in the Private Treatment is higher than the entropy of the answers of participants in the Public Treatment, as shown in Appendix Figure A5.

[43] The students' private issue-positions π are unobservable by definition: in both the Private and the Public Treatments, we can only observe the students' reports x about π. One could in principle impose the assumption that the answers of students assigned to the Private Treatment reflect the students' private-issue positions and, in fact, we do make such assumption in Appendix E.4. Most of the analysis, however, including the one in this section, is designed to be robust to violations of that assumption. In other words, most of the analysis is designed to accommodate the possibility that students assigned to the Private Treatment may not truthfully report their private issue-positions, which would happen for instance if they worried their answers would not be kept fully confidential.

According to the mutual information measure, the statements of participants in the Public Treatment are 25-to-50% less informative about the target random variables than the statements of participants in the Private Treatment.

As a robustness check, Appendix Figure A8 shows that the mutual information between the index of placebo statements and self-identifying as liberal or as moderate/conservative is not significantly different in the Private and the Public Treatments. Furthermore, Appendix Table A19 shows that, for virtually all the demographic characteristics elicited in the pre-screen survey, the mutual information between the index of sensitive attitudes and the demographic characteristic is at least as high in the Private Treatment as in the Public Treatment.

Appendix E shows that the results are unaltered if, rather than mutual information, one employs alternative measures of informativeness, ranging from Shannon Entropy to measures of classifier performance often used in the machine learning literature. The results about classifier performance have a particularly intuitive interpretation; specifically, the results show that, if one were to try to predict the students' demographic characteristics and incentivized behaviors from the students' answers to the sensitive statements, one would make more accurate predictions when using the answers of students assigned to the Private Treatment than when using the answers of students assigned to the Public Treatment.

Overall, social image concerns around topics related to political correctness distort the sensitive socio-political statements made by students assigned to the Public Treatment compared to the ones made by students assigned to the Private Treatment, and the nature of the distortions is such as to cause the statements made in the Public Treatment to be less informative than the ones made in the Private Treatment.

The model from Section 2 identified two distinct mechanisms whereby the distortions caused by social image may lead to information loss: pooling and scrambling. Appendix E.4 shows that most of the information loss in the experiment is due to scrambling rather than pooling. Furthermore, as shown in Appendix Figure A7, the distribution of answers from the Public Treatment can be closely mimicked by taking the distribution of answers from the Private Treatment and adding noise to them in a manner consistent with the theoretical framework. In other words, the answers of participants in the Public Treatment are close to being a *garbling* of the answers of participants in the Private Treatment (Blackwell, 1951, 1953).

6 Information Loss: Actual Audience

The previous section can be thought of as studying information loss from the perspective of a sophisticated audience who recognize the ways in which social image concerns distort the attitudes that students display in public and who adjust their inferences accordingly. If the natural audience in the environment, namely other UCSB students, are less sophisticated in understanding the ways

in which social image distorts the attitudes portrayed in public by their peers, they may be able to extract even less information from public statements. In order to study whether audience naivete leads to additional information loss, we employ a second experiment.

6.1 Experimental Design

In early May 2020, we recruited a new set of subjects through the Experimental and Behavioral Economics Laboratory (EBEL) portal at the University of California Santa Barbara (UCSB) to complete a short online survey.[44,45] The recruitment email did not mention the topic of the study. Before being directed to the consent form, subjects were asked to complete a brief pre-screen questionnaire similar to the one administered to subjects in the Social Image Experiment.[46] A subject was considered eligible to participate in the experiment if she reported being an undergraduate student at UCSB and had participated in neither the Exploratory Survey nor the Social Image Experiment.

After completing the pre-screen survey and consenting to take part in the experiment, subjects were informed that their task would be to forecast the answers of a group of UCSB undergraduate students who participated in an online survey experiment run a few months earlier. Subjects were furthermore informed that, on top of the $7 baseline payment for completing the experiment, each of them would have a chance to earn a $10 bonus payment with a probability increasing in the accuracy of their forecasts.

The details of the lottery system used to allocate the bonus payments to participants are as follows: at the end of the survey, each participant was entered into a lottery that paid either $10 or $0. The lotteries were independent across participants. For each participant, the probability with which the lottery paid $10 was a function of the average accuracy of the participant's forecasts across all the questions in the experiment. Specifically, the more accurate the participant's forecasts, the higher the probability of the $10 outcome. Such payment system has the property of being incentive compatible irrespective of the participant's degree of risk aversion.[47] Participants were informed of the details of the payment system, including the mathematical formulas used to generate the probability of being awarded the $10 bonus payments, and were told that it was in their best interest to give their most accurate guess to each question.

[44]The text of the survey in the Forecasting Experiment is included in the submission documents. It is also available from https://sites.google.com/view/lucabraghieri/research.

[45]Data collection for the Forecasting Experiment was completed before the death of George Floyd in late May 2020 and the ensuing protests.

[46]The only difference between the pre-screen survey from the Social Image Experiment and the one from the Forecasting Experiment is that the latter also included a question asking students to report their verbal and math SAT scores.

[47]The probability distribution over payments generated by the procedure above is equivalent to the one generated by a two-step procedure in which, first, a question is randomly selected for each participant, and, second, each participant is awarded a $10 bonus with a probability that depends on the accuracy of her answer to the selected question.

Subjects were then informed that the previous study employed two treatments and were shown both the instructions received by participants assigned to the Private Treatment and the instructions received by participants assigned to the Public Treatment. Furthermore, subjects were informed that they would be shown the distribution of answers of participants in the Public Treatment and be asked to forecast the average answers of participants in the Private Treatment.

In order to help subjects familiarize themselves with the forecasting task, the subsequent screen displayed, for two of the questions asked in the Social Image Experiment, the distribution of answers of both participants in the Private Treatment and of participants in the Public Treatment. Unbeknownst to subjects, one of the two questions involved a sensitive statement and one involved a placebo statement.

Next, subjects were shown, for each of the remaining 13 statements, the distribution of answers of participants in the Public Treatment and were asked to forecast the average answer of participants in the Private Treatment. Subjects reported their forecasts on a 0-10 slider that allowed them to select numbers with up to one decimal place. The order of the statements was randomized at the subject level.

Upon completing the first round of forecasts, subjects were informed that, for 10 of the statements, the experimenter had hypothesized that the average answers of participants in the Private and the Public Treatments would differ systematically. After being shown those 10 statements - the sensitive statements -, subjects were asked to forecast whether the effects of being assigned to the Public Treatment compared to the Private Treatment were, on average across the 10 statements, larger for participants belonging to a certain demographic group than for participants belonging to the complementary demographic group. The questions were multiple-choice: subjects could answer that the treatment effects were significantly larger for the first demographic group, significantly larger for the second demographic group, or that the treatment effects were not significantly different from one another. Subjects were informed that, in the context of the Social Image Experiment, a statistically significant difference corresponded to approximately 0.7 points on the 0-10 scale used by participants to report their answers. The demographic groups were the same as the ones pre-specified for the heterogeneity analysis in the Social Image Experiment, namely gender, race/ethnicity, political affiliation, year in school, and whether the student's major was in the humanities or in the sciences.[48] The order of the questions was randomized at the subject level.

After answering the questions about heterogeneity, subjects were asked how confident they were in the last round of forecasts. Specifically, they were asked to estimate, out of the 5 multiple-choice questions about treatment-effect heterogeneity they had just answered, how many of their responses

[48] When the composition of a certain demographic group was not self-explanatory, subjects were provided with additional information. For instance, subjects were informed that, among participants in the Social Image Experiment who did not identify as white, 5% identified as Black or African-American, 55% identified as Asian or Asian-American, 25% identified as Latino or Chicano, and 15% identified as some other race/ethnicity (e.g. Native American). Similarly, subjects were informed of the details of the classification into liberals and moderates/conservatives.

were correct. An answer was considered correct if it matched the results from the heterogeneous treatment effect analysis of the Social Image Experiment.

Next, subjects were asked a set of questions about the behaviors in the Social Image Experiment of participants who were classified as liberals and of participants who were classified as moderates/conservatives. Specifically, subjects were told to consider two of the sensitive statements, namely the ones about reparations for slavery and about illegal immigration.[49] For the statement about reparations for slavery, subjects were informed of the percentage of participants in the Private Treatment of the Social Image Experiment with a certain political affiliation - liberal or moderate/conservative - who reported agreeing with the statement, and were asked to forecast the corresponding percentage for participants in the Public Treatment. For the statement about illegal immigration, subjects engaged in a similar exercise, but rather than being asked to forecast the percentage of participants who agreed with the statement, they were asked to forecast the percentage of participants who disagreed with the statement. The difference is due to the fact that the direction of social desirability differed across the two statements.

The last few screens of the experiment were as follows. First, after being informed that their answers would be kept confidential, subjects were asked to state their own levels of agreement with each of the sensitive and placebo statements. The order of the questions was randomized at the subject level. Second, subjects were shown the aggregate levels of agreement with each of the sensitive and placebo statements of participants in the Private Treatment of the Social Image Experiment. Third, since the survey was administered in early May 2020, during the first wave of the COVID-19 pandemic, subjects were asked whether their answers would have been systematically different had the survey been run before the outbreak of the pandemic. Finally, upon completing the survey, subjects received a $7 Amazon gift-card for their participation and were redirected to a spreadsheet that contained the anonymized individual-level answers of participants in the Public Treatment of the Social Image Experiment.

Bonus payments were calculated and disbursed a few days after the end of the survey.

6.1.1 Discussion of the Experimental Design

Before proceeding, it is worth discussing two aspects of the experimental design. First, by showing subjects the instructions of both participants in the Private and the Public Treatments of the Social Image Experiment, the design of the Forecasting Experiment likely made social image more salient in the subjects' minds than it would be in a natural environment. Second, by showing subjects the full distribution of answers of both participants in the Private and the Public Treatments of the

[49] This section of the survey features only two of the sensitive statements, rather than all the sensitive statements, because of concerns about survey fatigue. The reason behind selecting the statements about reparations for slavery and about illegal immigration is that the students' levels of agreement with those statements, both in the Private and in the Public Treatments of the Social Image Experiment, are among the most highly correlated with self-reported political affiliation.

Social Image Experiment to two of the statements, the design may have helped subjects benchmark the magnitudes of their forecasts.

Both design choices aimed to mitigate the amount of information loss driven by the unfamiliar nature of the experimental environment. In other words, the design of the Forecasting Experiment aimed to ensure that any loss of information can be attributed to the students' incorrect beliefs about the ways in which social image distorts their peers' public attitudes rather than to the artificiality of the experimental environment.

6.2 Outcome Variables

The first set of outcomes is related to a subject's forecasts of the average answers of participants in the Private Treatment of the Social Image Experiment. Since, in the Social Image Experiment, the *direction of social acceptability* varied depending on the statement, the average answer of participants in the Private Treatment was sometimes larger than the average answer of participants in the Public Treatment and sometimes smaller. The first outcome variable of interest in the Forecasting Experiment is, for each statement, an indicator that takes value one whenever the actual average answer of participants in the Private Treatment and a subject's forecast are either both larger than the actual average answer of participants in the Public Treatment or both smaller. In other words, the indicator variable takes value one if and only if, for a sensitive statement, the subject's forecast reflects a correct understanding of the direction in which the answers of participants in the Public Treatment are on average skewed compared to the answers of participants in the Private Treatment.

The second outcome variable of interest is, for each statement, the absolute difference between the average answer of participants in the Public Treatment of the Social Image Experiment and a subject's forecast of the average answer of participants in the Private Treatment. Since some of the statements in the Social Image Experiment were considered sensitive and some placebo, comparing such absolute differences across the sensitive and placebo statements is informative of whether subjects in the Forecasting Experiment understand that some of the statements are sensitive, some placebo, and that they should treat the two differently.

The third outcome variable of interest is a measure of forecast accuracy that, for each statement, consists in taking the absolute value of the difference between a subject's forecast of the average answer of participants in the Private Treatment and the actual average answer.

In order to generate a measure of confidence in the answers to the questions about heterogeneity in treatment effects, we subtract the actual number of correct answers to the multiple-choice questions from the forecasted number of correct answers. The measure of confidence thus constructed is such that positive values indicate the degree of overconfidence, negative values indicate the degree of underconfidence, and a null value indicates calibrated beliefs.

Lastly, we consider the subjects' forecasts of the answers of self-identified liberals and self-

identified moderates/conservatives to the questions about reparations for slavery and illegal immigration. Each subject's answers, together with information about the fraction of participants in the Social Image Experiment who self-identify as liberals or as moderates/conservatives, allow us to obtain an individual-level measure of the perceived degree of mutual information between agreeing with each of the statements and self-identifying as liberal and a measure of the perceived diagnosticity of agreeing with each of the statements for self-identifying as liberal.

6.3 Results

6.3.1 Forecasts of the Average Answers of Participants in the Private Treatment

When shown the distribution of answers of participants in the Public Treatment of the Social Image Experiment and asked to forecast the average answers of participants in the Private Treatment, subjects appear to exhibit some degree of sophistication.[50] Figure 8 shows, for each statement, the median value of the absolute difference between the average answer of participants in the Public Treatment of the Social Image Experiment and the subjects' forecasts. Virtually across the board, the median value of the absolute difference is larger for the sensitive statements than for the placebo statements. The pattern of results suggests that subjects understand that the wedge between private and public reports may differ systematically across statements and correctly identify the statements where the wedge is relatively large (sensitive statements) and the statements where the wedge is relatively small (placebo statements).

Figure 9 shows that, for each of the sensitive statements, the forecasts of the majority of subjects reflect a correct understanding of the direction in which public statements are skewed compared to private statements. On average across all the sensitive statements, the fraction of subjects who correctly forecasted the direction in which public statements are skewed compared to private statements is 74%.

Finally, Figure 10 presents two scatter plots: the left panel plots, for all sensitive statements other than the one that was shown to participants in the instructions of the Forecasting Experiment, the average level of agreement of participants in the Private Treatment of the Social Image Experiment against the average level of agreement of participants in the Public Treatment of the Social Image Experiment. The right panel plots, for the same set of sensitive statements, the average level of agreement of participants in the Private Treatment of the Social Image Experiment against the median forecast of subjects in the Forecasting Experiment. In both panels, the answers are oriented in such a way that higher numbers always correspond to views that are perceived to be more socially acceptable at UCSB. The left panel of Figure 10 reiterates the results of the Social Image Experiment: it shows that, for virtually all the sensitive statements, the average level of

[50] The results of the Forecasting Experiment are virtually identical if we exclude from the analysis the few students who reported that their answers would have been systematically different had the survey been run before the outbreak of the COVID-19 pandemic.

agreement of participants in the Public Treatment is larger than the average level of agreement of participants in the Private Treatment. The right panel of Figure 10 is informative of the quality of the students' forecasts. If, despite the salient instructions, students remained entirely naive about the fact that social image distorts the attitudes their peers report in public, the left panel and the right panel would look identical. If the students were perfectly sophisticated about the ways in which social image distorts their peers' public attitudes, all the dots in the right panel would fall on the 45 degree line. The line of best fit in the right panel of Figure 10 is very close to the 45 degree line, thus suggesting that, modulo some noise at the level of the individual statements, the students are fairly accurate in forecasting the average treatment effects among their peers.

Despite being quite accurate at forecasting the average treatment effects on the sensitive statements, students are even more accurate at forecasting the average treatment effects on the placebo statements. As shown in Appendix Figure A9, the average absolute error on the sensitive attitudes is around 20% larger than the average absolute error on the placebo attitudes.

Who are the students that make the most accurate forecasts? For each of the three dimensions of accuracy described above, namely deflating the sensitive statements more than the placebo statements, deflating the sensitive statements in the correct direction, and deflating the sensitive statements by the correct amount, students with higher verbal SAT scores consistently outperform their peers. Specifically, for each of the three dimensions of accuracy, a LASSO regression always selects verbal SAT score as the first variable entering the model among the following variables: gender, age, race/ethnicity, political affiliation, religiosity, year in school, whether the subject's major is in the humanities or in the sciences, verbal SAT score, math SAT score, and the subject's own index of sensitive attitudes.[51]

6.3.2 Forecasts of Treatment Effect Heterogeneity

In day-to-day social interactions, students often have access to a rich set of cues that they can use to condition their inferences. For instance, they might know the gender of a speaker and will make inferences in light of that knowledge. As a consequence, it is important to study the students' beliefs about the extent to which different demographic groups skew their public statements. In this section, we show that students exhibit a much smaller degree of sophistication when forecasting heterogeneity than when forecasting average effects. In fact, the majority of students confidently holds incorrect beliefs about the extent to which different demographic groups skew their public statements. As shown in the next section, such incorrect beliefs can lead students to draw vastly erroneous inferences about the informativeness of public statements.

Figure 11 presents histograms of the subjects' answers to a set of questions asking whether the effects of being assigned to the Public Treatment compared to the Private Treatment are, on

[51] The result also holds for accuracy with respect to the multiple-choice questions about treatment-effect heterogeneity discussed in the next section.

average across the 10 sensitive statements, significantly larger for subjects belonging to a certain demographic group, significantly larger for subjects belonging to the complementary demographic group, or not significantly different across the two demographic groups. As mentioned, the instructions of the heterogeneity section gave subjects a precise numerical definition of when the treatment effect on one demographic group was considered significantly larger than the treatment effect on the complementary demographic group; specifically, a treatment effect was considered significantly larger than another if the absolute difference between the two was weakly larger than 0.7 points on the 0-10 scale used by participants in the Social Image Experiment to report their answers. The dimensions of heterogeneity explored in the Forecasting Experiment are the same as the ones discussed in the analysis of the Social Image Experiment, namely gender, race/ethnicity, political affiliation, year in school, and whether the student's major is in the humanities or in the sciences.

As shown in Figure 11, the subjects' beliefs about treatment effect heterogeneity are not very accurate. For three of the five questions, a plurality of subjects selected the answer that matches the results from the heterogeneous treatment effect analysis of the Social Image Experiment; however, the plurality margin is small and, in two cases out of three, not statistically significant. Importantly, a majority of participants in the Forecasting Experiment does not identify the key dimension of heterogeneity that emerges from the heterogeneous treatment effect analysis, namely self-reported political ideology. A plurality of participants (46%) does report believing that self-identified liberals exhibit smaller treatment effects than self-identified moderates/conservatives. However, a large fraction of participants (36%) thinks the opposite is true, and the two proportions are not significantly different from one another.

For two of the five questions, namely the ones related to race and gender, a plurality of subjects did not select the answer that matches the results from the heterogeneous treatment effect analysis of the Social Image Experiment. Specifically, a large majority of subjects (70%) reported believing that the difference between the average answers of participants in the Private and Public Treatments of the Social Image Experiment is significantly larger for participants who self-identified as white than for participants who self-identified as non-white. Similarly, a plurality of subjects (43%) reported believing that the difference between the average answers of participants in the Private and Public Treatments of the Social Image Experiment is significantly larger for participants who self-identified as male than for participants who self-identified as female.

Since the heterogeneous treatment effects in the Social Image Experiment are estimated with a margin of error, it is possible that, whenever there is a discrepancy between the answer given by the plurality of participants in the Forecasting Experiment and the answer suggested by the heterogeneous treatment effect analysis in the Social Image Experiment, the former, rather than the latter, is correct. Appendix F.1 shows such eventuality is unlikely.

Therefore, overall, students in the Forecasting Experiment seem to be partially naive about the dimensions of heterogeneity that drive the average treatment effects; specifically, they think whites

and males exhibit larger treatment effects than their counterparts and they fail to identify political ideology as the key dimension of heterogeneity driving the average treatment effects.

Is the extent of inaccuracy in forecasting heterogeneity in treatment effects reflected in the degree of confidence exhibited by the students? Figure 12 shows the fraction of subjects who are overconfident about the accuracy of their answers to the heterogeneity questions, the fraction of subjects who are underconfident about the accuracy of their answers, and the fraction of subjects whose beliefs about the accuracy of their answers are well-calibrated. More than 60% of subjects are overconfident, in the sense of their estimate of the number of multiple-choice questions about heterogeneity they answered correctly is strictly larger than the actual number of answers that match the results of the heterogeneous treatment effect analysis of the Social Image Experiment. Only about 20% of students are underconfident.

6.3.3 Inferences from Public Statements

The subjects' incomplete understanding of the dimensions of heterogeneity driving the treatment effects in the Social Image Experiment may lead them to draw erroneous inferences from their peers' public statements. In particular, the fact that subjects in the Forecasting Experiment do not seem to recognize self-reported political ideology as a key dimension of treatment-effect heterogeneity may induce them to overestimate the extent to which students' public statements are informative of their political leanings. This section shows that, indeed, subjects substantially overestimate the degree to which the answers of participants in the Public Treatment of the Social Image Experiment are informative of the participants' self-reported political affiliations.

Before proceeding, it is useful to introduce some notation. Let x_1 be an indicator for agreeing with the statement about reparations for slavery and x_2 be an indicator for disagreeing with the statement about illegal immigration.[52] Let τ be an indicator for being classified as a liberal in the Social Image Experiment. Participants in the Forecasting Experiment were asked to estimate, for participants in the Public Treatment of the Social Image Experiment, $P(x_1 = 1|\tau = 1)$, $P(x_1 = 1|\tau = 0)$, $P(x_2 = 1|\tau = 1)$, and $P(x_2 = 1|\tau = 0)$. In order to help them benchmark their forecasts, participants were given the corresponding proportions from the Private Treatment of the Social Image Experiment.

Table 4 shows the participants' estimates of $P(x_1 = 1|\tau = 1)$, $P(x_1 = 1|\tau = 0)$, $P(x_2 = 1|\tau = 1)$, and $P(x_2 = 1|\tau = 0)$, together with the actual fractions from the Public Treatment of the Social Image Experiment. On average, participants tend to overestimate the fraction of self-identified liberals who, when assigned to the Public Treatment, reported agreeing with the statement about reparations for slavery and to underestimate the fraction of self-identified moderates/conservatives

[52] Subjects in the Forecasting Experiment were informed that, in the Social Image Experiment, agreeing with a statement was defined as answering a number strictly greater than 5 on the 0-10 Likert scale and disagreeing with a statement was defined as answering a number strictly smaller than 5.

who, when assigned to the Public Treatment, reported agreeing with the statement. A similar pattern of results obtains when participants estimate the fractions of self-identified liberals and moderates/conservatives who, when assigned to the Public Treatment, report disagreeing with the statement about illegal immigration. The results are in line with the finding from the previous section that only a minority of participants thinks that treatment effects are larger for participants who are classified as moderates/conservatives than for participants who are classified as liberals.

As a result of such misperception, participants in the Forecasting Experiment think that agreeing with the statement about reparations for slavery and disagreeing with the statement about illegal immigration in the Public Treatment of the Social Image Experiment are more diagnostic of self-reported political ideology than they actually are. Figure 13 shows the distribution of Bayesian posteriors implied by the beliefs reported by subjects in the Forecasting Experiment. Overall, virtually all participants in the Forecasting Experiment believe that publicly agreeing with the statement about reparations for slavery or disagreeing with the statement about illegal immigration are more diagnostic of self-identifying as a liberal than they actually are.

Overall, the incorrect beliefs of participants in the Forecasting Experiment about the dimensions of heterogeneity that drive the average treatment effects lead them to draw erroneous inferences about the extent to which their peers' public statements are informative of their self-reported political ideology. As such, students in the Forecasting Experiment exhibit some naivete that leads to an additional degree of information loss.

7 Conclusion

An important argument in the cultural debate about political correctness seems to be that people may feel pressure to publicly espouse views on a set of sensitive socio-political topics that they may not privately hold, and that such misrepresentations may impoverish the quality of public discourse. This paper provided a formalization of the argument based on a signaling model with lying costs and tested it in an environment that is often at the center of the debate, namely a college campus. Broadly speaking, the results lend empirical support to the argument: students indeed skew their public utterances in the direction that is perceived to be more socially acceptable on campus, and such distortions cause public statements to be less informative than private statements according to a host of measures of informativeness suggested by the theoretical model. Furthermore, the students are in part naive about the ways in which social image distorts their peers' public attitudes and such naivete leads to an additional degree of information loss.

As discussed, this paper can be seen as a first step in a broader welfare analysis of the phenomenon of political correctness. Specifically, in environments where information transmission is desirable, the paper suggests that social image concerns around topics related to political correctness may exact a toll in terms of information transmission. Clearly, a full welfare analysis requires

a proper understanding of the benefit side of political correctness as well. Developing an empirically grounded understanding of the benefits of political correctness is an interesting and important avenue for future research.

References

Elliot Ackerman and Other Signatories. A Letter on Justice and Open Debate. *Harper's Magazine*, 2020.

Chunrong Ai and Edward C. Norton. Interaction Terms in Logit and Probit Models. *Economic Letters*, 80, 2003.

S. Nageeb Ali and Roland Bénabou. Image Versus Information: Changing Societal Norms and Optimal Privacy. *American Economic Journal: Microeconomics*, 12(3), 2020.

Francisco Alpizar, Fredrik Carlsson, and Olof Johansson-Stenman. Anonymity, Reciprocity, and Conformity: Evidence from Voluntary Contributions to a National Park in Costa Rica. *Journal of Public Economics*, 92(5-6):1047–1060, 2008.

American National Election Study. ANES. Technical report, 2016.

Michael L. Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495, 2008.

James Andreoni and B. Douglas Bernheim. Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica*, 77(5):1607–1636, 2009.

James Andreoni and Ragan Petrie. Public Goods Experiments Without Confidentiality: A Glimpse into Fundraising. *Journal of Public Economics*, 88(7-8):1605–1623, 2004.

Ian Ball. Scoring Strategic Agents. *Working Paper*, 2020.

Abhijit Banerjee, Emily Breza, Arun G. Chandrasekhar, and Benjamin Golub. When Less is More: Experimental Evidence on Information Delivery During India's Demonetization. *Working Paper*, 2020.

Jeffrey S. Banks and Joel Sobeli. Equilibrium Selection in Signaling Games. *Econometrica*, 55(3), 1987.

Kathleen Barker. To Be PC or Not To Be? A Social Psychological Inquiry into Political Correctness. *Journal of Social Behavior and Personality*, 9(2), 1994.

Lori Beaman and Andrew Dillon. Diffusion of agricultural information within social networks: Evidence on gender inequalities from Mali. *Journal of Development Economics*, 133:147–161, 2018.

Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

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

David Blackwell. Comparison of Experiments. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102, 1951.

David Blackwell. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

David Blackwell and Meyer A. Girshick. *Theory of Games and Statistical Decisions*. John Wiley & Sons, 1954.

Leonardo Bursztyn, Thomas Fujiwara, and Amanda Pallais. "Acting Wife": Marriage Market Incentives and Labor Market Investments. *American Economic Review*, 107(11):3288–3319, 2017.

Leonardo Bursztyn, Bruno Ferman, Stefano Fiorin, Martin Kanz, and Gautam Rao. Status Goods: Experimental evidence from platinum credit cards. *Quarterly Journal of Economics*, 133(3): 1561–1595, 2018.

Leonardo Bursztyn, Georgy Egorov, and Stefano Fiorin. From Extreme to Mainstream: The Erosion of Social Norms. *American Economic Review (forthcoming)*, 2020a.

Leonardo Bursztyn, Alessandra González, and David Yanagizawa-Drott. Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia. *American Economic Review*, 110(10), 2020b.

Arun G. Chandrasekhar, Benjamin Golub, and He Yang. Signaling, Shame, and Silence in Social Learning. *Working Paper*, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

Zoë Cullen and Ricardo Perez-truglia. The Salary Taboo: Privacy Norms and the Diffusion of Information. *Working Paper*, 2020.

Stefano DellaVigna and Devin Pope. What Motivates Effort? Evidence and Expert Forecasts. *Review of Economic Studies*, 85:1029–1069, 2018a.

Stefano DellaVigna and Devin Pope. Predicting Experimental Results: Who Knows What? *Journal of Political Economy*, 126:2410–2456, 2018b.

Stefano DellaVigna, John A. List, and Ulrike Malmendier. Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, 127(1):1–56, 2012.

Stefano DellaVigna, John A. List, Ulrike Malmendier, and Gautam Rao. Voting to Tell Others. *Review of Economic Studies*, 84(1):143–181, 2017.

Allen L. Edwards. *The Social Desirability Varible in Personality Assessment and Research*. Dryden Press, 1957.

Ruben Enikolopov, Alexey Makarin, Maria Petrova, and Leonid Polishchuk. Social Image, Networks, and Protest Participation. *Working Paper*, 2018.

Sebastian Fehrler and Niall Hughes. How Transparency Kills Information Aggregation: Theory and Experiment. *American Economic Journal: Microeconomics*, 10(1):181–209, 2018.

Ralph E. Folsom, Bernard G. Greenberg, Daniel G. Horvitz, and James R. Abernathy. The Two Alternate Questions Randomized Response Model for Human Surveys. *Journal of the American Statistical Association*, 68(343):525–530, 1973.

Alex Frankel and Navin Kartik. Muddled Information. *Journal of Political Economy*, 127(4), 2019.

Jana Friedrichsen, Tobias König, and Renke Schmacker. Social Image Concerns and Welfare Take-Up. *Journal of Public Economics*, 168:174–192, 2018.

Patricia Funk. Social Incentives and Voter Turnout: Evidence from the Swiss Mail Ballot System. *Journal of the European Economic Association*, 8(5):1077–1103, 2010.

Alan S. Gerber, Donald P. Green, and Chirstopher W. Larimer. Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. *American Political Science Review*, 102(1):33–48, 2008.

Ronen Gradwohl. Voting in the Limelight. *Economic Theory*, 66(1):65–103, 2018.

Ronen Gradwohl and Timothy J. Feddersen. Persuasion and Transparency. *The Journal of Politics*, 80(3):903–915, 2018.

Bernard G. Greenberg, Roy R. Kuebler, James R. Abernathy, and Daniel G. Horvitz. Application of the Randomized Response Technique in Obtaining Quantitative Data. *Journal of the American Statistical Association*, 66(334):243–250, 1971.

Yoel Inbar and Joris Lammers. Political Diversity in Social and Personality Psychology. *Perspectives on Psychological Science*, 7(5), 2012.

Anne Karing. Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone. *Working Paper*, 2018.

Navin Kartik. Strategic communication with lying costs. *Review of Economic Studies*, 76(4): 1359–1395, 2009.

Navin Kartik, Marco Ottaviani, and Francesco Squintani. Credulity, Lies, and Costly Talk. *Journal of Economic Theory*, 134, 2007.

Ivar Krumpal. Determinants of Social Desirability Bias in Sensitive Surveys: a Literature Review. *Quality and Quantity*, 47, 2013.

Timur Kuran. *Private Truths, Public Lies*. Harvard University Press, 1995.

John Lea. *Political Correctness and Higher Education*. Routledge, 2009.

Glenn C. Loury. Self-Censorship in Public Discourse. *Rationality and Society*, 6(4):428–461, 1994.

T. Clay McManus and Justin M. Rao. Signaling smarts? Revealed preferences for self and social perceptions of intelligence. *Journal of Economic Behavior and Organization*, 110:106–118, 2015.

Debrah Meloso, Salvatore Nunnari, and Marco Ottaviani. Looking into the Crystal Balls: A Laboratory Experiment on Reputational Cheap Talk. *Working Paper*, 2018.

Markus Mobius and Tanya Rosenblat. Social Learning in Economics. *Annual Review of Economics*, 6:827–847, 2014.

Markus Mobius, Tuan Phan, and Adam Szeidl. Treasure Hunt: Social Learning in the Field. *Working Paper*, 2015.

Felipe Montano-Campos and Ricardo Perez-Truglia. Giving to Charity to Signal Smarts: Evidence from a Lab Experiment. *Journal of Behavioral and Experimental Economics*, 78:193–199, 2019.

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

Alvaro Name-Correa and Huseyin Yildirim. Social Pressure, Transparency, and Voting in Committees. *Working Paper*, 2017.

Niche. Most Liberal Colleges in America, 2020. URL `https://www.niche.com/colleges/search/most-liberal-colleges/`.

Marco Ottaviani and Peter Norman Sørensen. Reputational Cheap Talk. *RAND Journal of Economics*, 37(1):155–175, 2006.

Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15, 2003.

D. Raghavarao and W. T. Federer. Block Total Response as an Alternative to the Randomized Response Method in Surveys. *Journal of the Royal Statistical Society*, 41(1):40–45, 1979.

Johannes Rauh, Pradeep Kr. Banerjee, Eckehard Olbrich, Jurgen Jost, Nils Bertschinger, and David Wolpert. Coarse-Graining and the Blackwell Order. *Entropy*, (19), 2017.

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

Sander Renes and Bauke Visser. Committees of Experts in the Lab. *Working Paper*, 2018.

Michael S. Roth. *Safe Enough Spaces: A Pragmatist's Approach to Inclusion, Free Speech, and Political Correctness on College Campuses*. Yale University Press, 2019.

Suart J. Russell and Peter Norvig. *Artificial Intelligence A Modern Approach*. Pearson, 1995.

Gonzalo Schwarz. Interview with James Heckman. *Medium*, 2020. URL `https://medium.com/archbridge-notes/nobel-prize-winning-economist-dr-5550da1df5c3`.

Paul M. Sniderman and Thomas Piazza. *The Scar of Race*. Harvard University Press, 1993.

Seymour Sudman and Norman M. Bradburn. *Response Effects in Surveys*. Aldine Publishing Company, 1974.

Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. *The Psychology of Survey Response*. Cambridge University Press, 2000.

University of California Santa Barbara. UCSB 2018-2019 Campus Profile, 2019.

Stanley Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Stanley Warner. The Linear Randomized Response Model. *Journal of the American Statistical Association*, 66(336):884–888, 1971.

Robert J Zimmer. Free Speech Is the Basis of a True Education. *Wall Street Journal*, 2016.

Table 1: **List of Sensitive and Placebo Statements**

	Sensitive Statements
1	All statues and memorials of Confederate leaders should be removed.
2	Adopting elements of other cultures, whether more or less dominant, is a perfectly acceptable practice.*
3	The UCSB administration should require professors to address students according to the students' preferred gender pronouns.
4	The Islamic religion is more likely than other religions to encourage violence among its believers.*
5	The UCSB administration should require professors to use trigger warnings in their classes.
6	Sexual harassment training should be mandatory for everybody who works or studies at UCSB.
7	People who immigrated to the U.S. illegally, when caught, should be deported and sent back to their countries of origin.*
8	The U.S. government should provide reparations for slavery.
9	Racial microaggressions are an important problem at UCSB.
10	The UCSB administration should allow students to wear blackface for Halloween.*
	Placebo Statements
11	Parents should limit the amount of time their kids spend on their smartphones.
12	The United States should increase tariffs on foreign imports.*
13	School uniforms help reduce clothing-related peer pressure.*
14	The one-cent coin (i.e. the penny) should be removed from circulation.
15	The members states of the European Union should cede more powers to the E.U.

Notes: The table above presents the 15 statements shown to participants in the Social Image Experiment. Some of the statements were preceded by a brief paragraph providing additional information. For instance, one such paragraph explained that "trigger warnings" are warnings that some work contains writing, images, or concepts that may be distressing to some people. See the survey instrument for details. A statement is marked with an asterisk (*) if the fraction of participants in the Exploratory Survey who answered that disagreeing with the statement is more socially acceptable at UCSB than agreeing with the statement is larger than the fraction of participants who answered the opposite. In the Exploratory Survey, answering that disagreeing with a statement is more socially acceptable at UCSB than agreeing with the statement is defined as selecting a number strictly smaller than 0 on the -5 to 5 scale; conversely, answering that agreeing with a statement is more socially acceptable at UCSB than disagreeing with the statement is defined as selecting a number strictly greater than 0 on the -5 to 5 scale.

Table 2: **Sample Sizes**

Phase		Sample size
Social Image Experiment	Completed Pre-screen	$N = 534$
	Met Eligibility Criteria	$N = 390$
	Passed Low Quality Screener	$N = 376$
	Consented to Participate	$N = 371$
	Reached Randomization Stage	$N = 360$
	Completed Survey	$N = 318$
	Included in Main Sample	$N = 304$
Forecasting Experiment	Completed Pre-screen	$N = 374$
	Met Eligibility Criteria	$N = 254$
	Passed Low Quality Screener	$N = 244$
	Consented to Participate	$N = 225$
	Completed Survey	$N = 219$
	Included in Main Sample	$N = 209$

Notes: The table above presents the size of the samples at different stages of the Main and Forecasting Experiments. The number of subjects included in the main sample is smaller than the number of subjects that completed the survey, because, as specified in the pre-analysis plans, the main sample excludes the bottom 5% of participants in terms of survey duration.

Table 3: **Average Treatment Effects: Open Responses**

	Log Odds		Average Marginal Effects	
	(1)	(2)	(3)	(4)
Public Treatment	0.41*	0.55*	0.09*	0.11**
	(0.24)	(0.29)	(0.05)	(0.05)
Controls	No	Yes	No	Yes
Observations	304	289	304	289
Fraction Private Treatment	0.30	0.30	0.30	0.30

Notes: The table above presents the results, both in log odds and in average marginal effects, of a logit model of whether a participant finds it is acceptable to disrupt the talks of controversial speakers invited to campus to deliver lectures on a treatment indicator and, depending on the specification, pre-specified controls. The Social Image Experiment included the following question with an open response text box: "Charles Murray is a political scientist who co-authored a controversial book that, among other things, discusses racial differences in intelligence and who wrote in a 2005 essay that 'no woman has been a significant original thinker in any of the world's great philosophical traditions'. If some student group at UCSB invited Charles Murray on campus to give a talk, would it be acceptable for other students to disrupt the talk and prevent Murray from delivering his lecture or would it not be acceptable?". The responses where independently coded by two Amazon Mechanical Turk (mTurk) workers who were blind to treatment status. The dependent variables equals 1 if both mTurk workers coded the open-ended response as supporting the view that it is acceptable to disrupt the talk of controversial speakers and 0 otherwise. Standard errors are in parentheses.

Table 4: **Forecast of Fraction Agreeing/Disagreeing given Political Ideology**

Agreeing with statement about reparations for slavery

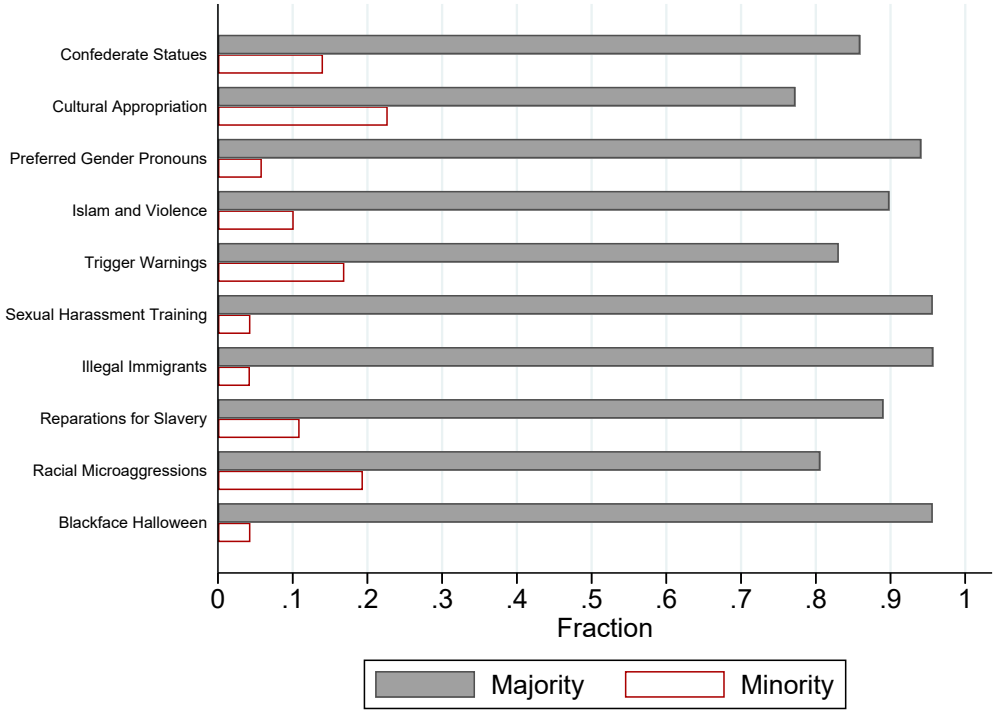
	(1) Actual Private	(2) Actual Public	(3) Forecasted Public
Liberals	0.78	0.76	0.81
Moderates/Conservatives	0.27	0.50	0.32

Disagreeing with statement about illegal immigration

	(1) Actual Private	(2) Actual Public	(3) Forecasted Public
Liberal	0.88	0.85	0.89
Moderates/Conservatives	0.44	0.62	0.48

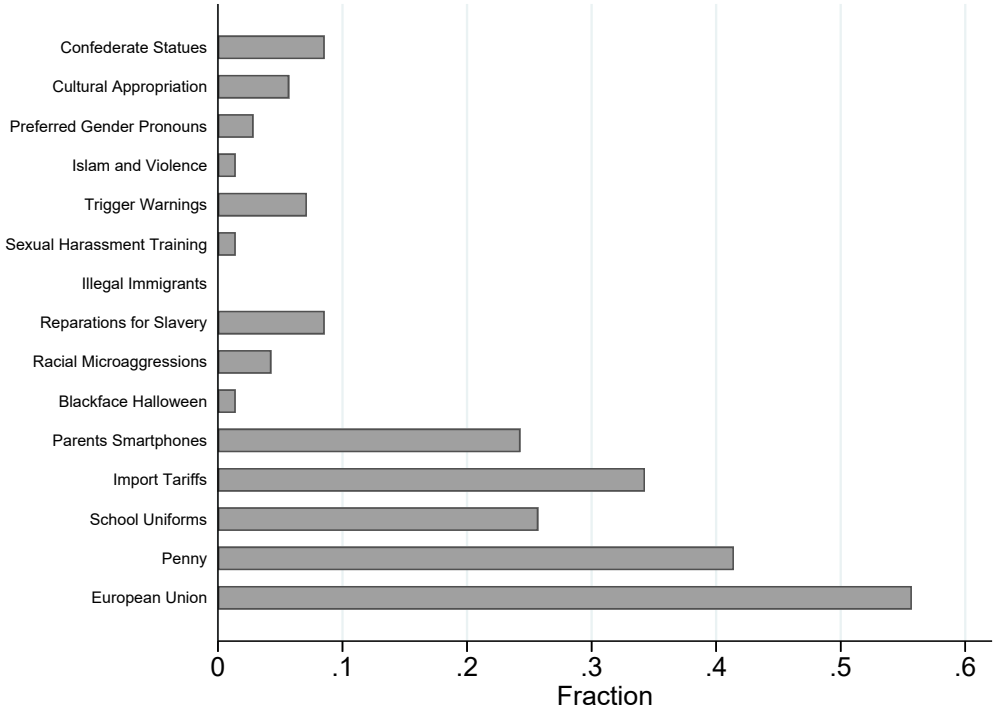
Notes: The first column of the top panel reports the fraction of participants, among the ones assigned to the Private Treatment that are classified as liberal (moderates/conservatives), who reported agreeing with the statement about reparations for slavery. The second column of the top panel reports the fraction of participants, among the ones assigned to the Public Treatment that are classified as liberal (moderates/conservatives), who reported agreeing with the statement about reparations for slavery. The third column of the top panel reports the average forecast, among participants in the Forecasting Experiment, of the numbers in the second column. The first column of the bottom panel reports the fraction of participants, among the ones assigned to the Private Treatment that are classified as liberal (moderates/conservatives), who reported disagreeing with the statement about illegal immigration. The second column of the bottom panel reports the fraction of participants, among the ones assigned to the Public Treatment that are classified as liberal (moderates/conservatives), who reported disagreeing with the statement about illegal immigration. The third column of the bottom panel reports the average forecast, among participants in the Forecasting Experiment, of the numbers in the second column. Disagreeing with a statement is defined as answering a number strictly smaller than 5 on the 0-10 Likert scale; agreeing with a statement is defined as answering a number strictly greater than 5. The classification into liberals and moderates/conservatives is based on the median split in terms of self-reported political ideology described in Section 4.3.

Figure 1: **Extent of Agreement about Socially Acceptable Position at UCSB**



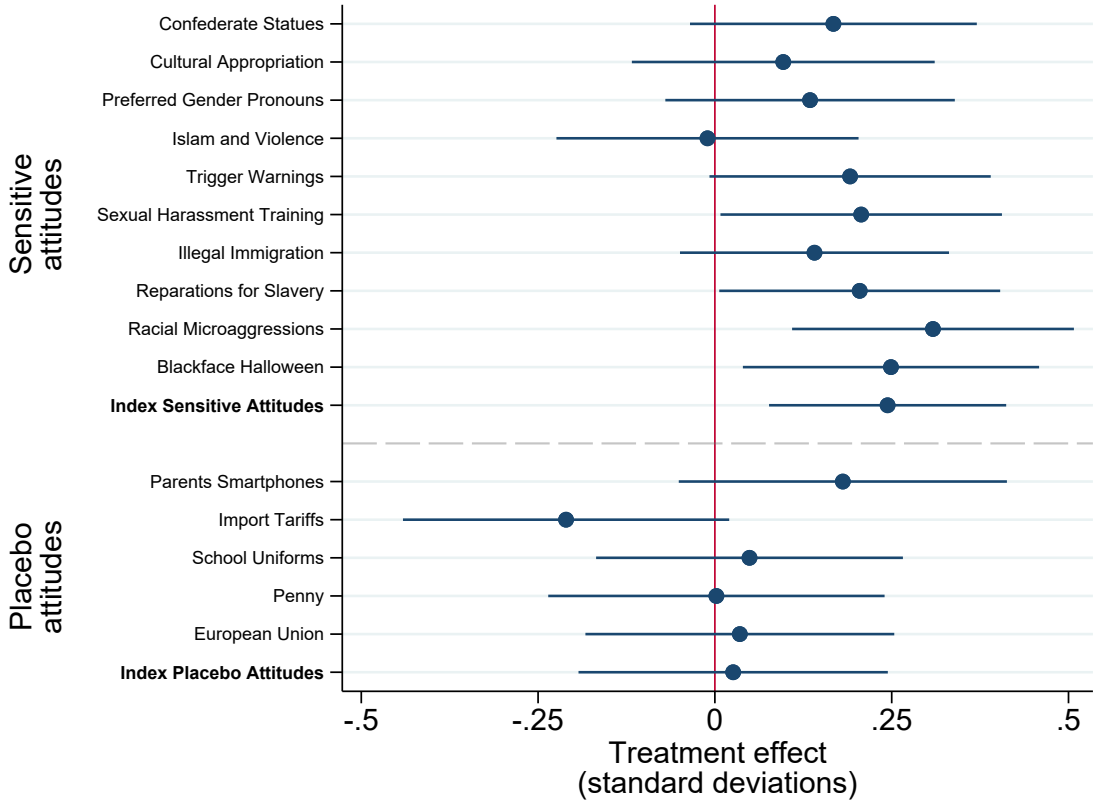
Notes: The figure above presents, for each of the sensitive statements, the relative proportions of participants in the Exploratory Survey whose opinion as to whether agreeing or disagreeing with the statement is more socially acceptable at UCSB aligns with that of the majority and of participants whose opinion aligns with that of the minority. The fractions in the figure are taken from the first two columns of Appendix Table A3 and normalized to sum to one.

Figure 2: **Degree of Social Image Engagement at UCSB by Statement**



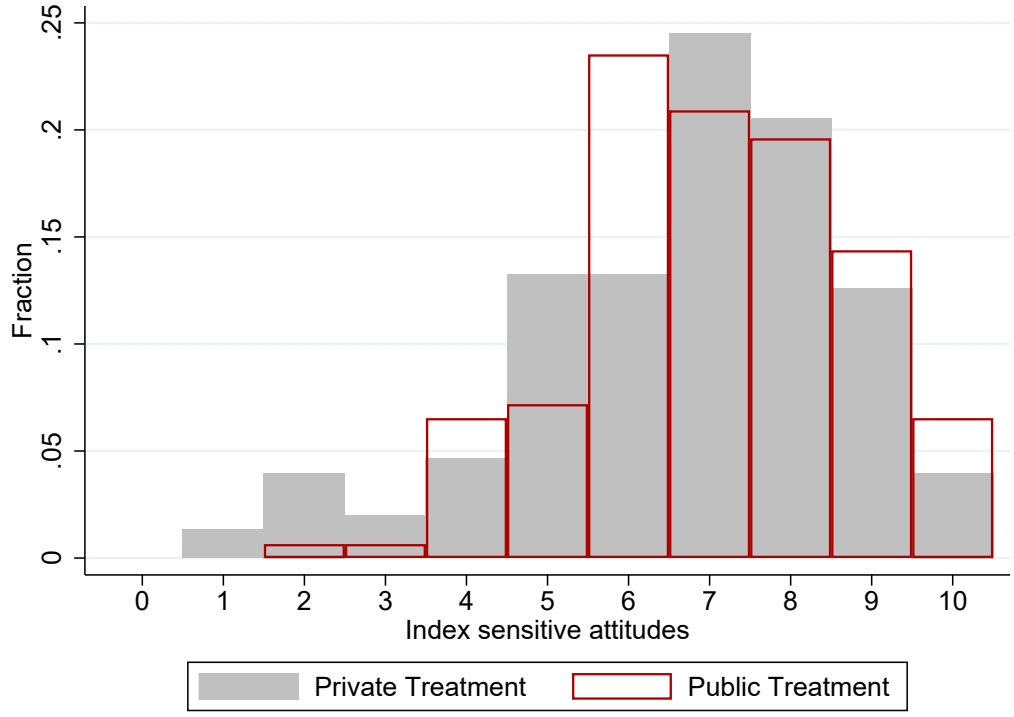
Notes: The figure above presents, for each of the sensitive and the placebo statements, the fraction of participants in the Exploratory Survey who answered that, at UCSB, agreeing and disagreeing with the statement is about the same in terms of social acceptability. Answering that, at UCSB, agreeing and disagreeing with a statement are about the same in terms of social acceptability is defined as selecting 0 on the -5 to 5 scale.

Figure 3: **Average Treatment Effects: Sensitive and Placebo Attitudes**



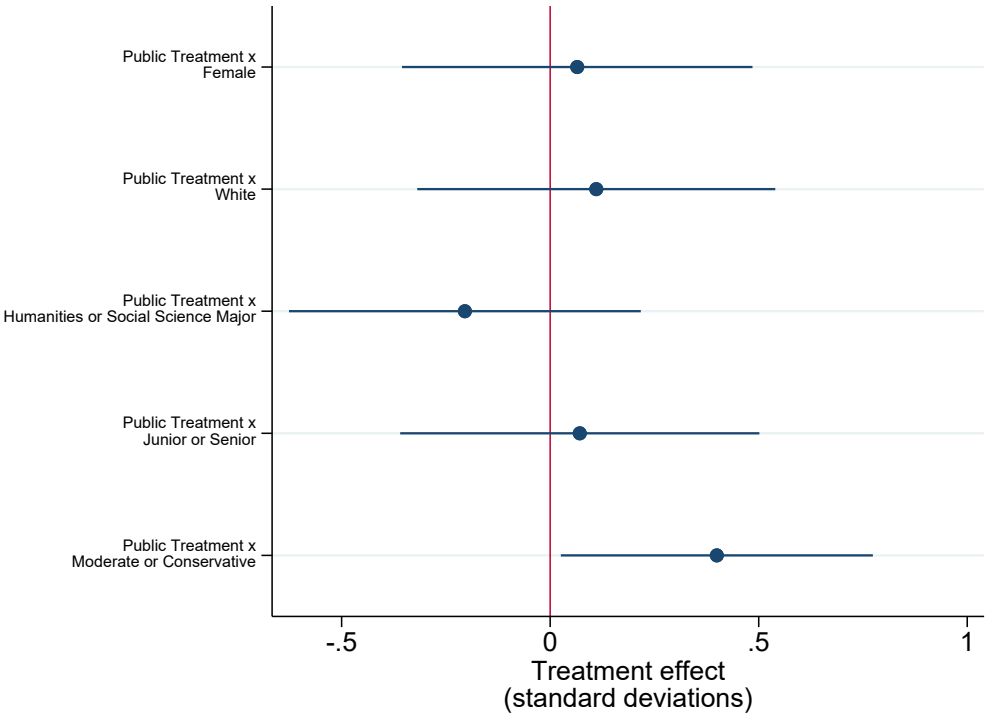
Notes: The figure above presents average treatment effects of being assigned to the Public Treatment using Equation 1. The students' reported levels of agreement with the sensitive and placebo statements are oriented in such a way that larger numbers always correspond to views that, according to the Exploratory Survey, are generally perceived to be more socially acceptable at UCSB. The index of sensitive (placebo) attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive (placebo) questions, with the answers already re-oriented. All variables are normalized so that the distribution of answers of participants in the Private Treatment has a standard deviation of one and a mean of zero. Error bars reflect 95 percent confidence intervals.

Figure 4: **Histogram of the Index of Sensitive Attitudes**



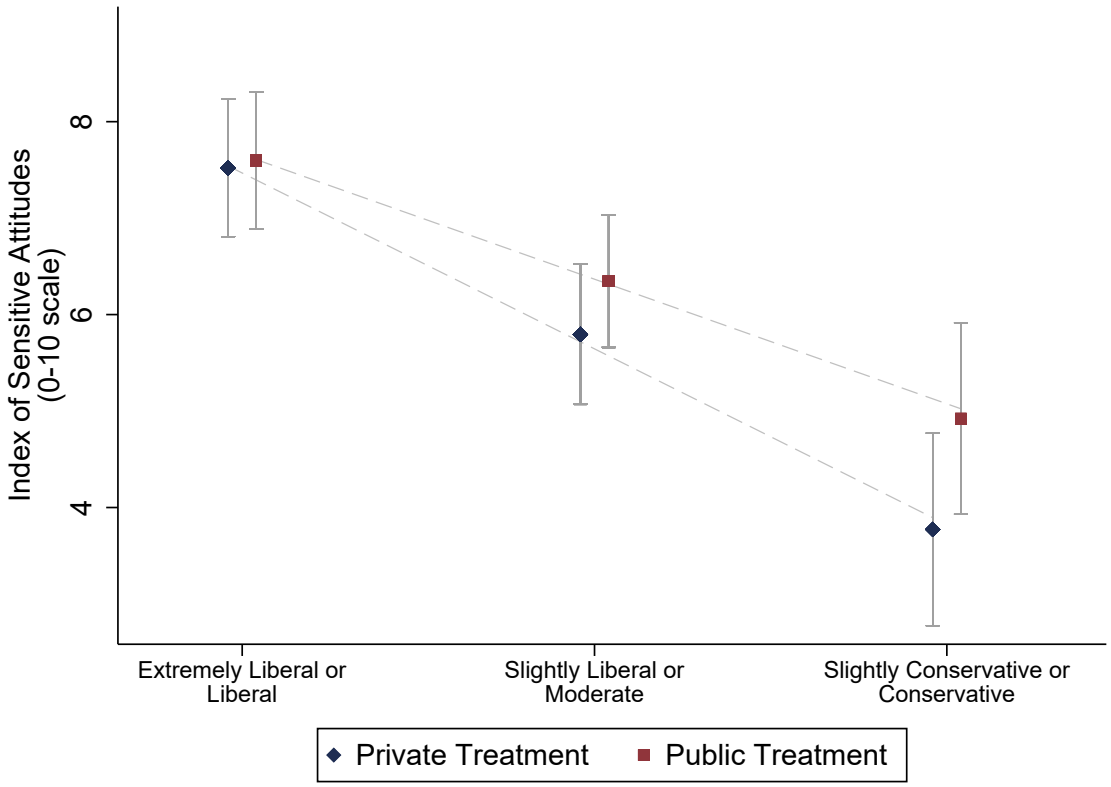
Notes: The figure above presents overlayed histograms of the index of sensitive attitudes for participants in the Private and in the Public Treatments. The index of sensitive attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey.

Figure 5: **Heterogeneous Treatment Effects on the Index of Sensitive Attitudes**



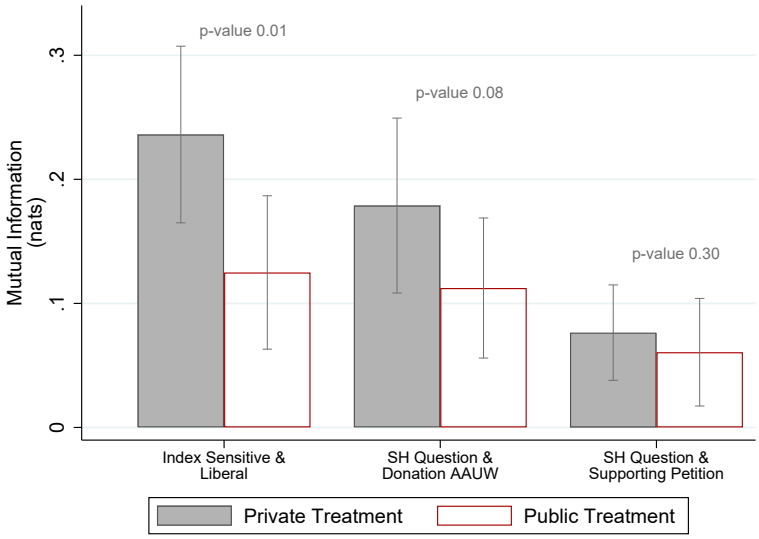
Notes: The figure above presents the coefficient on the interaction term in the following regression equation $Y_i = \alpha + \beta T_i + \delta M_i + \gamma T_i \times M_i + \varepsilon_i$, where Y_i denotes the standardized index of sensitive attitudes, T_i is an indicator that takes value 1 if individual i is assigned to the Public Treatment, M_i is an indicator for the category of the binary moderator participant i belongs to, and ε_i is an idiosyncratic error term. The index of sensitive attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. The normalization is achieved by first subtracting from an index the average value of the index among participants in the Private Treatment and then dividing the result by the standard deviation of the index among participants in the Private Treatment. Error bars reflect 95 percent confidence intervals.

Figure 6: **Index of Sensitive Attitudes by Treatment Status and Political Ideology**



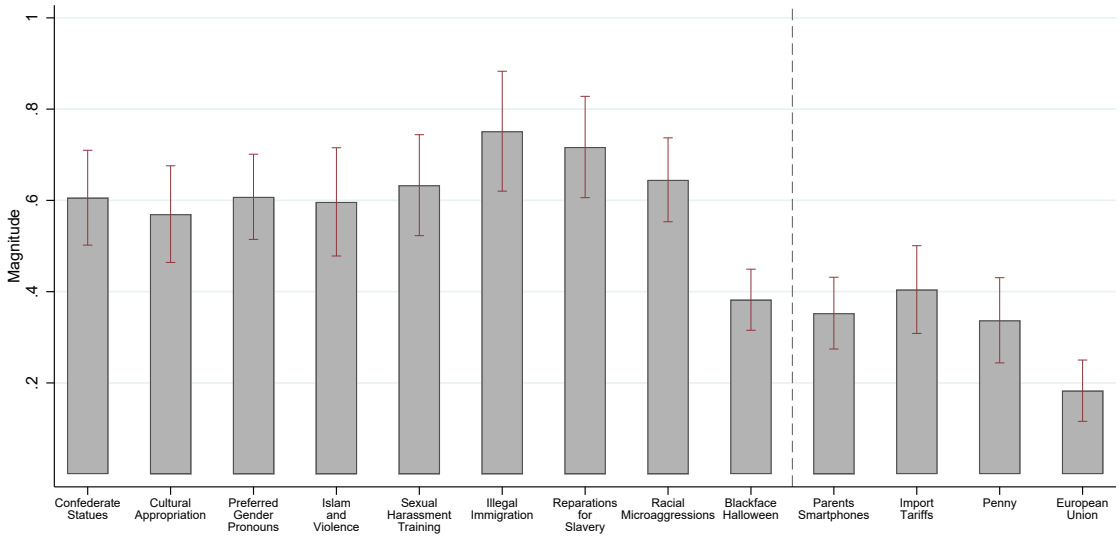
Notes: The figure above presents averages of the index of sensitive attitudes (in original units) by treatment status and by different categories of self-reported political ideology. Since no participant in the Social Image Experiment self-identified as Extremely Conservative, the Extremely Conservative category is omitted. The index of sensitive attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. Error bars reflect standard deviations of the underlying answers. See the discussion in Section 4.6 for details as to why standard deviations, rather than confidence intervals, are reported in the figure.

Figure 7: **Mutual Information**



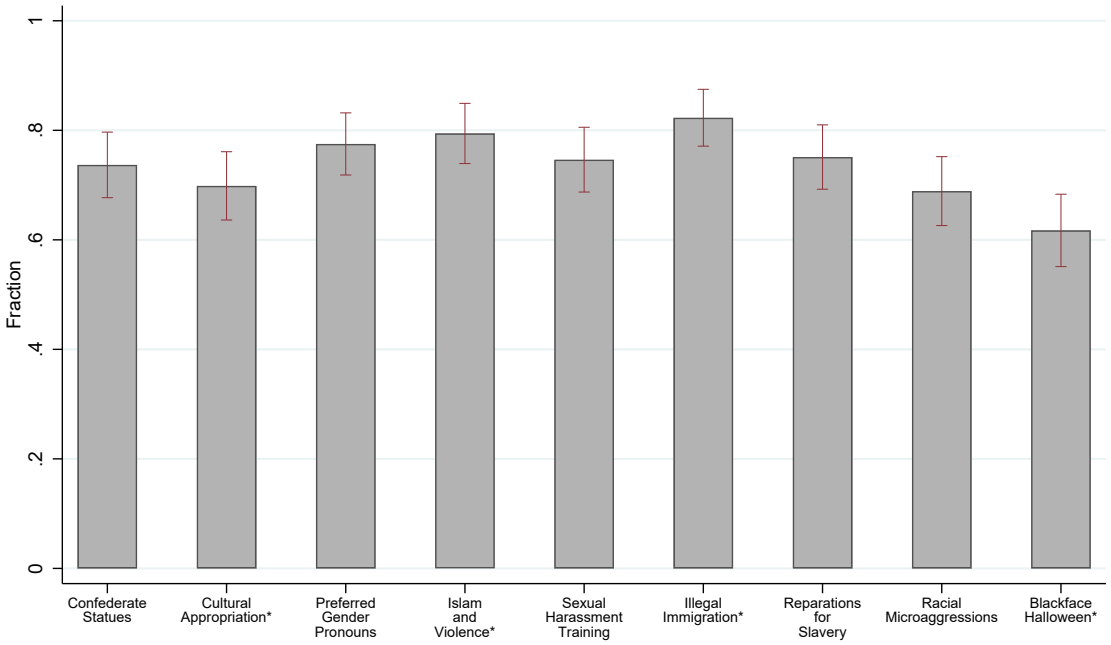
Notes: The figure above presents estimates of mutual information (in nats) for different duples of variables and separately for participants in the Private and the Public Treatments. The first set of columns shows estimates of the mutual information between a participant's index of sensitive attitudes, rounded to the nearest digit, and whether the participant self-identified as liberal. The index of sensitive attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. Self-identifying as liberal is defined according to the median split in terms of political ideology described in Section 4.3. The second set of columns shows estimates of the mutual information between a participant's reported level of agreement with the statement about sexual harassment (Statement 6 in Table 1) and the quartile of the participant's donation to the AAUW. The third set of columns shows estimates of the mutual information between a participant's reported level of agreement with the statement about sexual harassment (Statement 6 in Table 1) and whether the participant supported the anonymous petition to require mandatory yearly sexual harassment training at UCSB. Error bars reflect 95 percent confidence intervals. The p-values refer to one-sided tests where the null hypothesis is that mutual information in the Public Treatment is weakly larger than mutual information in the Private Treatment. A one-sided t-test is appropriate in light of the prediction of the model that mutual information should be higher in the Private Treatment than in the Public Treatment.

Figure 8: **Forecasted Magnitude of Distortions**



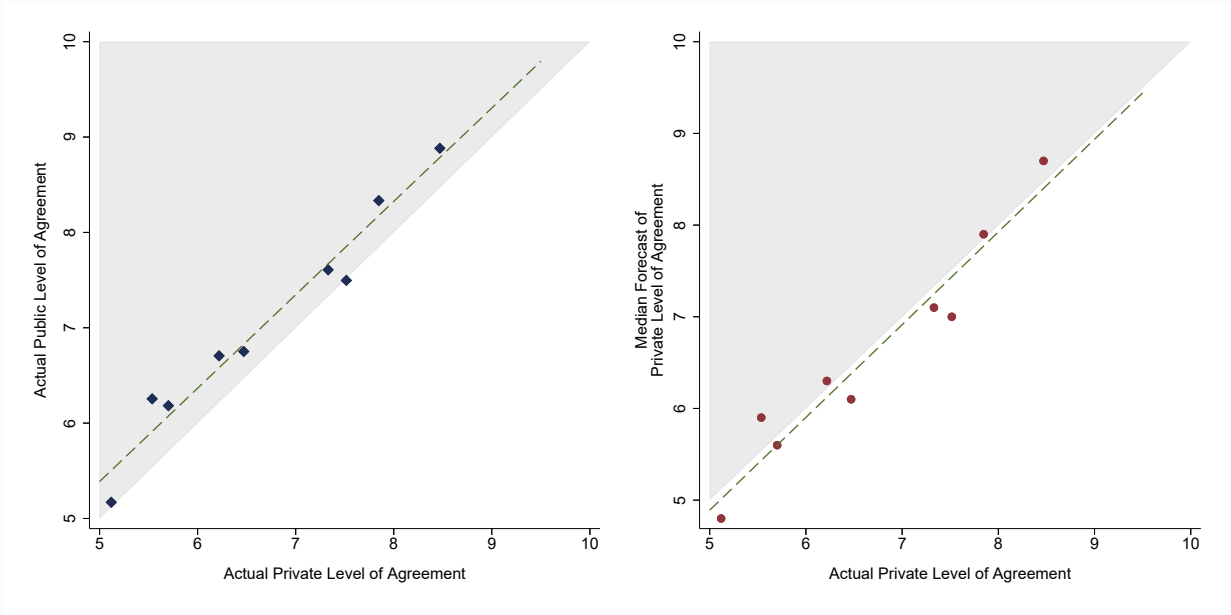
Notes: The figure above shows, for 13 out of the 15 statements shown to participants in the Social Image Experiment, the average absolute distance between the subjects' forecasts and the actual average answers of participants in the Public Treatment of the Social Image Experiment. One of the sensitive and one of the placebo statements are omitted from the figure because they were included as examples in the instructions of the Forecasting Experiment. Error bars reflect 95 percent confidence intervals.

Figure 9: **Forecasted Direction of Distortions**



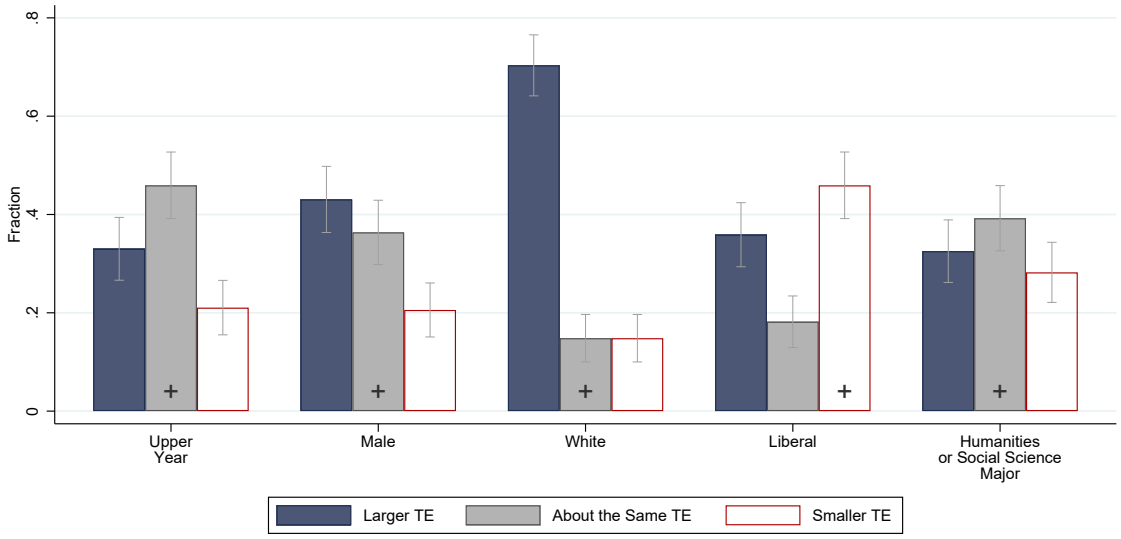
Notes: The figure above shows, for 9 out of the 10 sensitive statements, the fraction of participants in the Forecasting Experiment whose forecasts reflect a correct understanding of the direction in which public answers are on average skewed compared to private answers. For each statement, a subject's forecast is classified as reflecting a correct understanding of the direction in which public answers are on average skewed compared to private answers if the actual average answer of participants in the Private Treatment and the subject's forecast are either both larger than the actual average answer of participants in the Public Treatment or both smaller. One of the sensitive statements is omitted from the figure because it was included as an example in the instructions of the Forecasting Experiment. Error bars reflect 95 percent confidence intervals.

Figure 10: **Forecasts: Sensitive Statements**



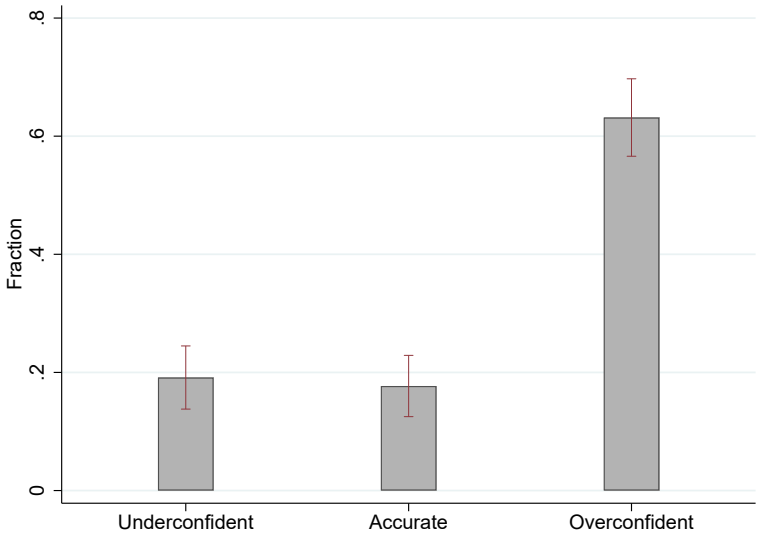
Notes: The left panel of the figure reiterates the results of the Social Image Experiment. It presents a scatter plot of the average level of agreement with each statement of participants in the Private Treatment of the Social Image Experiment against the average level of agreement with each statement of participants in the Public Treatment of the Social Image Experiment. Each diamond represents one of the sensitive statements shown to participants in the Social Image Experiment, excluding the statement shown in the instructions of the Forecasting Experiment. For each statement, the answers are oriented in such a way that larger numbers always correspond to views that, according to the Exploratory Survey, are generally perceived to be more socially acceptable at UCSB. Therefore, diamonds falling above the 45 degree line indicate statements for which the answers of participants in the Public Treatment of the Social Image Experiment are on average skewed in the direction that is perceived to be more socially acceptable at UCSB compared to the answer of participants in the Private Treatment of the Social Image Experiment. The right panel of the figure presents a scatter plot of the average level of agreement with each statement of participants in the Private Treatment of the Social Image Experiment against the median forecast of participants in the Forecasting Experiment. Each diamond represents one of the sensitive statements shown to participants in the Social Image Experiment, excluding the statement shown in the instructions of the Forecasting Experiment. For each statement, the answers are oriented in such a way that larger numbers always correspond to views that, according to the Exploratory Survey, are generally perceived to be more socially acceptable at UCSB. Diamonds falling on the 45 degree line indicate statements for which the median forecast of participants in the Forecasting Experiment is perfectly accurate. In each panel, the dashed line represents the line of best fit.

Figure 11: **Forecasted Heterogeneity in Treatment Effects**



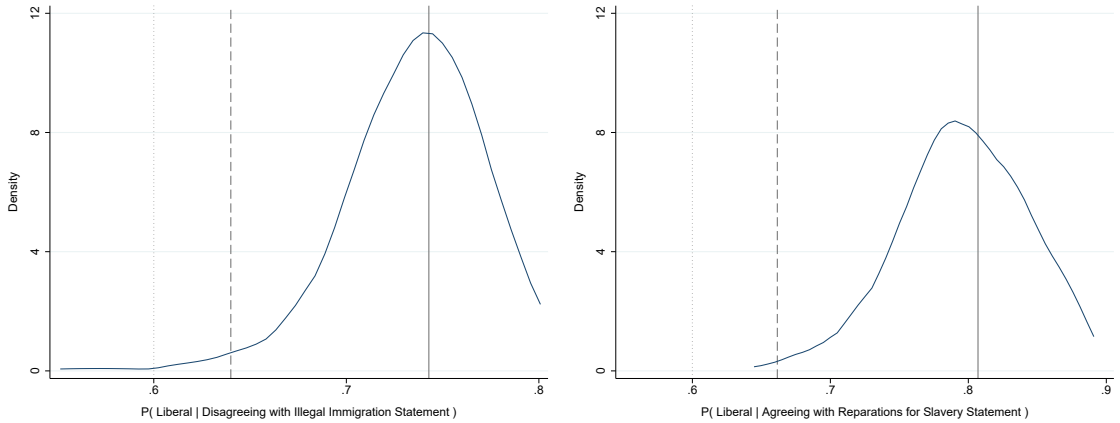
Notes: The figure above shows the distribution of answers of participants in the Forecasting Experiment to the multiple-choice questions about heterogeneity. The first bar in each set represents the fraction of participants in the Forecasting Experiment who answered that the demographic group on the x-axis exhibited a significantly larger treatment effect than its complement. The second bar in each set represents the fraction of participants in the Forecasting Experiment who answered that the demographic group on the x-axis and its complement exhibited treatment effects that are not significantly different from one another. Finally, the third bar in each set represents the fraction of participants in the Forecasting Experiment who answered that the demographic group on the x-axis exhibited a significantly smaller treatment effect than its complement. The pluses (+) represent the answers that match the results of the heterogeneous treatment effect analysis of the Social Image Experiment. Error bars reflect 95 percent confidence intervals.

Figure 12: **Confidence**



Notes: The figure above reports the fraction of participants in the Forecasting Experiment who are classified as underconfident, accurate, and overconfident. A participant is classified as underconfident if the difference between the forecasted number of correct answers to the multiple-choice questions about treatment effect heterogeneity and the actual number of correct answers is strictly negative. A participant is classified as accurate if the difference between the forecasted number of correct answers to the multiple-choice questions about treatment effect heterogeneity and the actual number of correct answers equals 0. Finally, a participant is classified as overconfident if the difference between the forecasted number of correct answers to the multiple-choice questions about treatment effect heterogeneity and the actual number of correct answers is strictly positive. The answer of a participant in the Forecasting Experiment is considered correct if it matches the result of the heterogeneous treatment effect analysis in the Social Image Experiment. Error bars reflect 95 percent confidence intervals.

Figure 13: **Perceived Diagnosticity of Statements for Political Ideology**



Notes: The figure above reports, for participants in the Forecasting Experiment, the empirical distributions of the perceived diagnosticity of publicly disagreeing with the statement about illegal immigration and of publicly agreeing with the statement about reparations for slavery for self-identifying as a liberal. "Publicly agreeing" and "publicly disagreeing" refer to the fact that the figure shows the diagnosticity, as perceived by participants in the Forecasting Experiment, of the answers of participants in the Public Treatment of the Social Image Experiment. Disagreeing with a statement is defined as answering a number strictly smaller than 5 on the 0-10 Likert scale; agreeing with a statement is defined as answering a number strictly greater than 5. Self-identifying as a liberal is defined according to the median split in terms of self-reported political ideology described in Section 4.3. For each participant in the Forecasting Experiment, the perceived diagnosticity of publicly disagreeing with the statement about illegal immigration and of publicly agreeing with the statement about reparations for slavery was calculated via Bayes' rule using the participant's answers. The dotted lines represent the unconditional proportion of participants in the Public Treatment of the Social Image Experiment who are classified as liberals. The dashed lines represent, for participants in the Public Treatment of the Social Image Experiment, the estimated probability of being classified as liberal given the fact that the subject reported disagreeing with the statement about illegal immigration or given the fact that the subject reported agreeing with the statement about reparations for slavery. The solid lines represent, for participants in the Private Treatment of the Social Image Experiment, the estimated probability of being classified as liberal given the fact that the subject reported disagreeing with the statement about illegal immigration or given the fact that the subject reported agreeing with the statement about reparations for slavery.

Online Appendix: Not for Publication

Political Correctness, Social Image, and Information Transmission

Luca Braghieri

A Proofs

We first prove the propositions and then the claim from Section 2.2.

A.1 Proof of Proposition 1

We prove the proposition by means of a series of lemmas.

Lemma 1. *$\forall e \in [0,1]$, a pure-strategy equilibrium exists.*

Proof
If $e = 0$, truth-telling, together with the corresponding beliefs, is trivially an equilibrium.
Consider $e > 0$. Let's start by fixing $\{E_\beta(\pi|x=1), E_\beta(\pi|x=2)..., E_\beta(\pi|x=\bar{\pi})\}$ and finding the agents' best response functions.
Plainly, every type $(\pi, \varsigma) \in \Pi \times [0, \bar{\varsigma}]$ prefers to choose $x = \pi$ to choosing any $\tilde{x}$ s.t. $E_\beta(\pi|\tilde{x}) < E_\beta(\pi|\pi)$. Therefore, it cannot be a best response for the agent to choose any $\tilde{x}$ s.t. $E_\beta(\pi|\tilde{x}) < E_\beta(\pi|\pi)$.
Fix $\pi \in \Pi$ and let $A_\pi = \{x \in X | E_\beta(\pi|x) \geq E_\beta(\pi|\pi)\}$. $\forall x \in A_\pi$, let's sort the $E_\beta(\pi|x)$ in descending order and, if there are any ties, let's consider only the x closest to π. If the x closest to π is not unique, let's consider the larger one. Call the set thus refined B_π. By construction, the $\forall x \in B_\pi$, the $E_\beta(\pi|x)$ can be strictly well ordered. Furthermore, by construction, the best response of any agent with type (π, ς) must be an element of B_π. Notice B_π is not empty, because $\pi \in B_\pi$. Finally, $\forall x \in B_\pi$, let $B^u_{\pi,x} = \{z \in B_\pi | E_\beta(\pi|z) > E_\beta(\pi|x)\}$ and $B^l_{\pi,x} = \{z \in B_\pi | E_\beta(\pi|z) < E_\beta(\pi|x)\}$.
Consider any $\pi \in \Pi$ for which $B_\pi = \{\pi\}$. Then, all types (π, ς) for $\varsigma \in [0, \bar{\varsigma}]$ prefer choosing $x = \pi$ to any $x \neq \pi$.
Now consider any $\pi \in \Pi$ for which $B_\pi \neq \{\pi\}$. We next determine what types (π, ς) of the agent, if any, choose each $x \in B_\pi$.
Type (π, ς) chooses $x = \pi$ if

$$e\varsigma E_\beta(\pi|\pi) - c(\pi, \pi) \geq e\varsigma E_\beta(\pi|\tilde{x}) - c(\tilde{x}, \pi) \ \ \forall \tilde{x} \in B^u_{\pi,\pi}$$

$$\varsigma \leq \min_{\tilde{x} \in B^u_{\pi,\pi}} \left\{ \frac{c(\tilde{x}, \pi)}{e\left[E_\beta(\pi|\tilde{x}) - E_\beta(\pi|\pi)\right]} \right\}$$

Notice that $x = \pi$ is a best response for a positive measure of types (π, ς).
Consider $x^M = \arg\max_{x \in B_\pi} E_\beta(\pi|x)$. Type (π, ς) chooses $x = x^M$ if

$$e\varsigma E_\beta\left(\pi|x^M\right) - c\left(x^M, \pi\right) \geq e\varsigma E_\beta(\pi|\tilde{x}) - c(\tilde{x}, \pi) \ \ \forall \tilde{x} \in B^l_{\pi,x^M}$$

$$\varsigma \geq \max_{\tilde{x} \in B^l_{\pi,x^M}} \left\{ \frac{c\left(x^M, \pi\right) - c(\tilde{x}, \pi)}{e\left[E_\beta(\pi|x^M) - E_\beta(\pi|\tilde{x})\right]} \right\}$$

Now consider any $x \in B_\pi \backslash \left\{\pi, x^M\right\}$. Type (π, ς) chooses x if

$$e\varsigma E_\beta(\pi|x) - c(x, \pi) \geq e\varsigma E_\beta(\pi|\tilde{x}) - c(\tilde{x}, \pi) \ \ \forall \tilde{x} \in B^u_{\pi,x}$$

and

$$e\varsigma E_\beta(\pi|x) - c(x, \pi) \geq e\varsigma E_\beta(\pi|\tilde{x}) - c(\tilde{x}, \pi) \ \ \forall \tilde{x} \in B^l_{\pi,x}$$

We can rewrite the inequalities above as

$$\varsigma \leq \min_{\tilde{x} \in B^u_{\pi,x}} \left\{ \frac{c(\tilde{x}, \pi) - c(x, \pi)}{e\left[E_\beta(\pi|\tilde{x}) - E_\beta(\pi|x)\right]} \right\}$$

and

$$\varsigma \geq \max_{\tilde{x}\in B^{l}_{\pi,x}} \left\{ \frac{c\left(x,\pi\right)-c\left(\tilde{x},\pi\right)}{e\left[E_{\beta}\left(\pi|x\right)-E_{\beta}\left(\pi|\tilde{x}\right)\right]} \right\}$$

Therefore, there exists a $\varsigma \in [0,\bar{\varsigma}]$ such that x is a best response for type (π,ς) iff the following condition is satisfied:

$$\max_{\tilde{x}\in B^{l}_{\pi,x}} \left\{ \frac{c\left(x,\pi\right)-c\left(\tilde{x},\pi\right)}{e\left[E_{\beta}\left(\pi|x\right)-E_{\beta}\left(\pi|\tilde{x}\right)\right]} \right\} \leq \min_{\tilde{x}\in B^{u}_{\pi,x}} \left\{ \frac{c\left(\tilde{x},\pi\right)-c\left(x,\pi\right)}{e\left[E_{\beta}\left(\pi|\tilde{x}\right)-E_{\beta}\left(\pi|x\right)\right]} \right\}$$

Let's refer to the condition above as condition $\star_{\pi,x}$.

If condition $\star_{\pi,x}$ is satisfied, the types (π,ς) for whom x is a best response are types with

$$\varsigma \in \left[\max_{\tilde{x}\in B^{l}_{\pi,x}} \left\{ \frac{c\left(x,\pi\right)-c\left(\tilde{x},\pi\right)}{e\left[E_{\beta}\left(\pi|x\right)-E_{\beta}\left(\pi|\tilde{x}\right)\right]} \right\}, \min_{\tilde{x}\in B^{u}_{\pi,x}} \left\{ \frac{c\left(\tilde{x},\pi\right)-c\left(x,\pi\right)}{e\left[E_{\beta}\left(\pi|\tilde{x}\right)-E_{\beta}\left(\pi|x\right)\right]} \right\} \right]$$

Notice that, for some types $(\pi,\varsigma)\in\Pi\times[0,\bar{\varsigma}]$, the best response is a correspondence rather than a function. In order to obtain a best-response function, pick the smallest x from the best-response correspondence.

This characterizes the agent's best response function to $\{E_{\beta}\left(\pi|x=1\right),E_{\beta}\left(\pi|x=2\right)...,E_{\beta}\left(\pi|x=\bar{\pi}\right)\}$.

Let's now consider at the audience's inferences. The audience's inferences are derived by applying Bayes' rule to the agents' best-response functions. $\forall\pi\in\Pi$ and $x\in X$, let $C_{\pi,x}=\{(\tilde{\pi},\varsigma)\in\Pi\times[0,\bar{\varsigma}]\,|\tilde{\pi}=\pi\ \wedge\ x^{*}\left(\tilde{\pi},\varsigma\right)=x\}$. By the construction of the agents' best-response functions, we know the $C_{\pi,x}$ are intervals. Let $|C_{\pi,x}|=\int_{C_{\pi,x}} f\left(z|\pi\right)dzP\left(\pi\right)$. Then, $E_{\beta}\left(\pi|x\right)$ can be written as:

$$E_{\beta}\left(\pi|x\right)=\frac{1}{\sum\limits_{z\in\Pi}|C_{z,x}|}\left(\sum_{m\in\Pi}m\,|C_{m,x}|\right)$$

We know the expression above is well-defined because, as already shown, $\forall\pi\in\Pi$, a positive measure of types (π,ς) chooses $x=\pi$. In other words, all statements are on the equilibrium path; therefore, all beliefs can be derived via Bayes' rule.

$\forall\pi\in\Pi\backslash\{\bar{\pi}\}$ and $\forall x\in X$, $|C_{\pi,x}|$ can be written as:

$$|C_{\pi,x}|=\begin{cases} 0 & if\ x\notin B_{\pi}\ or\ \left[x\in B_{\pi}\backslash\left\{\pi,x^{M}\right\}\ and\ \star_{\pi,x}\ is\ not\ satisfied\right] \\ \int_{\min\limits_{\tilde{x}\in B^{u}_{\pi,x}}\left\{\frac{c(\tilde{x},\pi)-c(x,\pi)}{e\left[E_{\beta}(\pi|\tilde{x})-E_{\beta}(\pi|x)\right]}\right\}}^{\max\limits_{\tilde{x}\in B^{l}_{\pi,x}}\left\{\frac{c(x,\pi)-c(\tilde{x},\pi)}{e\left[E_{\beta}(\pi|x)-E_{\beta}(\pi|\tilde{x})\right]}\right\}} f\left(z|\pi\right)dzP\left(\pi\right) & if\ x\in B_{\pi}\backslash\left\{\pi,x^{M}\right\}\ and\ \star_{\pi,x}\ is\ satisfied \\ \int_{0}^{\min\limits_{\tilde{x}\in B^{u}_{\pi,\pi}}\left\{\frac{c(\tilde{x},\pi)}{e\left[E_{\beta}(\pi|\tilde{x})-E_{\beta}(\pi|\pi)\right]}\right\}} f\left(z|\pi\right)dzP\left(\pi\right) & if\ x=\pi \\ \int_{\max\limits_{\tilde{x}\in B^{l}_{\pi,x^{M}}}\left\{\frac{c\left(x^{M},\pi\right)-c(\tilde{x},\pi)}{e\left[E_{\beta}\left(\pi|x^{M}\right)-E_{\beta}(\pi|\tilde{x})\right]}\right\}}^{\infty} f\left(z|\pi\right)dzP\left(\pi\right) & if\ x=x^{M} \end{cases}$$

noticing that $\forall\pi\in\Pi$, $f\left(\varsigma|\pi\right)=0$ whenever $\varsigma>\bar{\varsigma}$.

For $\pi=\bar{\pi}$, $|C_{\pi,x}|=0$ if $x\neq\bar{\pi}$ and $|C_{\pi,x}|=P\left(\bar{\pi}\right)$ if $x=\bar{\pi}$.

We can now apply Brouwer's fixed-point theorem. We know that $\forall x\in X$, $E_{\beta}\left(\pi|x\right)\in[1,\bar{\pi}]$. We define a function $q:\ [1,\bar{\pi}]^{n}\rightarrow[1,\bar{\pi}]^{n}$ as follows

$$q\left(E_{\beta}\left(\pi|x=1\right),...,E_{\beta}\left(\pi|x=\bar{\pi}\right)\right)=\left(\frac{1}{\sum\limits_{z\in\Pi}|C_{z,1}|}\left(\sum_{m\in\Pi}m\,|C_{m,1}|\right),...,\frac{1}{\sum\limits_{z\in\Pi}|C_{z,\bar{\pi}}|}\left(\sum_{m\in\Pi}m\,|C_{m,\bar{\pi}}|\right)\right)$$

where the $|C_{\pi,x}|$ are calculated according to the expressions above. q is a continuous function from a convex compact subset of a Euclidean space to itself; therefore, by Brouwer's fixed-point theorem, it has a fixed point. □

Lemma 2. *$\forall e\in[0,1]$, if any misreporting occurs in equilibrium, it occurs in the socially acceptable direction; i.e.*

$x^*(\pi,\varsigma) \geq \pi$.

Proof
We prove the lemma by induction. Base case: $\forall(\pi,\varsigma)$ with $\pi = \bar{\pi}$ and $\varsigma \in [0,\bar{\varsigma}]$, $x^*(\bar{\pi},\varsigma) = \bar{\pi}$.
Suppose, aiming towards contradiction, that $\exists$ an equilibrium in which, for some (π,ς) with $\pi = \bar{\pi}$, $\varsigma \in [0,\bar{\varsigma}]$, $x^*(\bar{\pi},\varsigma) \neq \bar{\pi}$. Consider any type $(\bar{\pi},\varsigma)$ s.t. $x^*(\bar{\pi},\varsigma) = \tilde{x} < \bar{\pi}$. It must be the case that, in equilibrium, $E_\beta(\pi|\tilde{x}) > E_\beta(\pi|\bar{\pi})$.
Notice it cannot be the case that, in equilibrium, $E_\beta(\pi|x = \bar{\pi}-1) > E_\beta(\pi|x=\bar{\pi})$. Suppose it was the case. Then, no type (π,ς) with $\pi \leq \bar{\pi}-1$, $\varsigma \in [0,\bar{\varsigma}]$ would choose $x = \bar{\pi}$, because

$$e\varsigma E_\beta(\pi|\bar{\pi}-1) - c(\bar{\pi}-1,\pi) \geq e\varsigma E_\beta(\pi|\bar{\pi}) - c(\bar{\pi},\pi)$$

But then, $E_\beta(\pi|\bar{\pi}) = \bar{\pi} > E_\beta(\pi|x=\bar{\pi}-1)$ and we would have a contradiction.
A similar line of argument shows that it cannot be the case that, in equilibrium, $E_\beta(\pi|x=\bar{\pi}-2) > E_\beta(\pi|x=\bar{\pi})$. Therefore, it must be the case that $\tilde{x} \leq \bar{\pi}-3$. We claim that no type (π,ς) with $\pi < \frac{1}{2}(\bar{\pi}+\tilde{x})$, $\varsigma \in [0,\bar{\varsigma}]$ chooses $x = \bar{\pi}$ in equilibrium. That's because choosing $\tilde{x}$ yields higher social image benefit and lower misrepresentation costs. Therefore, $E(\pi|\bar{\pi}) \geq \frac{1}{2}(\bar{\pi}+\tilde{x})$.
Consider type $(\bar{\pi},\varsigma)$ for which $x^*(\bar{\pi},\varsigma) = \tilde{x} < \bar{\pi}$. It must be the case that, in equilibrium,

$$e\varsigma E_\beta(\pi|\tilde{x}) - c(\tilde{x},\bar{\pi}) \geq e\varsigma E_\beta(\pi|\bar{\pi}) - c(\bar{\pi},\bar{\pi})$$

A necessary condition for the inequality above to be satisfied is

$$e\varsigma E_\beta(\pi|\tilde{x}) - c(\tilde{x},\bar{\pi}) \geq e\varsigma\frac{1}{2}(\bar{\pi}+\tilde{x})$$

$$c(\tilde{x},\bar{\pi}) \leq e\varsigma E_\beta(\pi|\tilde{x}) - e\varsigma\frac{1}{2}(\bar{\pi}+\tilde{x})$$

A necessary condition for the inequality above to be satisfied is

$$c(\tilde{x},\bar{\pi}) \leq e\varsigma\bar{\pi} - e\varsigma\frac{1}{2}(\bar{\pi}+\tilde{x})$$

$$c(\tilde{x},\bar{\pi}) \leq \frac{1}{2}e\varsigma(\bar{\pi}-\tilde{x})$$

Finally, a necessary condition for the inequality above to be satisfied is

$$c(\tilde{x},\bar{\pi}) \leq \frac{1}{2}\bar{\varsigma}(\bar{\pi}-\tilde{x})$$

Since we assumed $c(x,\pi) > \frac{1}{2}|x-\pi|\bar{\varsigma}$, we reached the desired contradiction.
Inductive step: for sufficiently convex misrepresentation costs, if $\forall\pi > \tilde{\pi}$ and $\varsigma \in [0,\bar{\varsigma}]$ $x^*(\pi,\varsigma) \geq \pi$, then $x^*(\tilde{\pi},\varsigma) \geq \tilde{\pi}$ $\forall\varsigma \in [0,\bar{\varsigma}]$.
Since $x^*(\pi,\varsigma) \geq \pi$ $\forall\pi > \tilde{\pi}$, $E(\pi|x=\pi') \leq \tilde{\pi}$ $\forall\pi' \leq \tilde{\pi}$. Suppose, aiming towards contradiction, that $\exists$ an equilibrium in which, for some (π,ς) with $\pi = \tilde{\pi}$, $\varsigma \in [0,\bar{\varsigma}]$, $x^*(\tilde{\pi},\varsigma) < \tilde{\pi}$. Consider any type $(\tilde{\pi},\varsigma)$ s.t. $x^*(\tilde{\pi},\varsigma) = \tilde{x} < \tilde{\pi}$. It must be the case that, in equilibrium, $E_\beta(\pi|\tilde{x}) > E_\beta(\pi|\tilde{\pi})$.
Notice it cannot be the case that, in equilibrium, $E_\beta(\pi|x=\tilde{\pi}-1) > E_\beta(\pi|x=\tilde{\pi})$. Suppose it was the case. Then, no type (π,ς) with $\pi \leq \tilde{\pi}-1$, $\varsigma \in [0,\bar{\varsigma}]$ would choose $x = \tilde{\pi}$, because

$$e\varsigma E_\beta(\pi|\tilde{\pi}-1) - c(\tilde{\pi}-1,\pi) \geq e\varsigma E_\beta(\pi|\tilde{\pi}) - c(\tilde{\pi},\pi)$$

The above, together with the inductive hypothesis, would imply $E_\beta(\pi|\tilde{\pi}) = \tilde{\pi} > E_\beta(\pi|x=\tilde{\pi}-1)$ which would yield a contradiction. If $\tilde{\pi} = 2$, we have reached the desired contradiction.

If $\tilde{\pi} > 2$, a similar line of argument shows that it cannot be the case that, in equilibrium, $E_\beta(\pi|x = \tilde{\pi} - 2) > E_\beta(\pi|x = \tilde{\pi})$. If $\tilde{\pi} = 3$, we have reached the desired contradiction.
Therefore, $\tilde{\pi} > 3$. But then, it must be the case that $\tilde{x} \leq \tilde{\pi} - 3$. We claim that no type (π, ς) with $\pi < \frac{1}{2}(\tilde{\pi} + \tilde{x})$, $\varsigma \in [0, \bar{\varsigma}]$ chooses $x = \tilde{\pi}$ in equilibrium. That's because choosing $\tilde{x}$ yields higher social image benefit and lower misrepresentation costs. Therefore, $E(\pi|\tilde{\pi}) \geq \frac{1}{2}(\tilde{\pi} + \tilde{x})$.
Consider type $(\tilde{\pi}, \varsigma)$ for which $x^*(\tilde{\pi}, \varsigma) = \tilde{x} < \tilde{\pi}$. It must be the case that, in equilibrium,

$$e\varsigma E_\beta(\pi|\tilde{x}) - c(\tilde{x}, \tilde{\pi}) \geq e\varsigma E_\beta(\pi|\tilde{\pi}) - c(\tilde{\pi}, \tilde{\pi})$$

A necessary condition for the inequality above to be satisfied is

$$e\varsigma E_\beta(\pi|\tilde{x}) - c(\tilde{x}, \tilde{\pi}) \geq e\varsigma \frac{1}{2}(\tilde{\pi} + \tilde{x})$$

$$c(\tilde{x}, \tilde{\pi}) \leq e\varsigma E_\beta(\pi|\tilde{x}) - e\varsigma \frac{1}{2}(\tilde{\pi} + \tilde{x})$$

A necessary condition for the inequality above to be satisfied is

$$c(\tilde{x}, \tilde{\pi}) \leq e\varsigma\tilde{\pi} - e\varsigma \frac{1}{2}(\tilde{\pi} + \tilde{x})$$

$$c(\tilde{x}, \tilde{\pi}) \leq \frac{1}{2} e\varsigma(\tilde{\pi} - \tilde{x})$$

Finally, a necessary condition for the inequality above to be satisfied is

$$c(\tilde{x}, \tilde{\pi}) \leq \frac{1}{2}\bar{\varsigma}(\tilde{\pi} - \tilde{x})$$

Since we assumed $c(x, \pi) > \frac{1}{2}|x - \pi|\bar{\varsigma}$, we reached the desired contradiction and proved the lemma.□
In order to complete the proof of the proposition, we need to show $\exists e^* \in (0, 1)$ s.t. $\forall e < e^*$, no misreporting occurs in equilibrium and $\forall e > e^*$ the equilibrium involves some misreporting.
Let $e^* = \min_{k \in \{1,..,\bar{\pi}\}} \left\{ \frac{1}{\bar{\varsigma}} \frac{c_k}{k} \right\}$. We begin by showing that, if $e > e^*$ there does not exist an equilibrium in which all agents truthfully report their private issue-position.
Assume, aiming towards contradiction, that no misreporting occurs in equilibrium. Let $\tilde{k} \in \arg\min_{k \in \{1,..,\bar{\pi}-1\}} \left\{ \frac{1}{\bar{\varsigma}} \frac{c_k}{k} \right\}$. Then, it must be the case that type $(1, \bar{\varsigma})$ is better off choosing $x = 1$ than $x = \tilde{k} + 1$; i.e.

$$e\bar{\varsigma}E_\beta(\pi|x = 1) - c(1, 1) \geq e\bar{\varsigma}E_\beta\left(\pi|x = \tilde{k} + 1\right) - c\left(\tilde{k} + 1, 1\right)$$

$$e\bar{\varsigma} \geq e\bar{\varsigma}\left(\tilde{k} + 1\right) - c_{\tilde{k}}$$

$$e \leq \frac{1}{\bar{\varsigma}} \frac{c_{\tilde{k}}}{\tilde{k}} = e^*$$

which contradicts the assumption that $e > e^*$.
Let's now show that $\forall e < e^*$ there does not exist an equilibrium in which at least one type of the agent misreports her private issue-position.
We prove the statement by induction. Base case: $\forall e < e^*$, no type (π, ς) with $\pi = 1$ is better off misreporting her private issue-position than truthfully reporting it. We know that, in every equilibrium, $E_\beta(\pi|x = 1) = 1$. $\forall k \in \{1, ..., \bar{\pi} - 1\}$, we want

$$e\varsigma E_\beta(\pi|x = 1) - c(1, 1) \geq e\varsigma E_\beta(\pi|x = 1 + k) - c(1 + k, 1)$$

$$c_k \geq e\varsigma\left[E_\beta(\pi|x = 1 + k) - 1\right]$$

We know that, in any equilibrium, $E_\beta(\pi|x = 1+k) \leq 1+k$. Therefore, a sufficient condition for the inequality above to be satisfied is

$$c_k \geq e\bar{\varsigma}[1+k-1]$$

$$e \leq \frac{1}{\bar{\varsigma}}\frac{c_k}{k} = e^*$$

which we assumed.

Inductive step: if no type (π, ς) with $\pi < \tilde{\pi}$ is better off misreporting her private issue-position than truthfully reporting it, then not type (π, ς) with $\pi = \tilde{\pi}$ is better off misreporting her private issue-position than truthfully reporting it. $\forall k \in \{1, ..., \bar{\pi} - \tilde{\pi}\}$, we want

$$e\varsigma E_\beta(\pi|x = \tilde{\pi}) - c(\tilde{\pi}, \tilde{\pi}) \geq e\varsigma E_\beta(\pi|x = \tilde{\pi}+k) - c(\tilde{\pi}+k, \tilde{\pi})$$

$$c_k \geq e\varsigma[E_\beta(\pi|x = \tilde{\pi}+k) - E_\beta(\pi|x = \tilde{\pi})]$$

By the inductive hypothesis, we know that $E_\beta(\pi|x = \tilde{\pi}) \geq \tilde{\pi}$. Furthermore, we know that, in any equilibrium, $E_\beta(\pi|x = \tilde{\pi}+k) \leq \tilde{\pi}+k$. Therefore, a sufficient condition for the inequality above to be satisfied is

$$c_k \geq e\bar{\varsigma}[\tilde{\pi}+k-\tilde{\pi}]$$

$$e \leq \frac{1}{\bar{\varsigma}}\frac{c_k}{k} = e^*$$

which we assumed.

Therefore, $\forall e < e^*$ there does not exist an equilibrium in which at least one type of the agent misreports her private issue-position. □

A.2 Proof of Proposition 2

First, we show that, in Case 1, no equilibrium in which a positive measure of types misreport their private issue-positions is fully informative about π. We have shown in the proof of Proposition 1 that, under the assumptions of Case 1, $\forall x \in X$, a positive measure of types (π, ς) with $\pi = x$ choose x in equilibrium. Since a positive measure of types misreports their private issue-positions in equilibria, $\exists \tilde{x} \in X$ s.t. $x^*(\pi, \varsigma) = \tilde{x}$ for a positive measure of types (π, ς) with $\pi = \tilde{\pi} \neq x$. But then, upon observing $\tilde{x}$, the audience's posterior belief $\beta|\tilde{x}$ will not be degenerate, which implies the equilibrium is Blackwell less informative about π than any fully informative equilibrium.

Now we show that, in Case 2, there exist equilibria involving misreporting that are fully informative about π.

Consider the following strategy

$$x^*(\pi, \bar{\varsigma}) = \begin{cases} \pi & for\ \pi = 1 \\ \pi + 1 & for\ \pi \in \{2, ..., \bar{\pi} - 1\} \end{cases}$$

and the following set of beliefs: $P(\pi|x = 1) = 1$ for $\pi = 1$ and $P(\pi|x = 1) = 0$ for $\pi \neq 1$; $P(\pi|x = 2) = 1$ for $\pi = 1$ and $P(\pi|x = 2) = 0$ for $\pi \neq 1$; $\forall x > 2$, $P(\pi|x) = 1$ for $\pi = x - 1$ and $P(\pi|x) = 0$ for $\pi \neq x - 1$.

Notice the beliefs are fully informative about π, because there exists a bijective mapping between π and x. Therefore, we only need to show that $\exists e \in [0, 1]$ s.t. the alleged equilibrium above is in fact an equilibrium.

First of all, notice the beliefs are such that $\forall x, x' \in X$, with $x > x'$, $E_\beta(\pi|x) \geq E_\beta(\pi|x')$. Therefore, it is not profitable for any type $(\pi, \bar{\varsigma})$ with $\pi > 1$ to deviate to $x < \pi$. When is it unprofitable for any type $(\pi, \bar{\varsigma})$ with $\pi > 1$ to deviate to $x = \pi$?

$$e\bar{\varsigma}E_\beta(\pi|x = \pi) - c(\pi, \pi) \leq e\bar{\varsigma}E_\beta(\pi|x = \pi + 1) - c(\pi + 1, \pi)$$

$$e\bar{\varsigma}(\pi - 1) \leq e\bar{\varsigma}\pi - c_1$$

$$e \geq \frac{c_1}{\bar{\varsigma}}$$

When is it unprofitable for any type $(\pi, \bar{\varsigma})$ with $\pi > 1$ to deviate to $x = \pi + k$ for $k \in \{2, ..., \bar{\pi} - \pi\}$?

$$e\bar{\varsigma}E_\beta(\pi|x = \pi + k) - c(\pi + k, \pi) \leq e\bar{\varsigma}E_\beta(\pi|x = \pi + 1) - c(\pi + 1, \pi)$$

$$e\bar{\varsigma}(\pi + k - 1) - c_k \leq e\bar{\varsigma}\pi - c_1$$

$$e \leq \frac{c_k - c_1}{\bar{\varsigma}(k-1)}$$

Furthermore, type $(\pi, \bar{\varsigma})$ with $\pi = 1$ does not want to deviate to $\pi = 2$, because she would reap no social image benefits and pay a strictly positive misrepresentation cost. Type $(\pi, \bar{\varsigma})$ with $\pi = 1$ does not want to deviate to $\pi = 1 + k$ for $k \in \{2, ..., \bar{\pi} - 1\}$, if $e \leq \frac{c_k}{\bar{\varsigma}(k-1)}$.
Therefore, a sufficient condition for the existence of equilibria involving misreporting that are fully informative about π is

$$\frac{c_1}{\bar{\varsigma}} \leq \frac{c_k - c_1}{\bar{\varsigma}(k-1)} \quad \forall k \in \{2, ..., \bar{\pi} - 1\}$$

$$c_k \geq kc_1 \quad \forall k \in \{2, ..., \bar{\pi} - 1\}$$

which is true because we assumed the cost schedule is convex. □

B Heterogeneous Interpretation of Natural Language

The model introduced in Section 2 assumes the interpretation of natural language is homogeneous across agents: absent social image concerns, all agents believe that statement $x \in X = \Pi$ has exogenous and commonly-understood meaning "my private issue-position is x". Such assumption, while plausible in many settings, need not always hold. In fact, it is easy to imagine situations in which the interpretation of natural language, in the absence of distortions due to social image concerns, exhibits a degree of heterogeneity across agents. Such heterogeneity would occur, for instance, if different individuals were brought up in environments that differed idiosyncratically in language use and were not fully aware of such differences. In the context of the model from Section 2, heterogeneity in the interpretation of natural language can be captured by assuming heterogeneity in beliefs about the mapping between x and π in the absence of social image concerns. In this section, we show that, if the interpretation of natural language is heterogeneous across agents, the distortions caused by social image concerns may in principle increase the informativeness of equilibrium statements rather than decrease it.

Formally, the setup is virtually the same as in Section 2.1. The only differences are as follows. First, much like in Section 2.4, we create some slack in the message space; specifically, we assume $\beta(\pi) = 0$ for $\pi \in \{\pi_a, ..., \pi_b\}$, where $\pi_a \leq \pi_b$.[53] Second, we assume that the agent's type has an additional dimension, captured by function $g : X \to \Pi$, that describes the way in which the agent interprets natural language in the absence of social image concerns. We assume g is a bijection, though the assumption can in principle be relaxed. We let (π, ς, g) be drawn from cumulative distribution F with full support on $\{\pi_a, ..., \pi_b\}^c \times \Sigma$, where subscript c stands for complement. In

[53] $\{\pi_a, ..., \pi_b\}$ denotes the set of consecutive natural numbers between π_a and π_b.

order to close the model, we assume that each agent believes her own understanding of natural language is shared by all other agents and plays one of the equilibria that would prevail if that was the case. We refer to a situation in which all agents behave this way as a *heterogeneous-natural-language-interpretation equilibrium* (HNLI equilibrium for short).

Before stating the next proposition, it is useful to define a stricter notion of Blackwell informativeness.[54] Specifically, we say that equilibrium $\mathcal{E}$ is strictly more informative about π than equilibrium $\mathcal{E}'$ if the distribution of $\beta|x$ under equilibrium $\mathcal{E}$ is Blackwell more informative than the distribution of $\beta|x$ under equilibrium $\mathcal{E}'$, and if there exists a decision problem in which an agent would be strictly better off observing the distribution of $\beta|x$ under equilibrium $\mathcal{E}$ than under equilibrium $\mathcal{E}'$.

The next proposition shows that there exist HNLI equilibria involving a degree of misreporting that are strictly more informative about π than HNLI equilibria involving no misreporting. For simplicity, the following proposition assumes $\bar{x}$ is even and $\beta(\pi) = 0$, $\forall \pi \leq \frac{\bar{x}}{2}$.

Proposition 3. *The following four statements hold true:*

1. $\forall e \in [0,1]$*, an HNLI equilibrium exists.*
2. $\exists e \in [0,1]$ *that sustains an HNLI equilibrium in which no type misreports her private issue-position (according to her interpretation of natural language). Denote the set of such HNLI equilibria by* $\mathcal{A}$*.*
3. $\exists e \in [0,1]$ *that sustains an HNLI equilibrium in which a positive measure of types misreport their private issue-positions (according to their interpretations of natural language). Denote the set of such HNLI equilibria by* $\mathcal{B}$*.*
4. $\exists$ *type distributions* F *that sustain equilibria* $\mathcal{E} \in \mathcal{B}$ *and* $\mathcal{E}' \in \mathcal{A}$ *such that* $\mathcal{E}$ *is strictly more informative about* π *than* $\mathcal{E}'$*.*

Proof

The first statement in the proposition follows from arguments similar to the ones in Lemma 1 of Proposition 1, with the difference that, in this case, we also have to specify off-the-equilibrium-path beliefs. We assume that, if any off-the-equilibrium-path statement is observed, the statement is attributed to any of the types with $\pi = \frac{\bar{\pi}}{2} + 1$. The second statement can be easily shown by considering $e = 0$. The third statement follows from the arguments in Proposition 1.

We will now prove the fourth statement. We fix the agents' interpretations of natural language as follows: for $\pi = \frac{\bar{\pi}}{2} + 1$, assume $g(x) = x$. For $\pi = \frac{\bar{\pi}}{2} + 1 + k$ with $k \in \left\{1, ..., \frac{\bar{\pi}}{2} - 1\right\}$, assume

$$g(x) = \begin{cases} x + k & if\ x \leq \bar{\pi} - k \\ x + k - \bar{\pi} & if\ x > \bar{\pi} - k \end{cases}$$

[54] The stricter notion is used, for instance, in Rauh et al. (2017).

Consider an HNLI equilibrium in which no type misreports her private issue-position (according to her interpretation of natural language). Denote this equilibrium by $\mathcal{E}'$. Notice the equilibrium is completely uninformative about π, because effectively all types $\pi \in \left\{\frac{\bar{\pi}}{2}+1, ..., \bar{\pi}\right\}$ report $x = \frac{\bar{\pi}}{2}+1$. Now consider an HNLI equilibrium in which a positive measure of types misreport their private issue-positions (according to their interpretations of natural language) and assume the off-the-equilibrium-path beliefs are such as to attribute all off-the-equilibrium-path statements to any of the types with $\pi = \frac{\bar{\pi}}{2}+1$. Denote this equilibrium by $\mathcal{E}$. Since $\mathcal{E}'$ is does not reveal any additional information about π beyond the prior, $\mathcal{E}$ is trivially more informative about π than $\mathcal{E}'$. Furthermore, notice that, in equilibrium $\mathcal{E}$, no type with $\pi = \bar{\pi}$ misreports her private issue-position (according to her interpretation of natural language). But then, if a positive measure of types misreport their private issue-positions (according to their interpretations of natural language), $x = \frac{\bar{\pi}}{2}+1$ necessarily becomes a relatively more informative signal that $\pi = \bar{\pi}$. Considering a decision problem involving a bet on whether the agent is of type $\pi = \bar{\pi}$ or type $\pi \neq \bar{\pi}$ shows that $\mathcal{E}$ is strictly more informative about π than $\mathcal{E}'$. □

C Exploratory Survey

Subjects were recruited from the Experimental and Behavioral Economics Laboratory (EBEL) portal at the University of California Santa Barbara (UCSB) to complete a short online survey. The recruitment email did not mention the topic of the study. Before being directed to the consent form, subjects were asked to complete a brief pre-screen questionnaire identical to the one described in Section 4.

After consenting to participate in the study, subjects were asked to rate the social acceptability of agreeing or disagreeing with each of the 20 statements from Table A2, 15 of which we hypothesized would engage the students' social image concerns and 5 of which we hypothesized would do so to a lesser extent or not at all. The statements were adapted from a variety of sources, ranging from newspaper and magazine articles to think-tank reports. Participants rated the social acceptability of agreeing or disagreeing with each statement on a slider from -5 to 5, where -5 indicated "among UCSB students, disagreeing with the statement is a lot more socially acceptable than agreeing with the statement", 5 indicated "among UCSB students, agreeing with the statement is a lot more socially acceptable than disagreeing with the statement", and 0 indicated "among UCSB students, agreeing and disagreeing with the statement are about the same in terms of social acceptability". The order of the statements was randomized at the participant level.

Upon completing the survey, subjects received a $5 electronic gift-card for their participation.

Table A1 shows the sample size for the Exploratory Survey.

For each statement, we categorized every answer strictly above 0 on the -5 to 5 scale as supporting the idea that "agreeing with the statement is more socially acceptable at UCSB than disagreeing with it", every answer strictly below 0 as supporting the idea that "disagreeing with the statement is more socially acceptable at UCSB than agreeing with it", and every answer equal to 0 as supporting the idea that, at UCSB, "agreeing and disagreeing with the statement are about the same in terms of social acceptability". Table A3 shows, for each statement, the proportion of

subjects in each of the three categories.

In order to establish, for each statement, whether agreeing or disagreeing with it is considered more socially acceptable at UCSB, we simply compared the fraction of subjects who answered a strictly positive number to the fraction of subjects who answered a strictly negative number. The larger of the two fractions determined, for each statement, whether the *perceived social-acceptability direction* was to agree or to disagree with the statement.

We then proceeded to select the 10 sensitive statements and the 5 placebo statements to show participants in the Social Image Experiment. The sensitive statements were selected as follows: first, each of the 20 statements was ranked in descending order by the proportion of subjects whose answers aligned with the *perceived social-acceptability direction*. For instance, if agreeing was designated as the perceived social acceptability direction for a certain statement, the proportion of subjects whose answers align with the *perceived social-acceptability direction* is simply the proportion of subjects who answered a number strictly greater than zero to the question. Second, the top 10 statements based on that order were selected.[55]

The placebo statements were selected as follows: first, each of the 20 statements in the Exploratory Survey was ranked in descending order by the proportion of students who answered that, at UCSB, "agreeing and disagreeing with the statement are about the same in terms of social acceptability".[56] Second, the top 5 statements based on that order were selected.

Table A1: **Sample Size Exploratory Survey**

	Sample size
Completed Pre-screen	$N = 109$
Met Eligibility Criteria	$N = 76$
Passed Low Quality Screener	$N = 72$
Consented to Participate	$N = 72$
Completed Survey	$N = 70$

Notes: The table above presents the size of the samples at different stages of the Exploratory Survey.

[55]The selection criterion described above does not make use of the intensity of the subjects' answers on the -5 to 5 issue-sensitivity scale. An alternative selection criterion for the sensitive statements is to rank the 20 statements by the absolute value of the average of the students' answers. The latter selection criterion rewards both the degree of agreement about the *perceived social acceptability direction* and the perceived intensity of the social desirability concerns vis-à-vis the statement. The two criteria select the same set of statements.

[56]A student's answer was classified as supporting the position that, at UCSB, "agreeing and disagreeing with the statement are about the same in terms of social acceptability" if the student answered a zero on the -5 to 5 scale.

Table A2: **List of All Statements in the Exploratory Survey**

Statement	
A	All statues and memorials of Confederate leaders should be removed.†
B	Adopting elements of other cultures, whether more or less dominant, is a perfectly acceptable practice.†
C	The UCSB administration should require professors to address students according to the students' preferred gender pronouns.†
D	The Islamic religion is more likely than other religions to encourage violence among its believers.†
E	The UCSB administration should require professors to use trigger warnings in their classes.†
F	Sexual harassment training should be mandatory for everybody who works or studies at UCSB.†
G	If Proposition 209 was repealed, universities in the UC system should adopt extensive affirmative action policies that explicitly take into account race in the admission process.
H	People who immigrated to the U.S. illegally, when caught, should be deported and sent back to their countries of origin.†
I	If some student group at UCSB invited Charles Murray on campus to give a talk, it would be perfectly acceptable for other students to disrupt the talk and prevent Murray from delivering his lecture.
J	As a result of different evolutionary pressures, men are naturally more promiscuous than women.
K	The U.S. government should provide reparations for slavery.†
L	At UCSB, students are often too quick to call others sexist or racist.
M	Racial microaggressions are an important problem at UCSB.†
N	The UCSB administration should allow students to wear blackface for Halloween.†
O	People of Color cannot be racist against white people.
P	Parents should limit the amount of time their kids spend on their smartphones.†
Q	The United States should increase tariffs on foreign imports.†
R	School uniforms help reduce clothing-related peer pressure.†
S	The one-cent coin (i.e. the penny) should be removed from circulation.†
T	The members states of the European Union should cede more powers to the E.U.†

Notes: the table above presents the 20 statements shown to participants in the Exploratory Survey. Some of the statements were preceded by a brief paragraph giving additional information. For instance, one such paragraph explained that "trigger warnings" are warnings that some work contains writing, images, or concepts that may be distressing to some people. See the survey instrument for additional details. A statement is marked with a dagger (†) if it was selected to be included in the Social Image Experiment.

Table A3: **Perceptions of the Socially Acceptable Position at UCSB**

	(1) Agreeing more socially acceptable than disagreeing	(2) Disagreeing more socially acceptable than agreeing	(3) Agreeing and disagreeing about the same
Statement A	0.79	0.13	0.09
Statement B	0.21	0.73	0.06
Statement C	0.91	0.06	0.03
Statement D	0.10	0.89	0.01
Statement E	0.77	0.16	0.07
Statement F	0.94	0.04	0.01
Statement G	0.50	0.40	0.10
Statement H	0.04	0.96	0.00
Statement I	0.66	0.29	0.06
Statement J	0.30	0.63	0.07
Statement K	0.81	0.10	0.09
Statement L	0.17	0.70	0.13
Statement M	0.77	0.19	0.04
Statement N	0.04	0.94	0.01
Statement O	0.44	0.49	0.07
Statement P	0.59	0.17	0.24
Statement Q	0.17	0.49	0.34
Statement R	0.21	0.53	0.26
Statement S	0.41	0.17	0.41
Statement T	0.23	0.21	0.56

Notes: The table above presents, for each of the statements in the Exploratory Survey , the fraction of participants who answered that agreeing with the statement is more socially acceptable at UCSB than disagreeing with the statement, the fraction of participants who answered the opposite, and the fraction of participants who answered that, at UCSB, agreeing and disagreeing with the statement are about the same in terms of social acceptability. Answering that disagreeing with a statement is more socially acceptable at UCSB than agreeing with the statement is defined as selecting a number strictly smaller than 0 on the -5 to 5 scale; conversely, answering that agreeing with a statement is more socially acceptable at UCSB than disagreeing with the statement is defined as selecting a number strictly greater than 0 on the -5 to 5 scale. Finally, answering that, at UCSB, agreeing and disagreeing with a statement are about the same in terms of social acceptability is defined as selecting 0 on the -5 to 5 scale.

D Additional Empirical Results: Social Image Experiment

D.1 Descriptive Statistics

Table A4: **Selected Universities in Top Quintile of Liberal-Conservative Ranking**

Boston College	Tufts University
Brandeis University	UC Berkeley
Brown University	UC Davis
Carnegie Mellon University	UC Los Angeles
Columbia University	UC San Diego
Duke University	UC Santa Cruz
Georgetown University	University of Chicago
Harvard University	University of Michigan - Ann Arbor
John's Hopkins University	University of Pennsylvania
New York University	University of Southern California
Northeastern University	Washington University in St. Louis
Northwestern University	Wellesley College
Stanford University	Wesleyan University

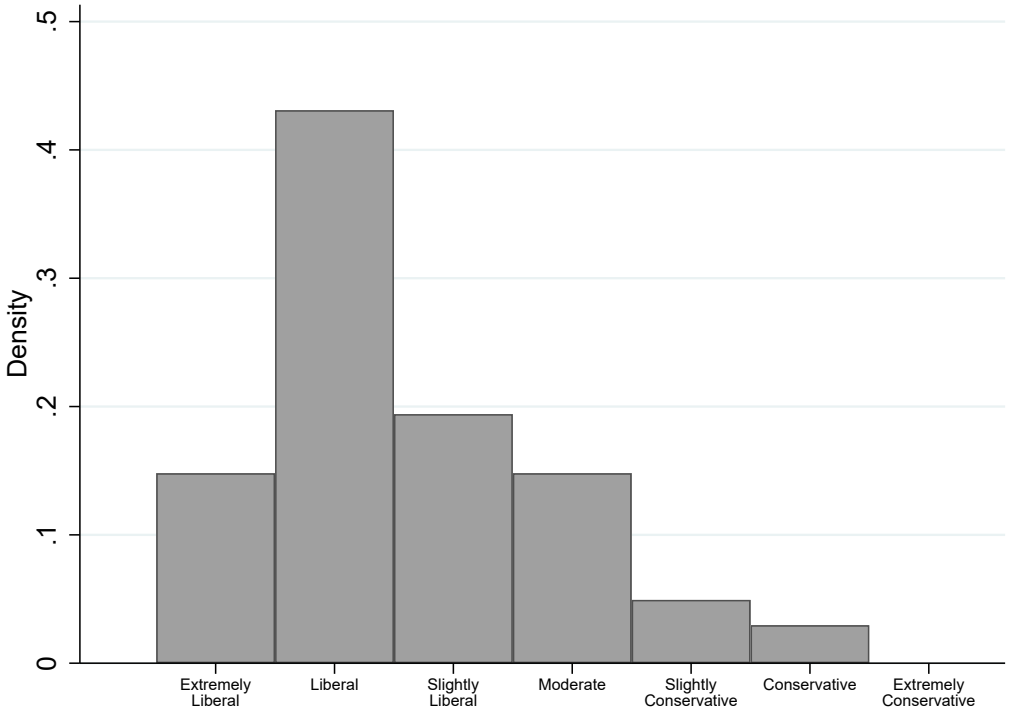
Notes: The table above presents, in alphabetical order, a non-exhaustive list of universities that fall in the top quintile of the Niche ranking of universities from most liberal to most conservative. The ranking is calculated by surveying a sample of students from each college and asking them both about their personal political leaning and about their beliefs about the political leanings of the other students at their college (Niche, 2020).

Table A5: **Sample Demographics**

	(1) Main sample	(2) UCSB population
Male	0.38	0.45
White	0.43	0.36
Junior/Senior	0.65	0.61
Science/Engineering Major	0.48	0.52

Notes: Column 1 presents the average demographics of participants in the Social Image Experiment. Column 2 presents the average demographics of UCSB undergraduate students as per the 2018-2019 Campus Profile report (UCSB, 2019).

Figure A1: **Histogram of Self-Reported Political Ideology**



Notes: The figure above shows a histogram of the self-reported political ideology of participants in the Social Image Experiment. The question about political ideology was asked before subjects were randomized into the Private and the Public Treatments. The wording of the question, taken from the American National Election Study (ANES, 2016), was: "We hear a lot of talk these days about liberals and conservatives. Here is a scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale?" The possible answers are shown on the x-axis of the figure.

Table A6: **Descriptive Statistics: Social Image Experiment**

	Mean	Standard deviation	Minimum value	Maximum value
Confederate Statues	6.22	2.97	0	10
Cultural Appropriation*	4.88	2.67	0	10
Preferred Gender Pronouns	7.33	3.11	0	10
Islam and Violence*	2.48	2.89	0	10
Trigger Warnings	5.90	3.01	0	10
Sexual Harassment Training	7.85	2.57	0	10
Illegal Immigration*	3.53	2.82	0	10
Reparations for Slavery	5.70	2.98	0	10
Racial Microaggressions	5.54	2.78	0	10
Blackface Halloween*	1.53	2.38	0	10
Parents Smartphones	7.03	2.00	0	10
Import Tariffs*	3.68	2.05	0	8
School Uniforms*	4.79	2.95	0	10
Penny	6.03	3.26	0	10
European Union	4.74	1.61	0	10

Notes: The table above presents the means, standard deviations, minimum values, and maximum values of the levels of agreement of participants in the Private Treatment of the Social Image Experiment with the 10 sensitive and 5 placebo statements. The statistics are reported in original units; therefore, larger numbers correspond to higher levels of agreement. A statement is marked with an asterisk (*) if the fraction of participants in the Exploratory Survey who answered that disagreeing with the statement is more socially acceptable at UCSB than agreeing with the statement is larger than the fraction of participants who answered the opposite.

Table A7: **Average Private Level of Agreement by Political Ideology**

	(1) Mean Moderate/Conservative	(2) Mean Liberal
Confederate Statues	4.60	7.35
Cultural Appropriation*	5.65	4.35
Preferred Gender Pronouns	5.85	8.36
Islam and Violence*	3.42	1.83
Trigger Warnings	4.84	6.64
Sexual Harassment Training	6.74	8.62
Illegal Immigration*	5.24	2.34
Reparations for Slavery	3.76	7.06
Racial Microaggressions	4.03	6.58
Blackface Halloween*	2.45	0.89

Notes: The table above reports, for participants assigned to the Private Treatment of the Social Image Experiment, the average level of agreement with each sensitive statement of students who were classified as liberals and students who were classified as moderates/conservatives according to the median split described in Section 4.3. The means are reported in original units; therefore, larger numbers correspond to higher levels of agreement. A statement is marked with an asterisk (*) if, according to the answers of students in the Exploratory Survey, disagreeing with the statement is generally considered to be more socially acceptable at UCSB than agreeing with it.

D.2 Balance and Attrition

Table A8: **Balance**

Variable	(1) Public Treatment Mean/SD	(2) Private Treatment Mean/SD	T-test P-value (1)-(2)
Female	0.63	0.58	0.36
	(0.49)	(0.50)	
White	0.41	0.46	0.36
	(0.49)	(0.50)	
Age	20.39	20.41	0.88
	(1.53)	(1.46)	
Junior or Senior	0.65	0.66	0.78
	(0.48)	(0.47)	
Liberal	0.57	0.59	0.71
	(0.50)	(0.49)	
Humanities/Social Science Major	0.44	0.52	0.17
	(0.50)	(0.50)	
N	153	151	
F-test of joint significance (p-value)			0.70
F-test, number of observations			304

Notes: Columns 1 and 2 present demographics for the Private and Public Treatments of the Social Image Experiment. Column 3 presents p-values of tests of differences in means between the two groups.

Table A9: **Attrition**

Variable	(1) Private Treatment Mean/SD	(2) Public Treatment Mean/SD	T-test P-value (1)-(2)
Completed survey	0.89	0.87	0.56
	(0.31)	(0.33)	
N	178	182	

Notes: Columns 1 and 2 present survey completion rates for individuals who were randomized into the Private and Public Treatments of the Social Image Experiment. Column 3 presents p-values of tests of differences in completion rates between the two groups.

Table A10: **Balance Check: Target Outcome Variables**

	Self-identifying as Liberal		Donation to AAUW		Signed Petition	
	(1)	(2)	(3)	(4)	(5)	(6)
Public Treatment	-0.02	0.00	0.57	0.13	0.06	0.07
	(0.06)	(0.00)	(3.47)	(3.66)	(0.05)	(0.05)
Controls	No	Yes	No	Yes	No	Yes
Observations	304	304	304	304	303	303
Dependent variable mean	0.58	0.58	33.53	33.53	0.26	0.26

Notes: The table above presents average treatment effects of being assigned to the Public Treatment using Equation 1. Depending on the specification, controls may be excluded. The dependent variables are: i) an indicator for self-identifying as liberal according to the median split described in Section 4.3; ii) the amount the student donated to the American Association of University Women (AAUW) in the dictator game; iii) an indicator for whether the student supported the anonymous petition to require that UCSB mandate yearly sexual harassment training for everybody who works or studies at UCSB. Standard errors are in parentheses.

D.3 Additional Empirical Results: Average Treatment Effects from the Social Image Experiment

Table A11: **Average Treatment Effects: Indices of Sensitive & Placebo Attitudes**

	Index Sensitive Attitudes		Index Placebo Attitudes	
	(1)	(2)	(3)	(4)
Public Treatment	0.19*	0.24***	-0.02	0.03
	(0.11)	(0.09)	(0.11)	(0.11)
Controls	No	Yes	No	Yes
Observations	304	304	304	304

Notes: The table above presents average treatment effects of being assigned to the Public Treatment using Equation 1. Depending on the specification, controls may be excluded. The dependent variables are the normalized indices of the level of agreement with the sensitive and placebo statements. The index of sensitive (placebo) attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive (placebo) questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. The normalization is achieved by first subtracting from an index the average value of the index among participants in the Private Treatment and then dividing the result by the standard deviation of the index among participants in the Private Treatment. Standard errors are in parentheses.

Table A12: **Average Treatment Effects: Statement by Statement**

	Treatment effect (original units)	Standard error (original units)	Treatment effect (SD units)	Standard error (SD units)	p-value	Sharpened FDR-adjusted q-value
Confederate Statues	0.50	0.31	0.17	0.10	0.10	0.12
Cultural Appropriation	0.26	0.29	0.10	0.11	0.37	0.21
Preferred Gender Pronouns	0.42	0.32	0.13	0.10	0.20	0.14
Islam and Violence	-0.03	0.31	-0.01	0.11	0.92	0.33
Trigger Warnings	0.58	0.30	0.19	0.10	0.06	0.11
Sexual Harassment Training	0.53	0.26	0.21	0.10	0.04	0.11
Illegal Immigration	0.40	0.27	0.14	0.10	0.15	0.14
Reparations for Slavery	0.61	0.30	0.20	0.10	0.04	0.11
Racial Microaggressions	0.86	0.28	0.31	0.10	0.00	0.03
Blackface Halloween	0.59	0.25	0.25	0.11	0.02	0.10

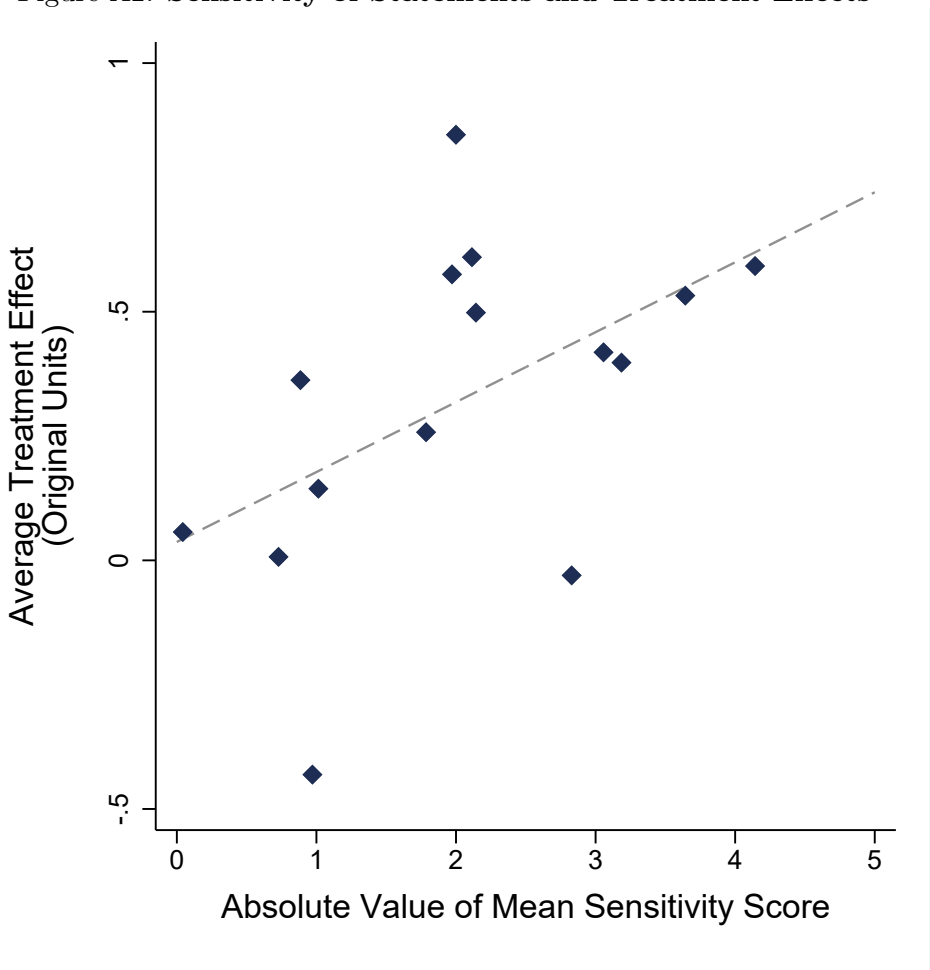
Notes: The table above presents average treatment effects of being assigned to the Public Treatment using Equation 1. The students' reported levels of agreement with the sensitive statements are oriented in such a way that larger numbers always correspond to views that, according to the Exploratory Survey, are generally perceived to be more socially acceptable at UCSB. Column 1 and Column 2 present the effects and standard errors in original units. Columns 3 and 4 present the effects and standard errors in standard deviation units, where outcomes are normalized so that the distribution of answers of participants in the Private Treatment has a standard deviation of one and a mean of zero. Columns 5 and 6 present the unadjusted p-value and sharpened False Discovery Rate-adjusted two-stage q-value, respectively.

Table A13: **Most Common Trigrams in the Open Responses**

Trigrams used relatively more often in the Private Treatment		Trigrams used relatively more often in the Public Treatment	
Trigram	Likelihood Ratio (Private:Public)	Trigram	Likelihood Ratio (Private:Public)
Freedom of speech	5:1	*The talk would*	1:5
Not be allowed	5:1	*For people to*	1:4
Also have the	4:1	*For them to*	1:4
Students should have	4:1	*Perfectly acceptable to*	1:4
Able to voice	3:1	*Talk would be*	1:4
Believe it is	3:1	*To give his*	1:4
Disagree with his	3:1	*A place for*	1:3
He has the	3:1	*Believe it would*	1:3
Right to say	3:1	*Has to say*	1:3
Speech is important	3:1	*He has to*	1:3

Notes: The Social Image Experiment included the following question with an open response text box: "Charles Murray is a political scientist who co-authored a controversial book that, among other things, discusses racial differences in intelligence and who wrote in a 2005 essay that 'no woman has been a significant original thinker in any of the world's great philosophical traditions'. If some student group at UCSB invited Charles Murray on campus to give a talk, would it be acceptable for other students to disrupt the talk and prevent Murray from delivering his lecture or would it not be acceptable?". The table above shows the trigrams that, according to a naive Bayes classifier, are most diagnostic of being in the Private Treatment compared to the Public Treatment and the trigrams that are most diagnostic of being in the Public Treatment compared to the Private Treatment.

Figure A2: **Sensitivity of Statements and Treatment Effects**



Notes: The figure above presents a scatter plot of the absolute value of the mean answer of participants in the Exploratory Survey (x-axis) against average treatment effects in the Social Image Experiment estimated using Equation 1 (y-axis). Each diamond represents one of the 15 statements shown to participants in the Social Image Experiment. The dashed line in the figure represents the line of best fit. When calculating the average treatment effects, the students' reported levels of agreement with the statements are oriented in such a way that larger numbers always correspond to views that, according to the Exploratory Survey, are generally perceived to be more socially acceptable at UCSB. The absolute value of the mean answer of participants in the Exploratory Survey is a measure of the extent to which the students who took the survey perceived a statement as being sensitive. Such sensitivity criterion rewards both the degree of agreement about the *perceived social acceptability direction* and the perceived intensity of the social desirability concerns vis-à-vis the statement. Defining sensitivity as the mean of the absolute value of answers of participants in the Exploratory Survey rather than the absolute value of the mean produces a similar figure.

D.4 Robustness Checks: Average Treatment Effects from the Social Image Experiment

Table A14: **Robustness to Excluding One Statement at a Time from Index of Sensitive Attitudes**

	Treatment effect	Standard error	p-value
Excluding Confederate Statues	0.24	0.09	0.01
Excluding Cultural Appropriation	0.25	0.08	0.00
Excluding Preferred Gender Pronouns	0.25	0.09	0.00
Excluding Islam and Violence	0.27	0.08	0.00
Excluding Trigger Warnings	0.24	0.09	0.01
Excluding Sexual Harassment Training	0.23	0.09	0.01
Excluding Illegal Immigration	0.25	0.09	0.01
Excluding Reparations for Slavery	0.24	0.09	0.01
Excluding Racial Microaggressions	0.22	0.09	0.01
Excluding Blackface Halloween	0.23	0.09	0.01

Notes: The figure above presents average treatment effects of being assigned to the Public Treatment using Equation 1. The dependent variable is always a version of the index of sensitive attitudes. Each row omits the statement listed from the index. Standard errors are in parentheses.

Table A15: **Ordered Probit on Indices of Sensitive and Placebo Attitudes**

	Index Sensitive Attitudes		Index Placebo Attitudes	
	(1)	(2)	(3)	(4)
Public Treatment	0.19	0.38***	-0.07	-0.01
	(0.12)	(0.13)	(0.12)	(0.13)
Controls	No	Yes	No	Yes
Observations	304	304	304	304

Notes: The table above presents the results of an ordered probit model of the index of sensitive or placebo attitudes, rounded to the nearest digit, on a treatment indicator and, depending on the specification, pre-specified controls. The index of sensitive (placebo) attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive (placebo) questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. Standard errors are in parentheses.

D.5 Robustness Checks: Heterogeneous Treatment Effects from the Social Image Experiment

Table A16: **Robustness to the Definition of Liberal**

	(1)	(2)
Public Treatment × Moderate/Conservative (def.1)	0.40**	
	(0.19)	
Public Treatment × Moderate/Conservative (def.2)		0.50**
		(0.24)
Observations	304	304

Notes: The table above presents the coefficient on the interaction term in the following regression equation $Y_i = \alpha + \beta T_i + \delta M_i + \gamma T_i \times M_i + \varepsilon_i$, where Y_i denotes the standardized index of sensitive attitudes,T_i is an indicator for whether participant i is assigned to the Public Treatment, M_i is an indicator for whether participant i is classified as a moderate/conservative, and ε_i is an idiosyncratic error term. The index of sensitive attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. The normalization is achieved by first subtracting from an index the average value of the index among participants in the Private Treatment and then dividing the result by the standard deviation of the index among participants in the Private Treatment. The first row in the table defines liberals and moderates/conservatives according to the median split described in Section 4.3. The second row in the table extends the definition of liberals to include subjects who, on the 7-point political ideology spectrum, self-identify as "slightly liberal". Standard errors are in parentheses.

Table A17: **Interacted Ordered Probit**

	Marginal effect	Standard error	p-value
Confederate Statues	0.08	0.07	0.27
Cultural Appropriation	-0.00	0.09	0.98
Preferred Gender Pronouns	0.10	0.08	0.23
Islam and Violence	0.04	0.09	0.68
Trigger Warnings	0.01	0.09	0.94
Sexual Harassment Training	0.16	0.08	0.06
Illegal Immigration	0.12	0.07	0.09
Reparations for Slavery	0.12	0.07	0.09
Racial Microaggressions	0.15	0.08	0.05
Blackface Halloween	0.12	0.10	0.23

Notes: The table above presents the results of an ordered probit model of the index of sensitive or placebo attitudes, rounded to the nearest digit, on a treatment indicator interacted with an indicator for being classified as a moderate/conservative according to the median split described in Section 4.3. The index of sensitive (placebo) attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive (placebo) questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. The marginal effects are calculated as described in Ai and Nortion (2003).

D.5.1 Ceiling Effects

One can imagine a world in which the treatment effects on liberals and conservatives would in principle be identical, but in which ceiling effects mechanically constrain the treatment effects on liberals and not on conservatives. Suppose it was indeed the case that the heterogeneous treatment effects on self-reported political ideology were entirely driven by ceiling effects. Consider the following regression equation

$$Y_{i,j} = \alpha_j + \beta_j T_i + \delta_j M_i + \gamma_j T_i \times M_i + \varepsilon_{i,j} \tag{2}$$

where $Y_{i,j}$ denotes participant i's reported level of agreement with sensitive statement j, T_i is an indicator for whether participant i is assigned to the Public Treatment, M_i is an indicator for whether participant i is classified as a moderate/conservative according to the median split described in Section 4.3, and $\varepsilon_{i,j}$ is an idiosyncratic error term. Let the $Y_{i,j}$ be oriented in such a way that larger numbers always correspond to views that are perceived to be more socially acceptable at UCSB.

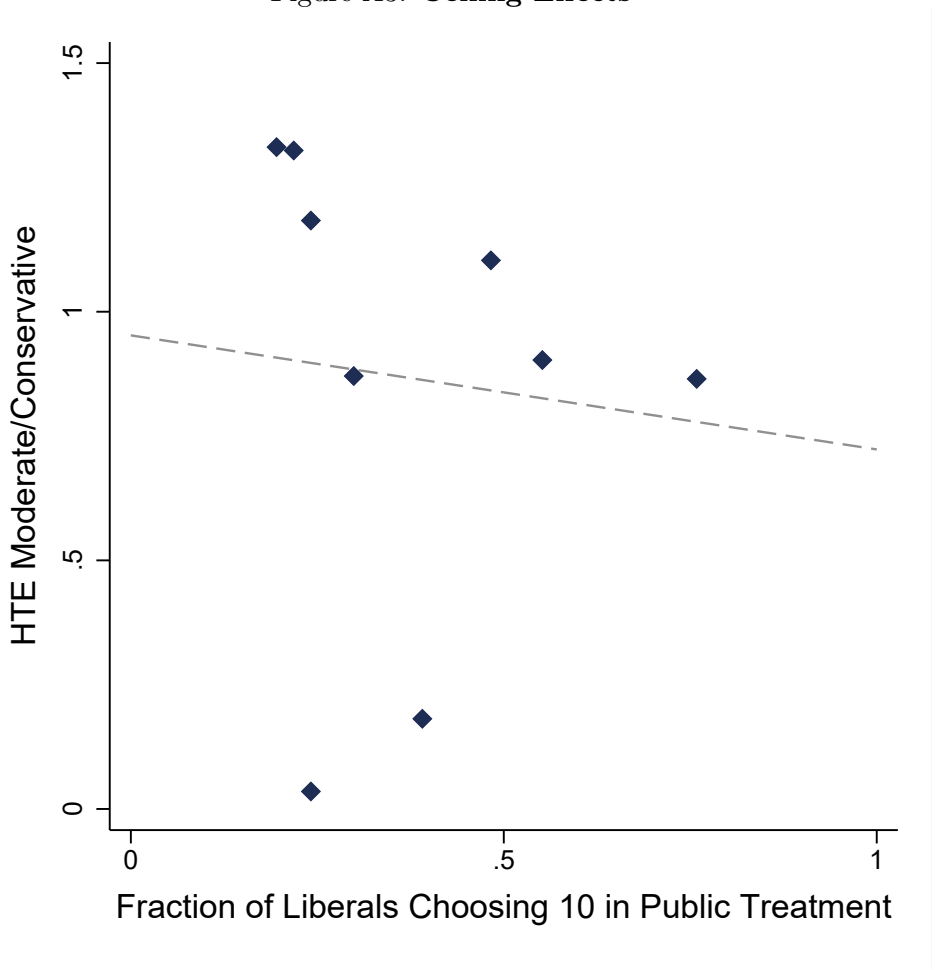
Consider all the sensitive statements for which γ_j in the equation above is estimated to be positive. The statements for which γ_j is estimated to be positive are the ones for which the ceiling on the 0-10 Likert scale may mechanically constrain the treatment effects on participants classified as liberals.[57] If the heterogeneous treatment effects on self-reported political ideology were entirely driven by ceiling effects, we would expect to see a positive relationship, for statements with $\gamma_j > 0$, between the size of the heterogeneous treatment effect (γ_j) and the fraction of participants in the Public Treatment classified as liberal who bunch at the socially-acceptable end of the Likert scale. The intuition is that a larger fraction of participants in the Public Treatment classified as liberal bunching at the socially-acceptable end of the Likert scale scale corresponds to a larger fraction of participants who are mechanically constrained due to ceiling effects.

Figure A3 presents a scatter plot of such relationship. As shown in Figure A3, the line of best fit is weakly negative; therefore, the heterogeneous treatment effects on self-reported political ideology are unlikely to be solely driven by ceiling effects.

The results of the interacted ordered probit model from Table A17 lend additional support to the argument that the treatment effect heterogeneity in terms of self-reported political ideology is not driven by ceiling effects.

[57] Notice once again that the end of the agreeing-disagreeing spectrum that, according to the answers of students in the Exploratory Survey, is generally considered to be more socially acceptable at UCSB always coincides with the end of the spectrum that is closer to the average position of students in the Private Treatment of the Social Image Experiment who are classified as liberals.

Figure A3: **Ceiling Effects**



Notes: The figure above presents a scatter plot of the fraction of participants classified as liberal in the Public Treatment of the Social Image Experiment who bunch at the socially-acceptable end of the Likert scale (which, after the statements are re-oriented in such a way that larger numbers correspond to views that are generally perceived to be more socially acceptable at UCSB, always equals 10) against the size of the γ coefficient in Equation 2. The diamonds represent the nine sensitive statements for which $\gamma_j > 0$. The dashed line represents the line of best fit.

E Additional Empirical Results: Information Loss

This appendix presents additional results about information loss in the Social Image Experiment. The first two sections report the analysis proposed in the pre-analysis plan.

E.1 Performance of Binary Classifiers

A simple implication of the theoretical framework from Section 2 is that binary classifiers should achieve lower expected loss, independently of how the loss function is defined, when based on the answers of participants in the Private Treatment than when based on the answers of participants in the Public Treatment. In this section, we construct: i) a binary classifier for whether the participant self-identified as a liberal based on the participant's index of sensitive attitudes, ii) a binary classifier for whether the participant made a positive donation to the AAUW based on the participant's reported level of agreement with the statement about sexual harassment, and, iii) a binary classifier for whether the participant supported the anonymous petition based on the participant's reported level of agreement with the statement about sexual harassment. In what follows, we refer to the dichotomous variable that the classifier is trying to forecast as the *binary target outcome variable* and the answer or index used to predict it as the *predictor*.

The construction of the classifiers is as follows: first, we specify a logit model relating the *binary target outcome variable* to a high-dimensional polynomial of the *predictor*. Second, we estimate the model separately for participants in the Private Treatment and participants in the Public Treatment. Third, we use the estimated models to generate, for each participant, the predicted probability of belonging to one of the two classes of the *binary target outcome variable* given the participant's value of the *predictor*. Importantly, the predicted probabilities for participants in the Private Treatment are calculated using the model estimated on the data from the Private Treatment, and the predicted probabilities for participants in the Public Treatment are calculated using the model estimated on the data from the Public Treatment. Fourth, we pick some cutoff value p^* and classify each participant as belonging to one of the two classes if, for that participant, the predicted probability of belonging to the class is above p^*. Conversely, if, for that participant, the predicted probability of belonging to the class is weakly below p^*, we classify the participant as belonging to the other class.

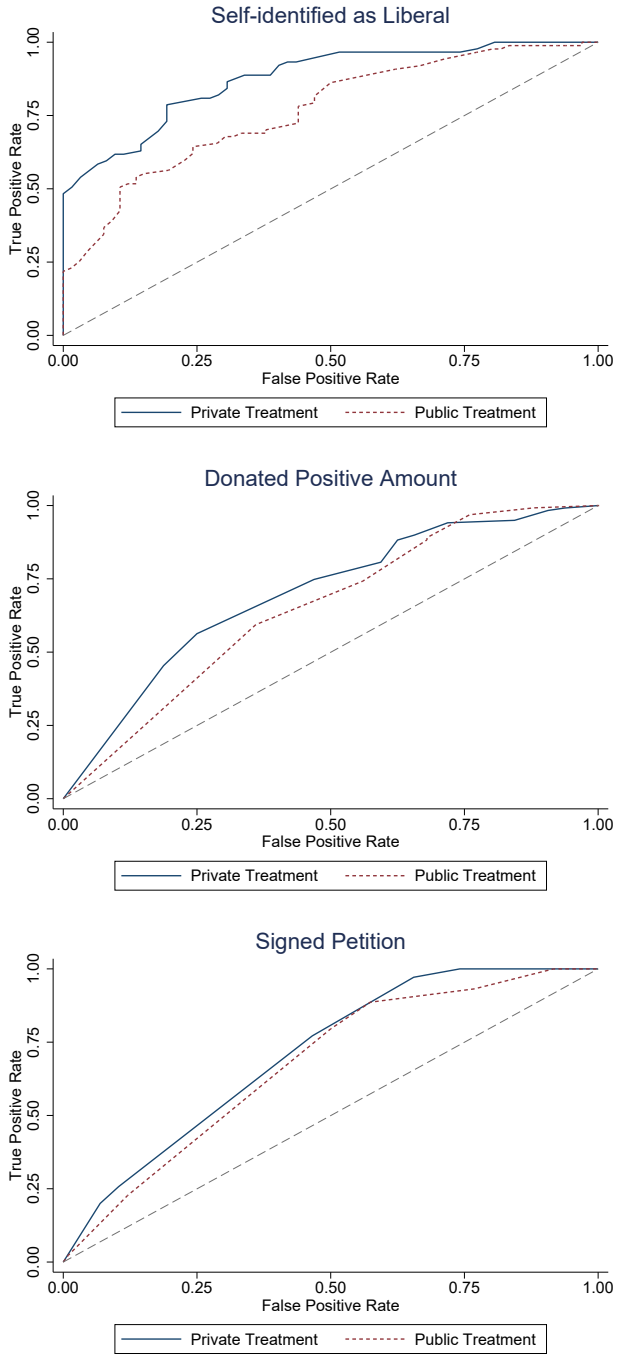
Given a particular loss function, one would choose p^* optimally to trade-off the differential penalties from *false positives* and *false negatives*. Specifically, if false positives were a lot more costly than false negatives, one would want to be quite sure the observation is a positive before classifying that observation as a positive; in other words, one would optimally choose a high p^*. Conversely, if false negatives were a lot more costly than false positives, one would optimally choose a low p^*. Imagine we had constructed two classifiers and were given a loss function. For a fixed loss function, a sufficient condition for one classifier to achieve a lower expected loss than the other is that the first classifier have a smaller false positive and false negative rate than the other. If we can show that, $\forall p^* \in [0, 1]$, the rate of false positives and false negatives from one classifier is smaller than for the other, we will have shown that the former classifier achieves a lower expected loss than the latter classifier independently of how the loss function is defined. The Receiver Operating Characteristic (ROC) curve analysis below shows that the classifier constructed using data from

the Private Treatment has a lower rate of false positives and false negatives than the classifier constructed using data from the Public Treatment $\forall p^* \in [0, 1]$.[58]

Figure A4 shows the Receiver Operating Characteristic (ROC) curves for the binary classifiers constructed according to the description above. The figure corroborates the hypothesis that, for our *binary target outcome variables*, the binary classifiers constructed using the answers of participants in the Private Treatment perform better than the binary classifiers constructed using the answers of participants in the Public Treatment. Specifically, for a fixed false positive rate, the binary classifiers constructed using the answers of participants in the Private Treatment achieves virtually always a higher true positive rate than the binary classifiers constructed using the answers of participants in the Public Treatment.

[58] Letting p^* vary between 0 and 1 generates a continuum of classifiers, which, when joined together, trace a curve in a two-dimensional space where the abscissa is the false positive rate and the ordinate is the true positive rate. That curve is known as the Receiver Operating Characteristic (ROC) curve.

Figure A4: **Receiver-Operating-Characteristic Curves**



Notes: The figure presents the Receiver-Operating-Characteristic curves of the binary classifiers constructed using the answers of participants in the Private and Public Treatments, for all the binary target outcome variables. The first panel shows the ROC curves of the binary classifiers for whether a participant self-identified as a liberal. The second panel shows the ROC curves of the binary classifiers for whether a participant made a positive donation to the AAUW. The third panel shows the ROC curves of the binary classifiers for whether a participant supported the anonymous petition to require that the UCSB administration mandate yearly sexual harassment training for everybody who works or studies at UCSB.

E.2 Regression Residuals

As an additional way to study whether the answers of participants in the Private Treatment are more informative than the answers of participants in the Public Treatment, we run the following regression, separately for participants in the Private Treatment and for participants in the Public treatment:

$$Y_i = \boldsymbol{\beta Z_i} + \varepsilon_i \tag{3}$$

where Y_i denotes individual i's donation to the AAUW, ε_i is an idiosyncratic error term, and $\boldsymbol{Z_i}$ is a vector of 11 indicator variables $Z_{i,j}$ $j \in \{0, ..., 10\}$, where $Z_{i,j}$ takes value 1 if individual i's level of agreement with the sexual harassment statement is j and 0 otherwise. We then compare the mean absolute deviations in the Private and Public Treatments to study whether the private answers to the sensitive statements about sexual harassment are better predictors of donation rates to the AAUW than the public answers.

Table A18 shows that the mean absolute deviations from Equation 3 are marginally larger in the Public Treatment than in the Private Treatment, but the results are quite noisy and insignificant.

Table A18: **Absolute Deviations**

	Absolute Deviation (1)
Public Treatment	0.02
	(0.06)
N	304
Dependent variable mean	0.77

Notes: The table above presents the results of a regression of the absolute value of the residuals from Equation 3 on a treatment indicator. Standard errors are in parentheses.

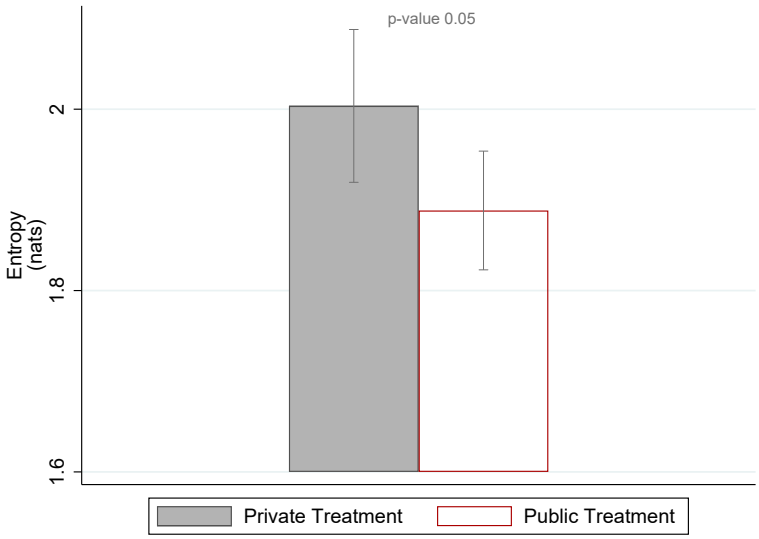
E.3 Shannon Entropy

The theoretical model from Section 2 does not make a general prediction about how the Shannon entropy of the answers of participants in the Private Treatment of the Social Image Experiment should compare to that of the answers of participants in the Public Treatment. Specifically, the model suggests that, depending on the initial distribution of types, the distortions generated by social image may reduce informativeness of equilibrium statements according to the Blackwell partial order and, at the same time, increase the Shannon entropy of the statements. Only when social image concerns are sufficiently strong as to generate a substantial degree of conformity, does the model imply that the distortions induced by social image decrease the entropy of equilibrium statements.

Figure A5 shows that, empirically, the entropy of the answers of participants in the Private

Treatment is higher than the entropy of the answers of participants in the Public Treatment.

Figure A5: **Shannon Entropy**



Notes: The figure above presents estimates of the Shannon Entropy of the distribution of the index of sensitive attitudes among participants in the Private and the Public Treatments of the Social Image Experiment. The index of sensitive attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. The index is rounded to the nearest digit. The entropy estimate applies the Miller-Madow correction as described in Paninski (2003). Error bars reflect 95 percent confidence intervals. The p-value refers to a one-sided test where the null hypothesis is that the Shannon Entropy in the Public Treatment is weakly larger than Shannon Entropy in the Private Treatment.

E.4 Garbling Matrix Analysis

In order to determine whether information loss in the experiment is primarily due to pooling or to scrambling, we leverage insights from the theoretical framework. Specifically, consider Case 1 of the model, namely the one in which social image susceptibility is heterogeneous. We know from Proposition 2 that equilibria in environments that do not substantially engage the agent's social image concerns ($e < e^*$) are Blackwell more informative than equilibria in environments that do substantially engage the agent's social image concerns ($e > e^*$). Let $\boldsymbol{v}_{private}$ be a row vector summarizing the distribution of equilibrium statements when $e < e^*$ and $\boldsymbol{v}_{public}$ be a row vector summarizing the distribution of equilibrium statements when $e > e^*$. As shown in Blackwell and Girshick (1954), the result about Blackwell informativeness in Proposition 2 can be equivalently stated as follows: $\exists$ a row-stochastic matrix M such that $\boldsymbol{v}_{public} = \boldsymbol{v}_{private} \cdot M$. From Proposition 1, we also know that M must be upper triangular.

Garbling matrix M describes the way in which the agents' private issue-positions are distorted by social image and, as such, it is a matrix that reveals information about the mechanisms driving information loss. In the remainder of this section, we present an approach aimed at obtaining an empirical counterpart to M. Let $\hat{\boldsymbol{v}}_{private}$ and $\hat{\boldsymbol{v}}_{public}$ denote the empirical distributions of the index of sensitive attitudes, rounded to the nearest digit, of participants in the Private and Public Treatment respectively. Suppose, only for this exercise, that participants in the Private Treatment truthfully report their private issue-positions.[59] In order to obtain an empirical counterpart to M, we solve

$$\min_{A \in \mathcal{M}_{10\times 10}} \|\hat{\boldsymbol{v}}_{private} \cdot A - \hat{\boldsymbol{v}}_{public}\| \tag{4}$$

subject to

$$A \text{ is upper triangular and row} - \text{stochastic}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\mathcal{M}_{10\times 10}$ denotes the space of 10×10 matrices.[60] Let M^* denote the 10×10 matrix that solves the minimization problem. Figure A7 compares $\hat{\boldsymbol{v}}_{public}$ and $\hat{\boldsymbol{v}}_{private} \cdot M^*$, and shows that the two distributions are almost identical. Therefore the distribution of the index of sensitive attitudes of participants in the Public Treatment of the Social Image Experiment is very close, under a suitable metric, to being a garbled version of the distribution of the index of sensitive attitudes of participants in the Private Treatment of the Social Image Experiment, where the garbling matrix has the additional feature, implied by the model, of being upper triangular.

M^* reveals information about the ways in which the answers of participants in the Public Treatment of the Social Image Experiment are distorted compared to the answers of participants in the Private Treatment. If information loss was primarily due to pooling at the top end of the Likert scale, we would expect the rows of M^* to be approximately equal to 0 in all but one position, we would expect the second to last row to have a 1 in the last position, and we would expect M^* to be weakly order preserving. Conversely, if information loss was primarily due to scrambling, we would expect a spread out distribution of mass on at least some of the rows of M^*, and we would expect M^* not to be order preserving.

Figure A6 shows a heat map of M^*, where darker colors signify larger numbers. As shown in Figure A6, M^* is more consistent with scrambling than with pooling: the overall mapping is not order preserving and, especially for rows 1 through 4, the distribution of mass in the row is rather spread out.

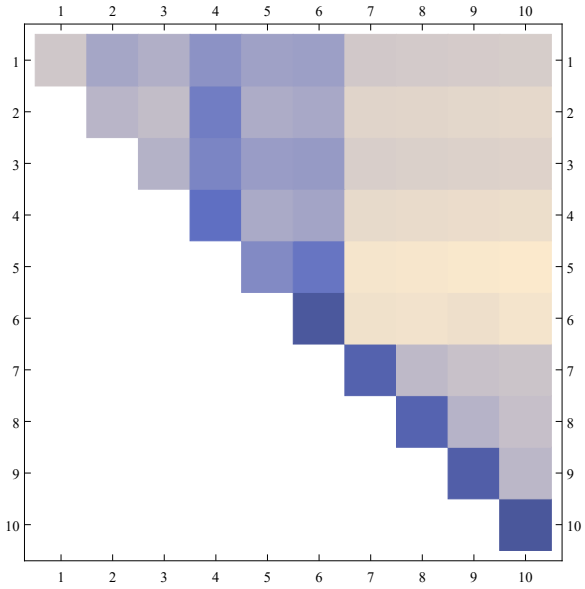
M^* corroborates and enriches the intuitions from the heterogeneous treatment effect analysis.

[59] This assumption is not required anywhere else in the empirical analysis.

[60] The Likert scale on which participants reported their levels of agreements with the sensitive and placebo statements is actually an 11-point scale. However, there are no participants whose index of sensitive attitudes, when rounded, equals zero. For the sake of compactness, we omit the zero category and consider a 10×10 matrix M rather than an 11×11 matrix.

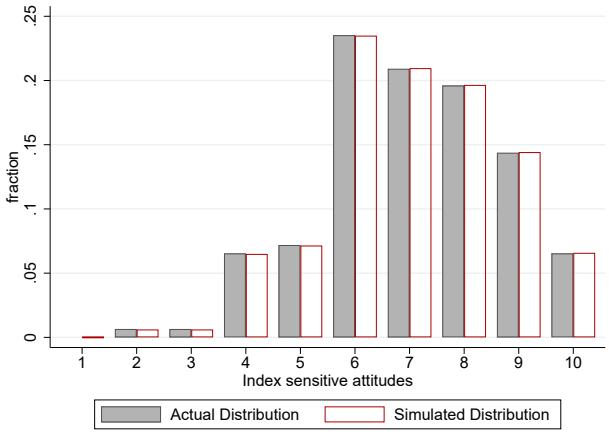
According to Figure A6, large values of the index of sensitive attitudes in the Private Treatment tend to map mostly onto themselves. Conversely, low values of the index of sensitive attitudes in the Private Treatment tend to map onto values that, according to the answers of participants in the Exploratory Survey, are perceived to be slightly more socially acceptable at UCSB. Therefore, M^* lends additional support to the idea that most of the misreporting in Public Treatment of the Social Image Experiment and most of the information loss is due to participants who, if assigned to the Private Treatment, would have given answers that are far from the end of the spectrum that is perceived to be more socially acceptable at UCSB. As shown in Figure 6, such participants tend to be students who self-identify as moderate or conservatives.

Figure A6: **Heat Map of Garbling Matrix M^***



Notes: The figure above presents a heat map of matrix M^* that solves Problem 4. Darker squares correspond to larger numbers.

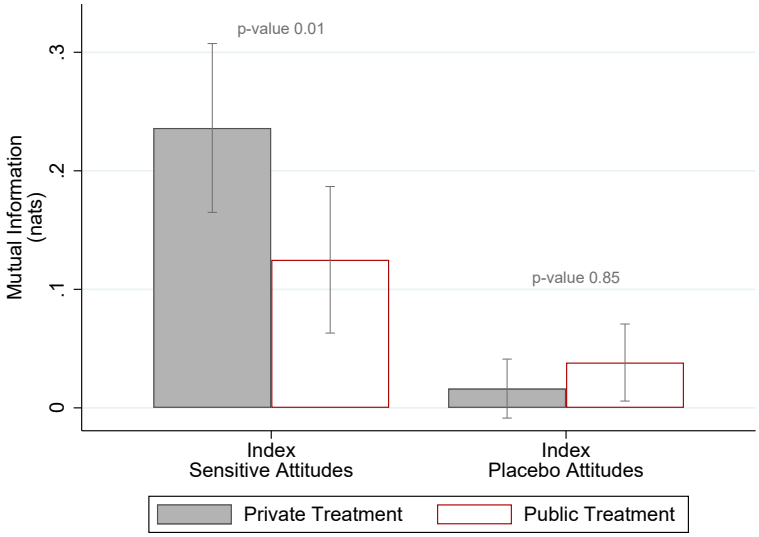
Figure A7: **Actual vs. Simulated Distribution of Answers from the Public Treatment**



Notes: The figure above compares the empirical distribution of the index of sensitive attitudes, rounded to the nearest digit, of participants in the Public Treatment of the Social Image Experiment ($\hat{\boldsymbol{v}}_{public}$) to the distribution of answers obtained by garbling the empirical distribution of the index of sensitive attitudes, rounded to the nearest digit, of participants in the Private Treatment of the Social Image Experiment ($\hat{\boldsymbol{v}}_{private}$) using matrix M^*.

E.5 Robustness Checks

Figure A8: **Mutual Information: Index of Sensitive vs. Index of Placebos Attitudes**



Notes: The figure above presents estimates of mutual information (in nats) for different duples of variables and separately for participants in the Private and the Public Treatments. The first set of columns shows estimates of the mutual information between a participant's index of sensitive attitudes, rounded to the nearest digit, and whether the participant self-identified as liberal. The second set of columns shows estimates of the mutual information between a participant's index of placebo attitudes, rounded to the nearest digit, and whether the participant self-identified as liberal. The index of sensitive (placebo) attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive (placebo) questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. Self-identifying as liberal is defined according to the median split in terms of political ideology described in Section 4.3. Error bars reflect 95 percent confidence intervals. The p-values refer to one-sided tests where the null hypothesis is that mutual information in the Public Treatment is weakly larger than mutual information in the Private Treatment.

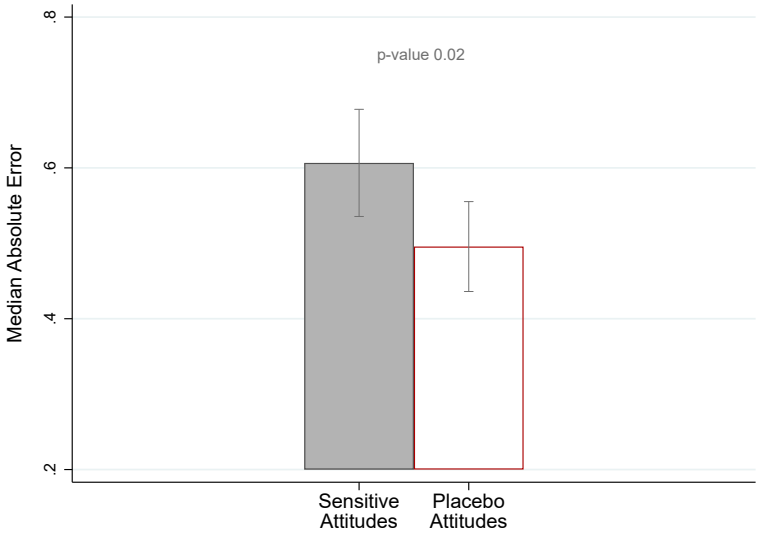
Table A19: **Mutual Information: All Demographic Characteristics**

	Mutual Information Private Treatment	Mutual Information Public Treatment	Standard Error Private Treatment	Standard Error Public Treatment	p-value
Gender	0.12	0.14	0.04	0.04	0.63
Race	0.13	0.13	0.04	0.04	0.49
Age	0.16	0.16	0.04	0.04	0.54
Self-identified as Democrat	0.14	0.07	0.03	0.03	0.06
Self-identified as Liberal	0.24	0.12	0.04	0.03	0.01
Religious	0.06	0.05	0.03	0.03	0.45
Year in School	0.10	0.08	0.03	0.03	0.37
Major (HSS vs. SE)	0.07	0.05	0.02	0.02	0.27

Notes: The table above presents estimates of mutual information (in nats) between a participant's index of sensitive attitudes, rounded to the nearest digit, and each of the demographic characteristics elicited in the pre-screen survey of the Social Image Experiment. The index of sensitive attitudes is calculated by taking, for each participant, a simple average of the participant's answers to the sensitive questions, with the answers oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB according to the Exploratory Survey. Standard errors are in parentheses.

F Additional Empirical Results: Forecasting Experiment

Figure A9: **Forecast Accuracy to Sensitive and Placebo Statements**



Notes: The figure above compares, for participants in the Forecasting Experiment, the accuracy of the median forecast across the sensitive and placebo attitudes. Specifically, for each sensitive and placebo statement, we first calculate the absolute value of the difference between a participant's forecast of the average level of agreement with the statement among participants in the Private Treatment of the Social Image Experiment and the actual average level of agreement with the statement among participants in the Private Treatment of the Social Image Experiment. Second, we average the absolute value of the difference across the sensitive statements and, separately, across the placebo statements. Third, we plot the median value of the measures thus obtained for the sensitive and the placebo statements. Error bars reflect 95 percent confidence intervals. The p-value refers to a two-sided t-test.

F.1 Robustness Checks: Forecasting Experiment

It is worthwhile calculating, for the dimensions of race and gender, the probability of observing heterogeneous treatment effects as extreme as the ones observed in the Social Image Experiment under the null hypothesis that the answers of the plurality of subjects in the Forecasting Experiment is correct. In the Social Image Experiment, the difference between the treatment effect on the index of sensitive attitudes for males and for females is approximately -0.1 points on the 0-10 Likert scale. Under the null hypothesis that the treatment effect on males is at least 0.7 points larger than the treatment effects on females, the probability of observing a difference in treatment effects at least as extreme as the one observed in the Social Image Experiment is 0.03. Similarly, in the Social Image Experiment, the difference between the treatment effect on the index of sensitive attitudes for whites and for non-whites is approximately 0.2 points on the 0-10 Likert scale. Under the null hypothesis that the treatment effect on whites is at least 0.7 points larger than the treatment effects

on non-whites, the probability of observing a difference in treatment effects at least as extreme as the one observed in the Social Image Experiment is 0.12. Therefore, we can be fairly confident that participants in the Forecasting Experiment have inaccurate beliefs about the extent to which treatment effects in the Social Image Experiment are heterogeneous across the race and gender dimensions.