Online Appendix: Not for Publication

Political Correctness, Social Image, and Information Transmission

Luca Braghieri

A Signaling Model

As discussed, when studying information loss from the perspective of a sophisticated audience, it is possible to embed a signaling model with lying costs along the lines of Kartik (2009) in the simple communication framework described in the main body of the paper. For ease of exposition, I relegate all proofs to Appendix B.

A.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 garners greater social approbation (Sudman and Bradburn, 1974). One way of thinking about this model is that it provides a microfoundation for the phenomenon of social desirability bias in survey responses.

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 transgender women should be allowed to participate in women's sports. 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 sophisticated 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, I think of statements $x \in X = \Pi$ as having exogenous commonly-understood meaning "my private issue-position is x".[1] As a consequence, I say that an agent

[1] I relax this assumption in Appendix C and allow the interpretation of natural language to be heterogeneous

is misrepresenting her private issue-position whenever she chooses a statement $x \neq \pi$. I assume the agent pays a cost whenever she misrepresents her private issue-position and I 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. I assume $c(x, \pi)$ is single-troughed, symmetric around $x = \pi$, and convex.[2] For simplicity, I 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. I 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 $E_\beta(\pi|x)$.

I let the agent's utility depend positively on $E_\beta(\pi|x)$; i.e. I 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 A.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, I 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, I let $e = s \cdot r \in [0, 1]$ and consider comparative statics directly on e. I 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 I refer to

across agents.

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

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.

A.2 Defining Informativeness

Equilibrium statements reveal information about the agent's type (π, ς). In the context of the debate about political correctness, I assume the audience is interested in learning about the agent's private issue-position π.[3] 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 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.

I formalize the notion of relative informativeness by appealing to the canonical partial order of Blackwell (Blackwell, 1951, 1953). Specifically, I 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 $\mathcal{E}'$. It is worthwhile 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.

I conclude this section by stating a simple result that I leverage in the experiment. Specifically, in

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

the experiment I 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, I will study the extent to which statements (x) are informative about observable characteristics and private behaviors (τ) that I expect to be correlated with the subjects' private issue-positions (π). The result below links information loss about π to information loss about τ.

Lemma 1. *For all equilibria $\mathcal{E}$ and $\mathcal{E}'$ with $\mathcal{E}$ more informative about π than $\mathcal{E}'$ and for all 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 $\mathcal{E}'$.*[4,5]

A.3 Equilibrium Analysis

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

Assumption 1. *For every statement $x \in X$ and private issue-position $\pi \in \Pi$, with $x \neq \pi$, I 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.[6]

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. I refer to this phenomenon as "misreporting in the socially acceptable direction". Finally,

[4] 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, I 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 $\mathcal{E}'$.

[5] 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.

[6] 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 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.

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

A.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, I say an equilibrium is fully informative about π when types with different private issue-positions make different equilibrium statements.

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

When the environment does engage the agent's social image concerns ($e > e^*$), I 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, I consider:

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, I 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, the 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.[7]

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 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 I 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.[8] Intuitively,

[7] 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 A.5 for further details.

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

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, I assumed the interpretation of natural language - i.e., the mapping between x and π in the absence of distortions - is homogeneous across agents. Appendix C 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.

A.5 Discussion

Overall, the signaling model introduced in this section 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 C 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, 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).

B Proofs

B.1 Proof of Lemma 1

Let $|\Omega| = m$ and consider the experiment of drawing a statement from equilibrium $\mathcal{E}$ to in order to gather information about τ. We can represent such experiment by means of matrix A below:

$$A = \begin{pmatrix} P(x = 1|\tau_1) & \cdots & P(x = \bar{x}|\tau_1) \\ \vdots & \ddots & \vdots \\ P(x = 1|\tau_m) & \cdots & P(x = \bar{x}|\tau_m) \end{pmatrix}$$

We know

$$P(x|\tau) = \sum_{\pi \in \Pi} \left[\int_0^{\bar{\varsigma}} P(x|\tau, \pi, \varsigma) f(\varsigma|\tau, \pi) \, d\varsigma \right] P(\pi|\tau) =$$

$$= \sum_{\pi \in \Pi} \left[\int_0^{\bar{\varsigma}} P(x|\pi, \varsigma) f(\varsigma|\pi) \, d\varsigma \right] P(\pi|\tau) =$$

$$= \sum_{\pi \in \Pi} P(x|\pi) P(\pi|\tau)$$

where the second step follows because, once we condition on π and ς, the equilibrium distribution of x does not depend directly on τ, and because we assumed that, conditional on π, τ is independent of ς.

Letting

$$C = \begin{pmatrix} P(\pi_1|\tau_1) & \cdots & P(\bar{\pi}|\tau_1) \\ \vdots & \ddots & \vdots \\ P(\pi_1|\tau_m) & \cdots & P(\bar{\pi}|\tau_m) \end{pmatrix} \; and \; P = \begin{pmatrix} P(x_1|\pi_1) & \cdots & P(\bar{x}|\pi_1) \\ \vdots & \ddots & \vdots \\ P(x_1|\bar{\pi}) & \cdots & P(\bar{x}|\bar{\pi}) \end{pmatrix}$$

we know $A = CP$.

Denote by A', C' and P' the counterparts to matrices A, C and P under equilibrium $\mathcal{E}'$. Notice $C' = C$. Notice furthermore that

$$A' = C'P' = CP' = CPM = AM$$

where M is a garbling matrix. The third equality follows because we assumed $\mathcal{E}$ is Blackwell more informative about π than $\mathcal{E}'$, which, by the standard characterization of Blackwell informativeness in terms of garbling matrices (Theorem 12.2.2 in Blackwell and Girshick, 1954), implies there exists a garbling matrix M such that $P' = PM$.

But then, A' is a garbled version of A, which, by the same theorem, implies that $\mathcal{E}$ is Blackwell more informative about τ than $\mathcal{E}'$. □

B.2 Proof of Proposition 1

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

Lemma 2. *For all $e \in [0,1]$, 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 $\{E_\beta(\pi|x=1), E_\beta(\pi|x=2)..., E_\beta(\pi|x=\bar{\pi})\} \in [1,\bar{\pi}]^n$ 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 $A_\pi = \{x \in X | E_\beta(\pi|x) \geq E_\beta(\pi|\pi)\}$. For all $x \in A_\pi$, 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, for all $x \in B_\pi$, 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, for every $x \in B_\pi$, 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 $x^M = \underset{x\in B_\pi}{\arg\max} E_\beta(\pi|x)$. 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

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

and

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

We can rewrite the inequalities above as

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

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. For every $\pi \in \Pi$ and $x \in X$, let $C_{\pi,x} = \{(\tilde{\pi}, \varsigma) \in \Pi \times [0, \bar{\varsigma}] \,|\tilde{\pi} = \pi \;\wedge\; x^*\left(\tilde{\pi}, \varsigma\right) = x\}$ and $|C_{\pi,x}| = P\left((\pi, \varsigma) \in C_{\pi,x}\right)$.

Consider any $\pi \in \Pi$ for which $B_\pi = \{\pi\}$. Given the agent's best-response function derived above, we know $|C_{\pi,x}| = 0$ if $x \neq \pi$ and $|C_{\pi,x}| = P\left(\pi\right)$ if $x = \pi$.

Now consider any $\pi \in \Pi$ for which $B_\pi \neq \{\pi\}$. Given the agent's best-response function derived above, we know that for every $\pi \in \Pi$ for which $B_\pi \neq \{\pi\}$ and for every $x \in X$, $|C_{\pi,x}|$ can be written as:

$$|C_{\pi,x}| = \begin{cases} 0 & if\ x \notin B_\pi\ or\ \left[x \in B_\pi \backslash \left\{\pi, x^M\right\}\ and\ \star_{\pi,x}\ is\ not\ satisfied\right] \\ \int_{\max\limits_{\tilde{x}\in B^l_{\pi,x}} \left\{\frac{c(x,\pi)-c(\tilde{x},\pi)}{e\left[E_\beta(\pi|x)-E_\beta(\pi|\tilde{x})\right]}\right\}}^{\min\limits_{\tilde{x}\in B^u_{\pi,x}} \left\{\frac{c(\tilde{x},\pi)-c(x,\pi)}{e\left[E_\beta(\pi|\tilde{x})-E_\beta(\pi|x)\right]}\right\}} f\left(z|\pi\right) dzP\left(\pi\right) & if\ x \in B_\pi \backslash \left\{\pi, x^M\right\}\ and\ \star_{\pi,x}\ is\ satisfied \\ \int_0^{\min\limits_{\tilde{x}\in B^u_{\pi,\pi}} \left\{\frac{c(\tilde{x},\pi)}{e\left[E_\beta(\pi|\tilde{x})-E_\beta(\pi|\pi)\right]}\right\}} f\left(z|\pi\right) dzP\left(\pi\right) & if\ x = \pi \\ \int_{\max\limits_{\tilde{x}\in B^l_{\pi,x^M}} \left\{\frac{c\left(x^M,\pi\right)-c(\tilde{x},\pi)}{e\left[E_\beta\left(\pi|x^M\right)-E_\beta(\pi|\tilde{x})\right]}\right\}}^{\infty} f\left(z|\pi\right) dzP\left(\pi\right) & if\ x = x^M \end{cases}$$

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

Then, $E_\beta\left(\pi|x\right)$ can be written as:

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

The expression above is well-defined because, as already shown, for every

$$\left\{E_\beta\left(\pi|x=1\right), E_\beta\left(\pi|x=2\right) ..., E_\beta\left(\pi|x=\bar{\pi}\right)\right\} \in \left[1,\bar{\pi}\right]^n$$

and for every $x \in X$, there exist a positive measure of types for whom x is a best response. In other words, all statements are on the equilibrium path; therefore, all beliefs can be derived via Bayes' rule.

We can now apply Brouwer's fixed-point theorem. We know that for every $x \in X$, $E_\beta\left(\pi|x\right) \in [1,\bar{\pi}]$. We define a function $q:\ [1,\bar{\pi}]^n \to [1,\bar{\pi}]^n$ 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 3. *For all $e \in [0,1]$, if any misreporting occurs in equilibrium, it occurs in the socially acceptable direction; i.e. $x^*\left(\pi,\varsigma\right) \geq \pi$.*

Proof

We prove the lemma by induction. Base case: for every (π,ς) with $\pi = \bar{\pi}$ and $\varsigma \in [0,\bar{\varsigma}]$, $x^*\left(\bar{\pi},\varsigma\right) = \bar{\pi}$. Suppose, aiming towards contradiction, that there exists an equilibrium in which, for some (π,ς) with $\pi = \bar{\pi}$, $\varsigma \in [0,\bar{\varsigma}]$, $x^*\left(\bar{\pi},\varsigma\right) \neq \bar{\pi}$. Consider any type $(\bar{\pi},\varsigma)$ s.t. $x^*\left(\bar{\pi},\varsigma\right) = \tilde{x} < \bar{\pi}$. It must be the case that, in equilibrium, $E_\beta\left(\pi|\tilde{x}\right) > E_\beta\left(\pi|\bar{\pi}\right)$.

Notice it cannot be the case that, in equilibrium, $E_{\beta}\left(\pi|x=\bar{\pi}-1\right)>E_{\beta}\left(\pi|x=\bar{\pi}\right)$. 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

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

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

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

A necessary condition for the inequality above to be satisfied is

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

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

A necessary condition for the inequality above to be satisfied is

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

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

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

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

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

Let $e^* = \min_{k \in \{1,..,\bar{\pi}\}} \left\{\frac{1}{\bar{\varsigma}}\frac{c_k}{k}\right\}$, where $c_k \doteq c(\pi + k, \pi)$. 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 $\tilde{k} \in$

$\underset{k\in\{1,..,\bar{\pi}-1\}}{\arg\min}\left\{\frac{1}{\bar{\varsigma}}\frac{c_k}{k}\right\}$. Then, it must be the case that type $(1,\bar{\varsigma})$ is better off choosing $x=1$ than $x=\tilde{k}+1$; i.e.

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

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

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

which contradicts the assumption that $e>e^*$.

Let's now show that for all $e<e^*$ 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: for all $e<e^*$, 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, $E_\beta\left(\pi|x=1\right)=1$. For all $k\in\{1,...,\bar{\pi}-1\}$, we want

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

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

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

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

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

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 no type (π,ς) with $\pi=\tilde{\pi}$ is better off misreporting her private issue-position than truthfully reporting it. For all $k\in\{1,...,\bar{\pi}-\tilde{\pi}\}$, we want

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

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

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

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

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

which we assumed.

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

B.3 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, for all $x \in X$, 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, there exist $\tilde{x} \in X$ 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 $\pi \neq 1$; for every $x > 2$, $P(\pi|x) = 1$ 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 there 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 for every $x, x' \in X$, 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 $k \in \{2, ..., \bar{\pi} - \pi\}$?

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

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

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

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

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

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

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

C Heterogeneous Interpretation of Natural Language

The model introduced in Appendix A 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 Appendix A, 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, I 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 Appendix A.1. The only differences are as follows. First, much like in Appendix A.4, I create some slack in the message space; specifically, I assume $\beta\left(\pi\right) = 0$ for $\pi \in \{\pi_a, ..., \pi_b\}$, where $\pi_a \leq \pi_b$.[9] Second, I assume that the agent's type has an additional dimension, captured by function $g : X \rightarrow \Pi$, that describes the way in which the agent

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

interprets natural language in the absence of social image concerns. I assume g is a bijection, though the assumption can in principle be relaxed. I 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 order to close the model, I 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. I 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.[10] Specifically, I 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 $\beta(\pi) = 0$, for every $\pi \leq \frac{\bar{x}}{2}$.

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

1. *For all $e \in [0, 1]$, an HNLI equilibrium exists.*
2. *There 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. *There 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. *There exist 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 2 of Proposition 1, with the difference that, in this case, I 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.

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

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 $g\left(x\right) = x$. For $\pi = \frac{\bar{\pi}}{2} + 1 + k$ with $k \in \left\{1, ..., \frac{\bar{\pi}}{2} - 1\right\}$, assume

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

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 $x = \frac{\bar{x}}{2} + 1$. 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), $x = \frac{\bar{x}}{2} + 1$ 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 $\mathcal{E}'$. $\square$

D Additional Empirical Results: Encoding Experiment

D.1 Descriptive Statistics

Table A1: **Selected Universities in Top Quintile of Liberal-Conservative Ranking**

Boston College
Brandeis University
Brown University
Carnegie Mellon University
Columbia University
Duke University
Georgetown University
Harvard University
John's Hopkins University
New York University
Northeastern University
Northwestern University
Stanford University
Tufts University
UC Berkeley
UC Davis
UC Los Angeles
UC San Diego
UC Santa Barbara
University of Chicago
University of Michigan - Ann Arbor
University of Pennsylvania
University of Southern California
Washington University in St. Louis
Wellesley College
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 A2: **Differences between the waves of the Encoding Experiment at UCSB and the one at UCSD**

UCSB	UCSD
The manipulation in the Public Treatment involved showing students a screen containing the following text: "**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 UCSB laboratory recruitment system.**"	The manipulation in the Public Treatment involved showing students a screen containing the same text as the one shown in the left column (replacing UCSB with UCSD) and, right after, a screen containing the following text: "According to the UCSD laboratory recruitment system, your first and last name are: **First Name:** [participant's first name] **Last Name:** [participant's last name]."
The recipient of the donation was 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.	The recipient of the donation was Athlete Ally, a national non-for-profit organization that, among other activities, advocates for the inclusion of transgender women in women's sports.
One of the ten sensitive statements was about cultural appropriation and one was about the relationship between Islam and violence.	The sensitive statement about cultural appropriation and the one about the relationship between Islam and violence were replaced by two statements that, in 2023, were more topical. Specifically, they were replaced by a statement about the slogan "defund the police" and by a statement about the participation of transgender women in women's sports.
After the experiment, students received an email asking them whether they wanted to sign a petition to require yearly mandatory sexual harassment training at UCSB.	At UCSD, sending students an email after the end of the experiment was logistically infeasible. Therefore, I did not ask students to sign a petition.

Notes: The table above describes the differences between the first two waves of the Encoding Experiment run at UCSB in November 2019 and May 2022 and the third wave run at UCSD in May 2023.

Table A3: **Sample Sizes**

		Sample size
Encoding Experiment	Wave 1	$N = 320$
	Wave 2	$N = 551$
	Wave 3	$N = 828$
	Total	$N = 1699$
Decoding Experiment		$N = 656$

Notes: The table above presents the size of the impact evaluation sample for each wave of the Encoding Experiment and for the Decoding Experiment. The impact evaluation samples for both the Encoding and the Decoding experiments include individuals who: i) were eligible to participate in the study, ii) completed the relevant parts of the survey, and iii) were not in the fastest 10% of respondents in terms of survey duration. The eligibility criteria were: i) being an undergraduate student at UCSB or UCSD, and ii) taking more than 30 seconds to complete a brief news knowledge quiz administered before the consent form.

Table A4: **Descriptive Statistics: Encoding Experiment**

	Mean	Standard deviation	Minimum value	Maximum value
Confederate Statues	6.43	2.85	0	10
Preferred Gender Pronouns	7.18	3.01	0	10
Trigger Warnings	6.07	2.78	0	10
Sexual Harassment Training	8.16	2.11	0	10
Illegal Immigration*	3.42	2.71	0	10
Reparations for Slavery	5.87	2.90	0	10
Racial Microaggressions	5.99	2.62	0	10
Blackface Halloween*	1.58	2.45	0	10
Parents Smartphones	7.33	1.96	0	10
Import Tariffs	5.24	1.98	0	10
School Uniforms*	4.90	2.71	0	10
Penny*	5.33	2.90	0	10
European Union	4.84	1.35	0	10

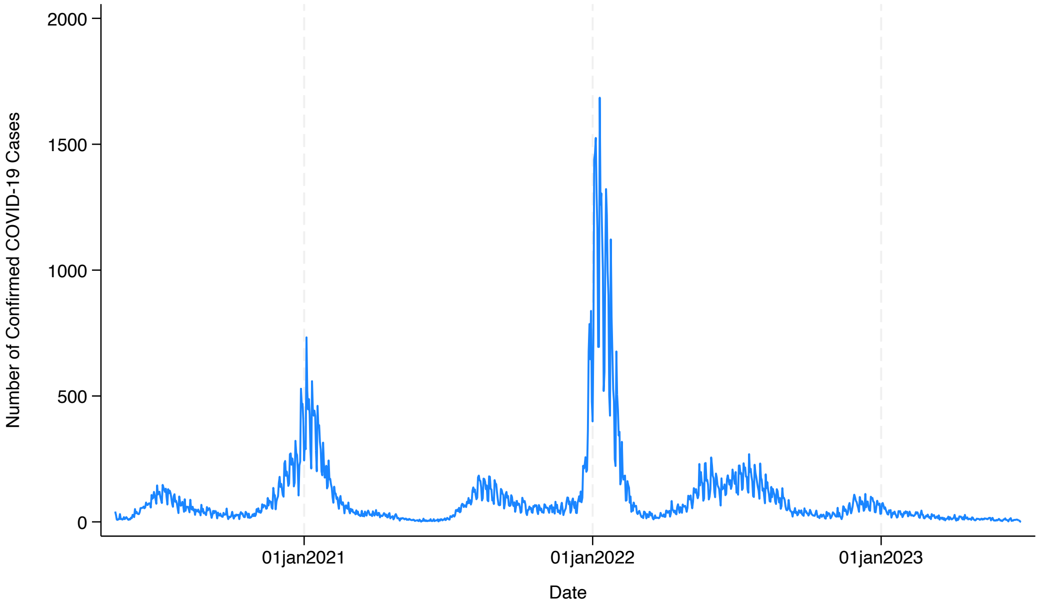
Notes: The table above presents the means, standard deviations, minimum values, and maximum values of the levels of agreement of participants in the Private Treatment of the Encoding Experiment with the eight sensitive and the five placebo statements that are common to the three waves of the Encoding Experiment. The statistics are reported in original units; therefore, larger numbers correspond to higher levels of agreement. A statement is marked with an asterisk (*) if, in the context of an exploratory survey run in support of wave three of the Encoding Experiment, the fraction of participants who answered that disagreeing with the statement is more socially acceptable at UCSD than agreeing with the statement is larger than the fraction of participants who answered the opposite.

Table A5: **Average Private Level of Agreement by Political Affiliation**

	(1) Mean Independent/Republican	(2) Mean Democrat
Confederate Statues	5.44	7.59
Preferred Gender Pronouns	6.00	8.31
Trigger Warnings	5.37	6.80
Sexual Harassment Training	7.68	8.81
Illegal Immigration*	4.50	2.28
Reparations for Slavery	4.93	6.94
Racial Microaggressions	5.35	6.86
Blackface Halloween*	2.46	0.75

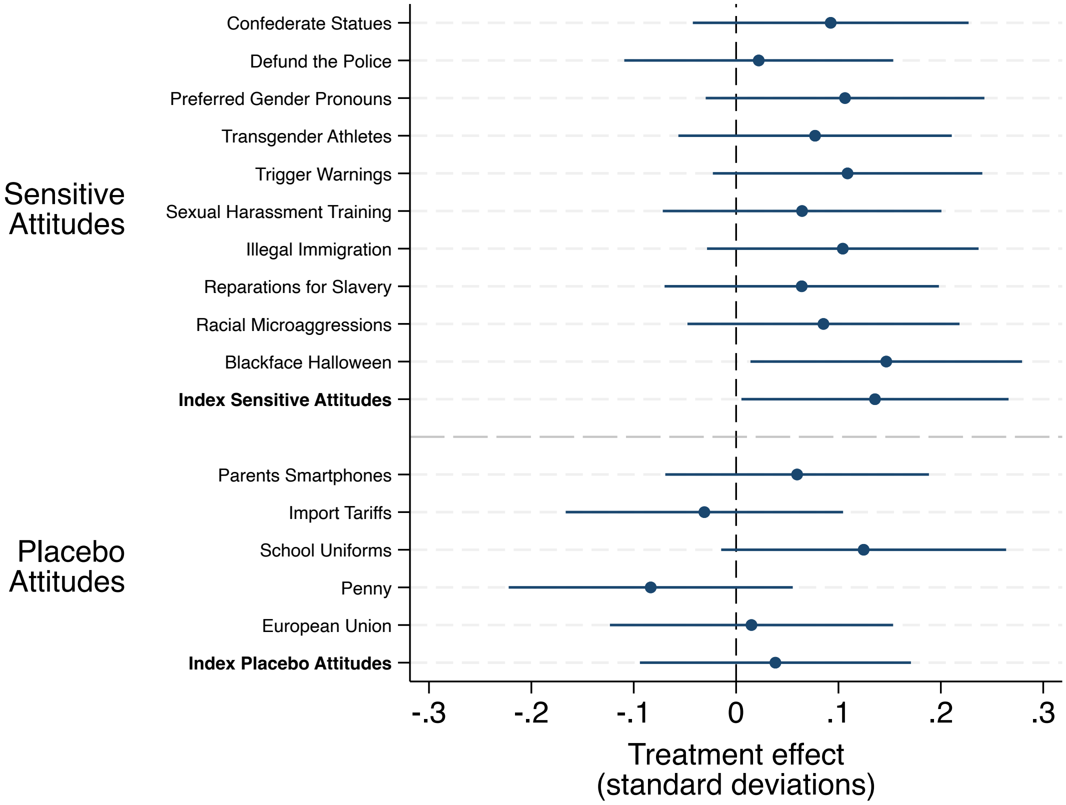
Notes: The table above reports the average level of agreement with each of the eight sensitive statements that were common to all three waves of the Encoding Experiment for students who self-identified as Democrats and students who self-identified as Independents/Republicans. The statistics are obtained using data form wave three of the Encoding Experiment. The means are reported in original units; therefore, larger numbers correspond to higher levels of agreement. A statement is marked with an asterisk (*) if, in the context of an exploratory survey run in support of wave three of the Encoding Experiment, the fraction of participants who answered that disagreeing with the statement is more socially acceptable than agreeing with the statement is larger than the fraction of participants who answered the opposite.

Figure A1: **Confirmed Number of COVID-19 Cases in Santa Barbara County over Time**



Notes: The figure shows a time series of the number of confirmed COVID-19 cases in Santa Barbara county, where I ran the first two waves of the Encoding Experiment. As shown in the figure, the 2022 wave of the Encoding Experiment run at UCSB was close to the peak of the COVID-19 pandemic. The data is from the California Health and Human Services Agency (CalHHS, 2023).

Figure A2: **Average Treatment Effects: Wave Three of Encoding Experiment**



Notes: The figure above presents average treatment effects of being assigned to the Public Treatment using Equation (1) from Section III. The results are obtained using data from wave three of the Encoding Experiment. 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 are generally perceived to be more socially acceptable at UCSD. 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 listed in the figure, with the answers already re-oriented. All variables are normalized so that the distribution of answers of participants in the Private Treatment has a mean of zero and a standard deviation of one. The normalization is achieved by first subtracting from a variable the average value of the variable among participants in the Private Treatment and then dividing the result by the standard deviation of the variable among participants in the Private Treatment. Error bars reflect 95 percent confidence intervals.

D.2 Robustness

Table A6: **Robustness to Excluding One Statement at a Time from the Index of Sensitive Attitudes**

	Treatment effect	Standard error	p-value
Excluding Confederate Statues	0.13	0.07	0.05
Excluding Preferred Gender Pronouns	0.13	0.07	0.04
Excluding Trigger Warnings	0.13	0.07	0.05
Excluding Sexual Harassment Training	0.14	0.07	0.04
Excluding Illegal Immigration	0.13	0.07	0.05
Excluding Reparations for Slavery	0.14	0.07	0.04
Excluding Racial Microaggressions	0.14	0.07	0.04
Excluding Blackface Halloween	0.12	0.07	0.07

Notes: The figure above presents average treatment effects of being assigned to the Public Treatment using Equation (1) from Section III. The results are obtained from the dataset that pools all three waves of the Encoding Experiment. The dependent variable is always a version of the index of sensitive attitudes constructed using seven out of the eight sensitive statements common to all three waves of the Encoding Experiment. Each row of the table omits the statement listed from the construction of the index.

Table A7: **Ordered Probit on the Indices of Sensitive and Placebo Attitudes**

	Index Sensitive Attitudes	Index Placebo Attitudes
	(1)	(2)
Average Treatment Effect	0.11^{**}	0.01
	(0.05)	(0.05)
Observations	1,699	1,699

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. The results are obtained from the dataset that pools all three waves of the Encoding Experiment. 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 that are common across the three survey waves, 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 and UCSD. Standard errors are in parentheses.

Table A8: **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	0.25	0.14	0.10	0.05	0.07	0.16
Preferred Gender Pronouns	0.18	0.15	0.07	0.05	0.21	0.19
Trigger Warnings	0.21	0.13	0.09	0.05	0.11	0.16
Sexual Harassment Training	0.16	0.10	0.07	0.05	0.12	0.16
Illegal Immigration	0.11	0.13	0.05	0.05	0.39	0.19
Reparations for Slavery	0.23	0.14	0.09	0.05	0.10	0.16
Racial Microaggressions	0.33	0.12	0.13	0.05	0.01	0.07
Blackface Halloween	0.23	0.12	0.10	0.05	0.05	0.16

Notes: The table above presents average treatment effects of being assigned to the Public Treatment using Equation (1) from Section III. The results are obtained from the dataset that pools all three waves of the Encoding Experiment. The table includes the eight sensitive statements that were common to all three waves of the Encoding Experiment. The students' reported levels of agreement with the sensitive statements are oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB and UCSD. 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.

D.2.1 Ceiling Effects

One can imagine a world in which the treatment effects on Democrats and Republicans would in principle be identical, but in which ceiling effects mechanically constrain the treatment effects on Democrats and not on Republicans. Suppose it was indeed the case that the heterogeneous treatment effects on self-reported political affiliation in wave three of the Encoding Experiment were entirely driven by ceiling effects. Consider the following regression equation

$$Y_{i,j} = \alpha_j + \beta_j T_i + \delta_j M_i + \gamma_j T_i \times M_i + \varepsilon_{i,j} \tag{1}$$

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 identified as an Independent/Republican, 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 UCSD.[11]

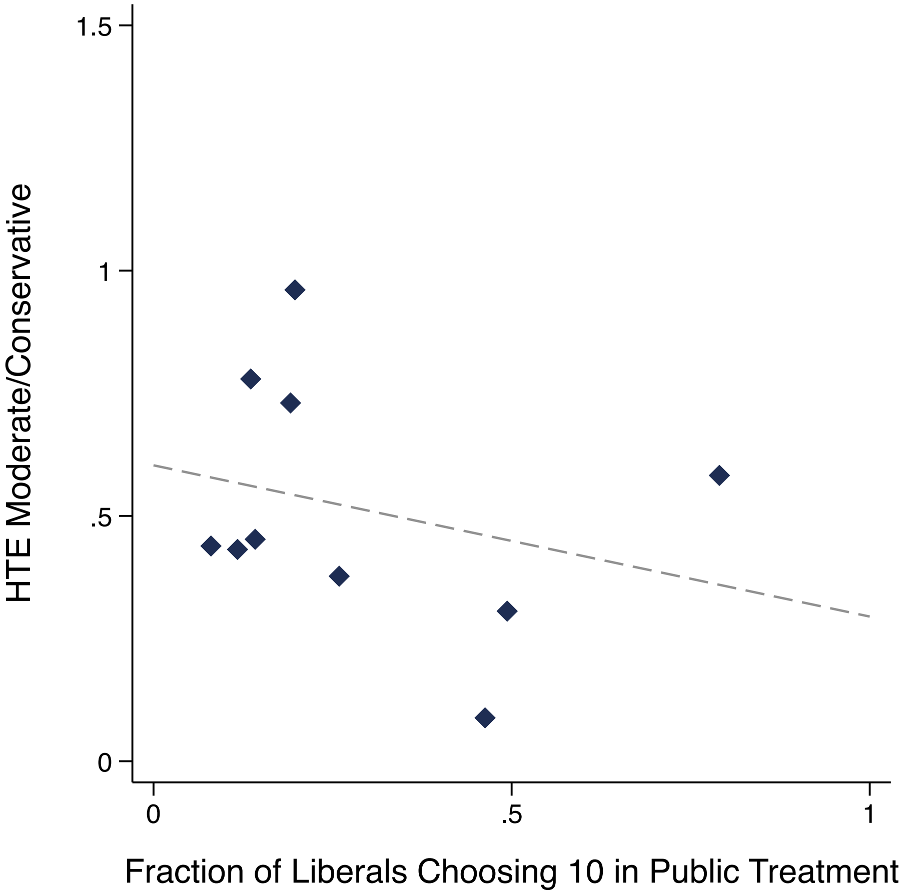
If the heterogeneous treatment effects on self-reported political affiliation were entirely driven by

[11] Notice once again that the end of the agreeing-disagreeing spectrum that is generally considered to be more socially acceptable at UCSD always coincides with the end of the spectrum that is closer to the average position of students in the Private Treatment of the Encoding Experiment who self-identify as Democrats.

ceiling effects, I would expect to see a positive relationship between the size of the heterogeneous treatment effect (γ_j) and the fraction of Democrats in the Public Treatment who bunch at the socially-acceptable end of the Likert scale. The intuition is that a larger fraction of Democrats in the Public Treatment 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 negative; therefore, the heterogeneous treatment effects on self-reported political affiliation are unlikely to be solely driven by ceiling effects.

Figure A3: **Ceiling Effects**



Notes: The figure above presents a scatter plot of the fraction of Democrats in the Public Treatment of wave three of the Encoding 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 UCSD, always equals 10) against the size of the γ coefficient in the equation from this section. The diamonds represent the ten sensitive statements. The dashed line represents the line of best fit.

E Additional Empirical Results: Information Loss

This appendix presents additional results about information loss in wave three of the Encoding Experiment.

Table A9: **Balance Check: Target Outcome Variables**

	Self-identifying as Democrat		Donation to Athlete Ally	
	(1)	(2)	(3)	(4)
Public Treatment	-0.03	-0.03	1.09	1.09
	(0.03)	(0.03)	(2.01)	(2.01)
Controls	No	Yes	No	Yes
Observations	828	828	828	828
Dependent variable mean	0.40	0.40	23.03	23.03

Notes: The table above presents average treatment effects of being assigned to the Public Treatment using Equation (1) from Section III. Depending on the specification, controls may be included. The dependent variables are: i) an indicator for self-identifying as a Democrat; ii) the amount the student donated to Athlete Ally in the dictator game. Standard errors are in parentheses.

E.1 Performance of Binary Classifiers

A simple implication of the theoretical framework from Appendix A 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 of the Encoding Experiment. In this section, I construct: i) a binary classifier for whether the participant self-identified as a Democrat based on the participant's index of sensitive attitudes, and ii) a binary classifier for whether the participant made a positive donation to Athlete Ally based on the participant's reported level of agreement with the statement about transgender athletes. In what follows, I 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, I specify a logit model relating the *binary target outcome variable* to a high-dimensional polynomial of the *predictor.* Second, I estimate the model separately for participants in the Private Treatment and participants in the Public Treatment. Third, I 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, I pick some cutoff value

p^* and classify each participant as belonging to one class 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^*, I 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 have strong evidence that 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^*.

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 I can show that, for all $p^* \in [0, 1]$, the rate of false positives and false negatives from one classifier is smaller than for the other, I will have shown that the former classifier achieves a lower expected loss than the latter classifier independently of how the loss function is defined.

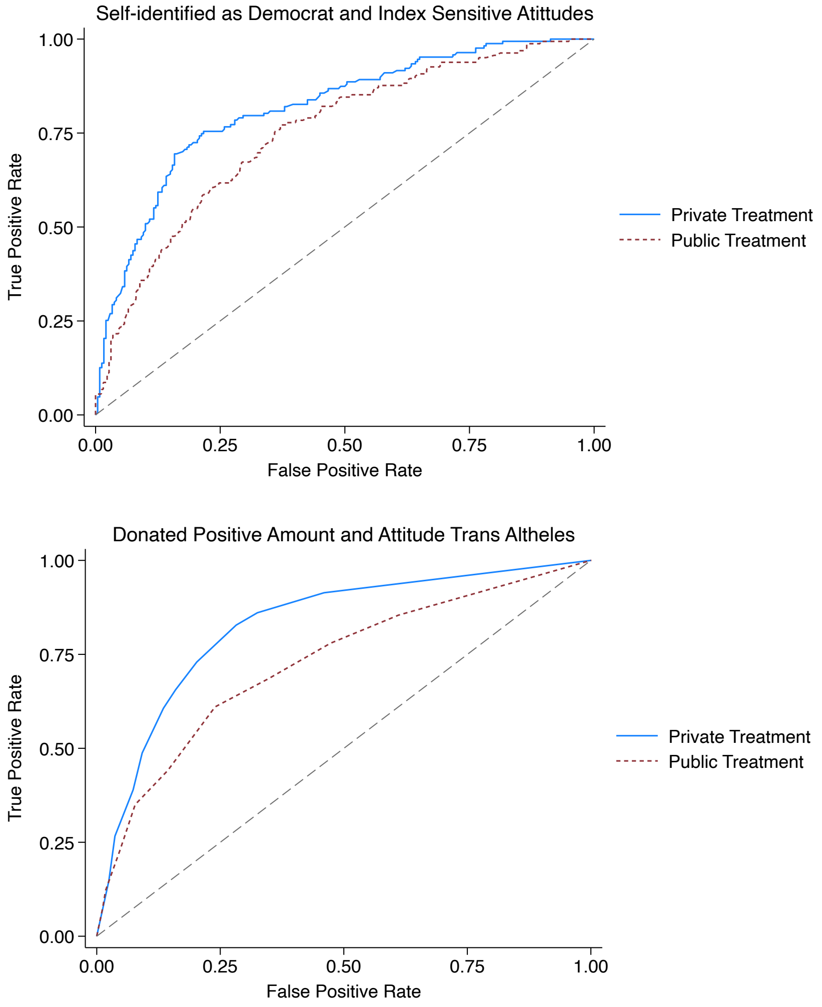
The Receiver Operating Characteristic (ROC) curve analysis below shows that the classifier constructed using data from the Private Treatment has a lower rate of false positives and false negatives than the classifier constructed using data from the Public Treatment for all $p^* \in [0, 1]$.[12] Figure 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 always achieves a higher true positive rate than the binary classifiers constructed using the answers of participants in the Public Treatment.

The performance of the binary classifiers from the Private and the Public Treatment can be compared non-parametrically by computing the areas under the ROC curves as shown by DeLong, DeLong, and Clarke-Pearson (1988). A larger area indicates better classifier performance. Table A10 shows that the areas under the ROC curves are always significantly larger for binary classifiers constructed using the answers of participants in the Private Treatment than for binary classifiers constructed using the answers of participants in the Public Treatment.

Figure A5 and Table A11 show that the accuracy of the prediction of whether students self-identify as Democrats made using the index of placebo rather than the index of sensitive attitudes are not significantly different across the Private and the Public Treatment.

[12] 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 x-axis is the false positive rate and the y-axis is the true positive rate. That curve is known as the Receiver Operating Characteristic (ROC) curve.

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



Self-identified as Democrat and Index Sensitive Atittudes

Donated Positive Amount and Attitude Trans Altheles

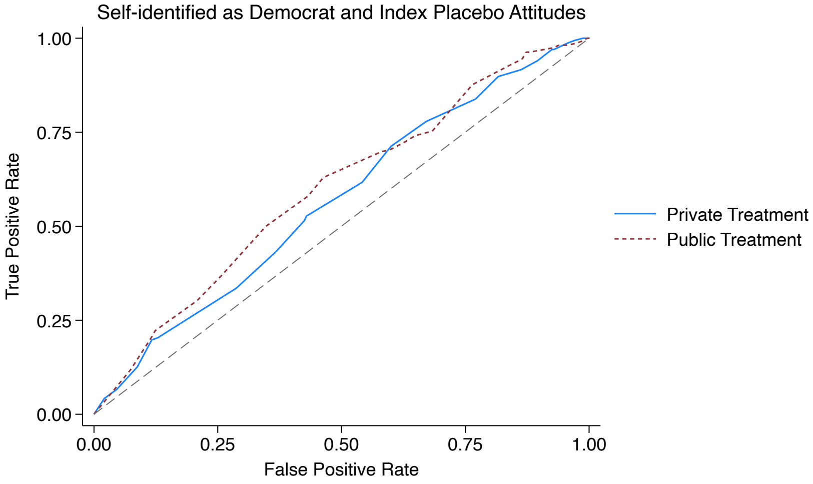
Notes: The figure presents the Receiver-Operating-Characteristic curves of the binary classifiers constructed using the answers of participants in the Private and Public Treatments, for the two binary target outcome variables. The first panel shows the ROC curves of the binary classifiers for whether a participant self-identified as a Democrat using the index of sensitive attitudes as a predictor. The second panel shows the ROC curves of the binary classifiers for whether a participant made a positive donation to Athlete Ally using the attitude towards trans athletes as a predictor. The results are obtained using data from wave three of the Encoding Experiment.

Table A10: **Comparison of Receiver-Operating-Characteristic Curves: Sensitive Attitudes**

	ROC area Private Treatment	ROC area Public Treatment	Standard Error Private Treatment	Standard Error Public Treatment	p-value
Self-identified as Democrat	0.82	0.75	0.02	0.02	0.04
Donated Positive Amount	0.83	0.73	0.02	0.02	0.00

Notes: The table above shows estimates of the areas under the Receiver-Operating-Characteristic curves from Figure A4 and compares the estimates non-parametrically as shown in DeLong, DeLong, and Clarke-Pearson (1988). The results are obtained using data from wave three of the Encoding Experiment.

Figure A5: **Receiver-Operating-Characteristic Curves: Placebo Attitudes**



Notes: The figure presents, separately for participants in the Private and the Public Treatments, the Receiver-Operating-Characteristic curves of the binary classifiers for whether a participant self-identified as a Democrat using the index of placebo attitudes as a predictor. The results are obtained using data from wave three of the Encoding Experiment.

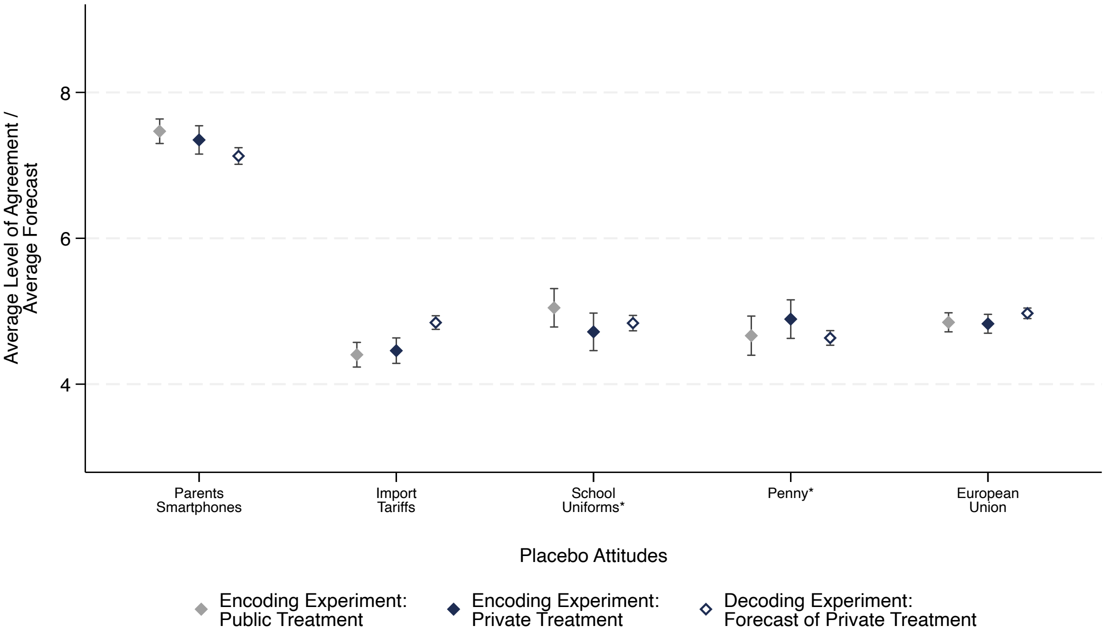
Table A11: **Comparison of Receiver-Operating-Characteristic Curves: Placebo Attitudes**

	ROC area Private Treatment	ROC area Public Treatment	Standard Error Private Treatment	Standard Error Public Treatment	p-value
Self-identified as Democrat	0.57	0.60	0.03	0.03	0.45

Notes: The table above shows estimates of the areas under the Receiver-Operating-Characteristic curves from Figure A5 and compares the estimates non-parametrically as shown in DeLong, DeLong, and Clarke-Pearson (1988). The results are obtained using data from wave three of the Encoding Experiment.

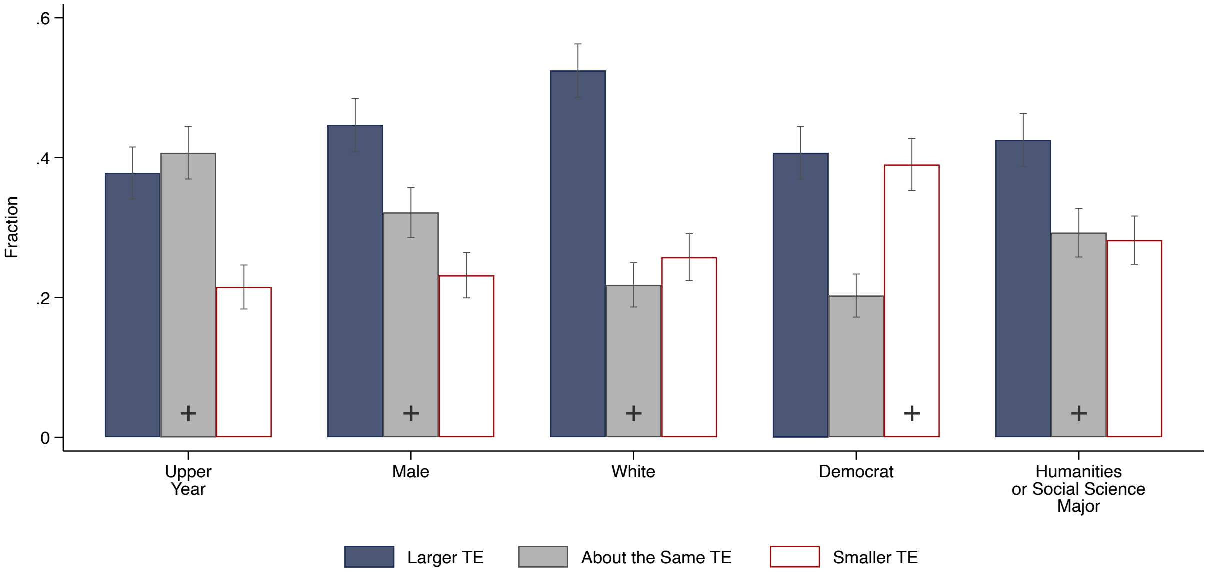
F Additional Empirical Results: Decoding Experiment

Figure A6: **Actual Average Levels of Agreement vs. Forecasts: Placebo Statements**



Notes: The figure above shows, for each placebo statement, the average level of agreement of participants in the Public Treatment of the third wave of the Encoding Experiment, the average level of agreement of participants in the Private Treatment of the third wave of the Encoding Experiment, and the mean forecast, by participants in the Decoding Experiment, of the average private level of agreement from the third wave of the Encoding Experiment. All answers are oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSD. Error bars reflect 95 percent confidence intervals.

Figure A7: **Forecasted Heterogeneity in Treatment Effects**



Notes: The figure above shows the distribution of answers of participants in the Decoding Experiment to the multiple-choice questions about heterogeneity. The first bar in each set represents the fraction of participants in the Decoding 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 Decoding 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 Decoding 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 in wave three of the Encoding Experiment. Error bars reflect 95 percent confidence intervals.

G First Version of the Decoding Experiment

I ran a first version of the Decoding Experiment in May 2020 at the University of California Santa Barbara, during the COVID-19 pandemic. Table A12 describes the differences between the version of the Decoding Experiment run at UCSB in May 2020 and the one run at UCSD in November and December 2023. The main differences between the two versions of the Decoding Experiment are: i) that the first version of the Decoding Experiment did not contain a manipulation to study social image neglect, and ii) that the first version of the Decoding Experiment showed subjects two examples of how the distribution of answers of participants in the Public Treatment of the first wave of the Encoding Experiment mapped to the average answer of participants in the Private Treatment.

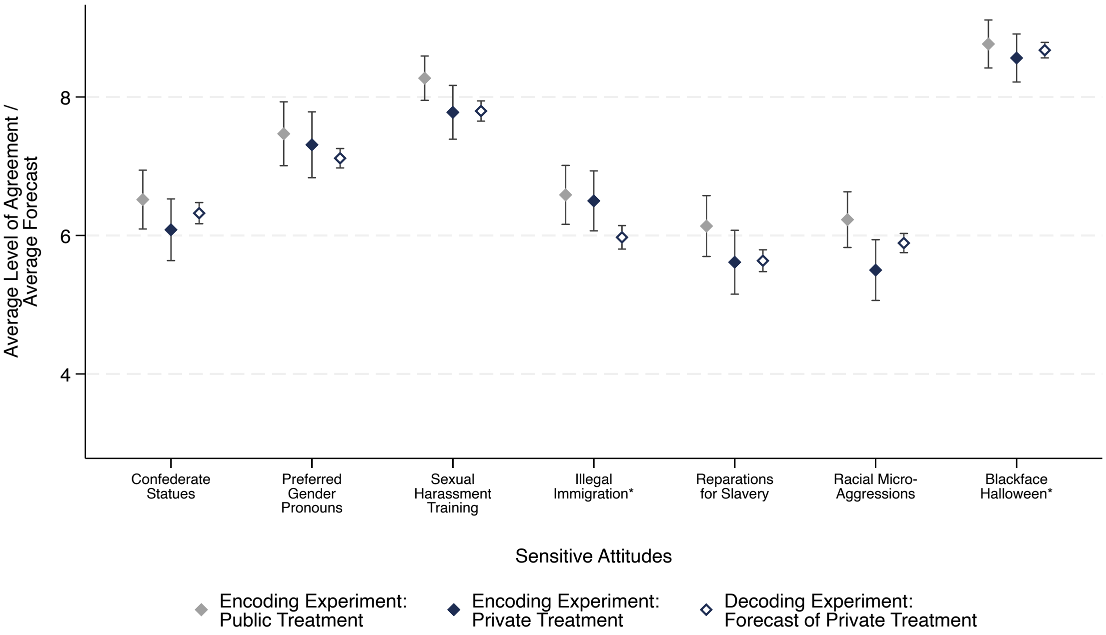
As shown in Appendix Figure A8, A9, and A10 the main difference in terms of the results of the two versions of the Decoding Experiment is that, likely due to the two examples that students were given to help them benchmark their forecasts, participants in the first version of the Decoding Experiment did not systematically overestimate the extent to which social image distorts their peers' public answers to the sensitive questions.

Table A12: **Differences between versions of the Decoding Experiment at UCSB and the one at UCSD**

UCSB	UCSD
The Decoding Experiment did not include the manipulation to study social image neglect.	The Decoding Experiment included the manipulation to study social image neglect.
Subjects were shown two examples of how the distribution of answers of participants in the Public Treatment of the first wave of the Encoding Experiment mapped to the average answer of participants in the Private Treatment.	Subjects were not shown any examples of how the distribution of answers of participants in the Public Treatment of the first wave of the Encoding Experiment mapped to the average answer of participants in the Private Treatment.
At the end of the survey, subjects were asked to report their levels of agreement with the sensitive and placebo statements.	At the end of the survey, subjects were not asked to report their levels of agreement with the sensitive and placebo statements.

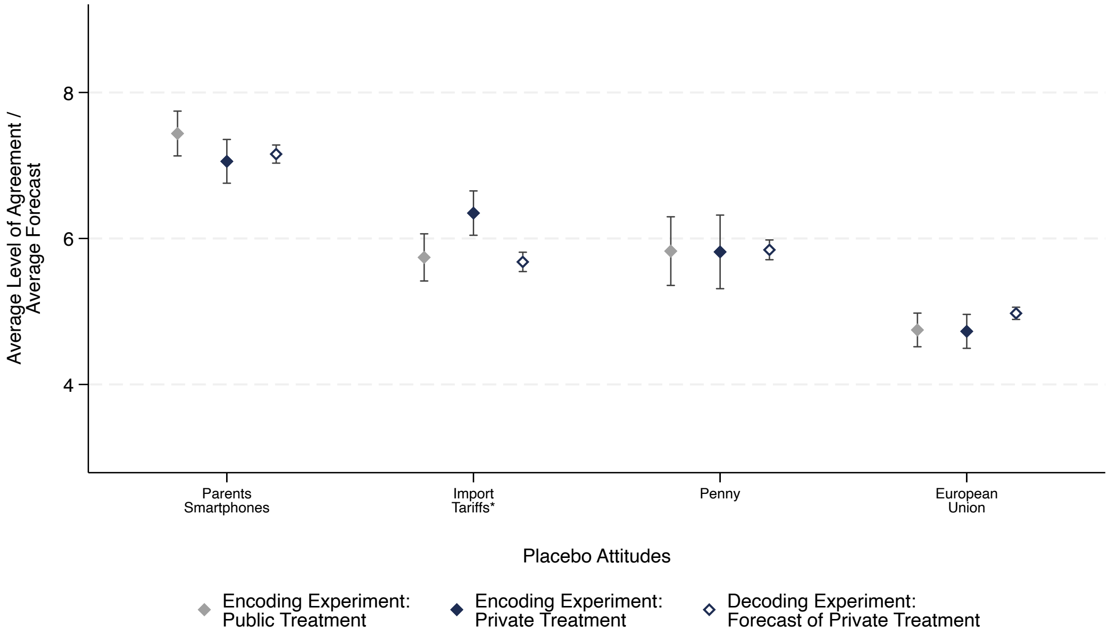
Notes: The table above describes the differences between the version of the Decoding Experiment run at UCSB in May 2020 and the version run at UCSD in November and December 2023.

Figure A8: **Actual Average Levels of Agreement vs. Forecasts: Sensitive Statements**



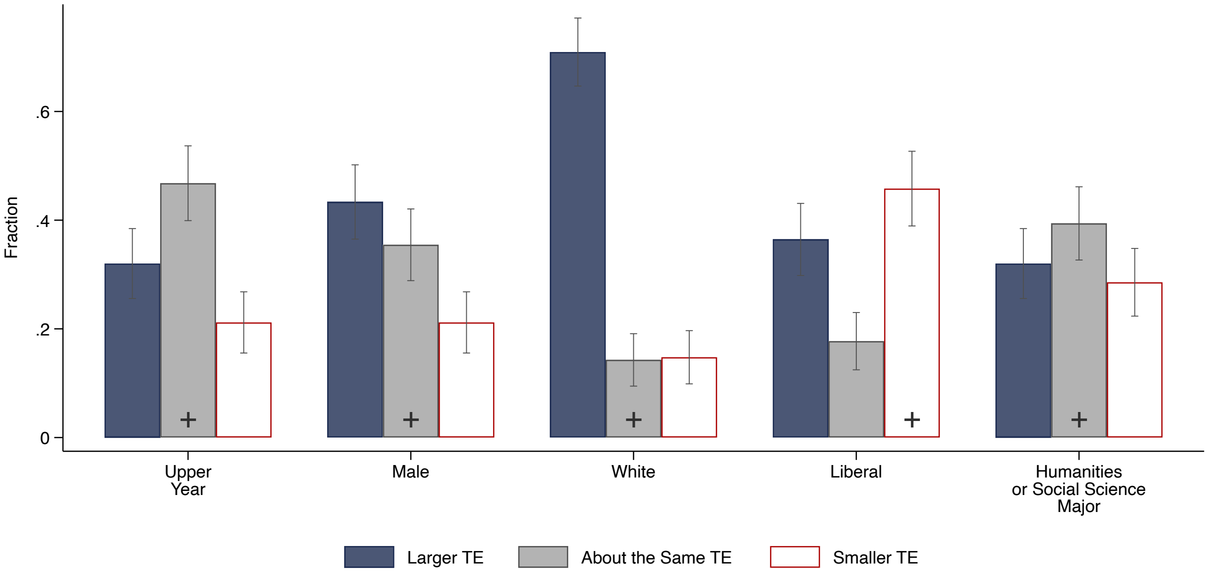
Notes: The figure above shows, for all but one of the sensitive statements that were common to all three waves of the Encoding Experiment, the average level of agreement of participants in the Public Treatment of the first wave of the Encoding Experiment, the average level of agreement of participants in the Private Treatment of the first wave of the Encoding Experiment, and the mean forecast, by participants in the first version of the Decoding Experiment, of the average private level of agreement from the first wave of the Encoding Experiment. One of the sensitive statements that were common to all three waves of the Encoding Experiment is omitted from the figure because it was shown to subjects in the Decoding Experiment as an example to help them benchmark their forecasts. All answers are oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB. Error bars reflect 95 percent confidence intervals.

Figure A9: **Actual Average Levels of Agreement vs. Forecasts: Placebo Statements**



Notes: The figure above shows, for all but one of the placebo statements that were common to all three waves of the Encoding Experiment, the average level of agreement of participants in the Public Treatment of the first wave of the Encoding Experiment, the average level of agreement of participants in the Private Treatment of the first wave of the Encoding Experiment, and the mean forecast, by participants in the first version of the Decoding Experiment, of the average private level of agreement from the first wave of the Encoding Experiment. One of the placebo statements that were common to all three waves of the Encoding Experiment is omitted from the figure because it was shown to subjects in the Decoding Experiment as an example to help them benchmark their forecasts. All answers are oriented in such a way that larger numbers always correspond to views that are generally perceived to be more socially acceptable at UCSB. Error bars reflect 95 percent confidence intervals.

Figure A10: **Forecasted Heterogeneity in Treatment Effects**



Notes: The figure above shows the distribution of answers of participants in the first version of the Decoding Experiment to the multiple-choice questions about heterogeneity. The first bar in each set represents the fraction of participants in the first version of the Decoding 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 first version of the Decoding 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 first version of the Decoding 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 in wave one of the Encoding Experiment. Error bars reflect 95 percent confidence intervals.

H Encoding Experiment: Instructions

Encoding Experiment

As we re-direct you to the main page of the study, we will enter you into a lottery. In particular, one of the participants in the study will be randomly selected to receive **$100**.

The selected participant will be given the opportunity to share some of the $100 with **Athlete Ally**, a national non-for-profit organization that, among other things, **advocates for the inclusion of transgender women** (i.e., individuals who were male at birth and transitioned later in life) **in competitive women's sports.**

If you turn out to be the participant that is randomly selected to receive the $100, how much of the $100 would you share with Athlete Ally? Please enter the amount of money that you would give to Athlete Ally below. If you are indeed selected, we will automatically implement your decision.

Notice: you can choose to give Athlete Ally anything between $0 and $100. For example, you can give Athlete Ally $0 and keep $100, or give Athlete Ally $20 and keep $80, or give Athlete Ally $80 and keep $20. These are only examples; the decision of how much to give is entirely yours.

____

After clicking the red arrow below, you will be automatically re-directed to the main page of the study and entered into the lottery.

Welcome and thank you for participating in this study!

Public Treatment Only

We will start by asking you a set of questions about socio-political topics that have sparked public debates in recent years.

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 UCSD 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 UCSD laboratory recruitment system.**

According to the UCSD laboratory recruitment system, your first and last name are:

First Name: [*First Name*]

Last Name: [*Last Name*]

Please acknowledge you understand that **your individual answers** to the questionnaire on the next screen will be shared with participants in the next phase of the study and that the UCSD laboratory system contains your **personal information**.

○ I acknowledge that **my individual answers** to the questionnaire on the next screen will be shared with participants in the next phase of the study and that the UCSD laboratory system already contains my **personal information**.

Private Treatment Only

We will start by asking you a set of questions about socio-political topics that have sparked public debates in recent years.

IMPORTANT: We will share **aggregate-level answers** to the questionnaire on the next screen with approximately 200 UCSD 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.**

Please acknowledge you understand that **aggregate-level answers** to the questionnaire on the next screen will be shared with participants in the next phase of the study and that your individual answers will remain **completely anonymous**.

○ I acknowledge that **aggregate-level answers** to the questionnaire on the next screen will be shared with participants in the next phase of the study and that my individual answers will remain **completely anonymous**.

Common to both Treatments

On a scale from 0 to 10, to what extent do you agree with the following statement: “**All statues and memorials of Confederate leaders should be removed**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10

On a scale from 0 to 10, to what extent do you agree with the following statement: “**Defunding the police is a bad idea because it will inevitably lead to increased crime rates**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10

A "**preferred gender pronoun**" is the personal pronoun (he, she, they, zhe) that a person chooses to use to identify that person's gender.

On a scale from 0 to 10, to what extent do you agree with the following statement: “**The UCSD administration should require professors to address students according to the students' preferred gender pronouns**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10

On a scale from 0 to 10, to what extent do you agree with the following statement: “**Transgender women (i.e., individuals who were male at birth and transitioned later in life) should be allowed to participate in women's sports**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10
					●					

"**Trigger warnings**" are warnings that some work contains writing, images, or concepts that may be distressing to some people.

On a scale from 0 to 10, to what extent do you agree with the following statement: “**The UCSD administration should require professors to use trigger warnings in their classes**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10
					●					

On a scale from 0 to 10, to what extent do you agree with the following statement: “**Sexual harassment training should be mandatory for everybody who works or studies at UCSD**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10
					●					

On a scale from 0 to 10, to what extent do you agree with the following statement: “**People who immigrated to the U.S. illegally, when caught, should be deported and sent back to their countries of origin**”?

Strongly Disagree	Neither Agree nor Disagree	Strongly Agree

0 1 2 3 4 5 6 7 8 9 10

Reparations for slavery is the idea that some form of compensatory payment needs to be made to the descendants of Africans trafficked to and enslaved in the Americas as a consequence of the Atlantic slave trade.

On a scale from 0 to 10, to what extent do you agree with the following statement: “**The U.S. government should provide reparations for slavery**”?

Strongly Disagree	Neither Agree nor Disagree	Strongly Agree

0 1 2 3 4 5 6 7 8 9 10

Racial microaggressions are brief and commonplace daily verbal, behavioral, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults toward people of color.

On a scale from 0 to 10, to what extent do you agree with the following statement: “**Racial microaggressions are an important problem at UCSD**”?

Strongly Disagree	Neither Agree nor Disagree	Strongly Agree

0 1 2 3 4 5 6 7 8 9 10

On a scale from 0 to 10, to what extent do you agree with the following statement: “**The UCSD administration should allow students to wear blackface for Halloween**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10

On a scale from 0 to 10, to what extent do you agree with the following statement: “**Parents should limit the amount of time their kids spend on their smartphones**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10

On a scale from 0 to 10, to what extent do you agree with the following statement: “**The United States should increase tariffs on foreign imports**”?

Strongly Disagree					Neither Agree nor Disagree					Strongly Agree
0	1	2	3	4	5	6	7	8	9	10

On a scale from 0 to 10, to what extent do you agree with the following statement: **“School uniforms help reduce clothing-related peer pressure**”?

Strongly Disagree

Neither Agree nor Disagree

Strongly Agree

0 1 2 3 4 5 6 7 8 9 10

On a scale from 0 to 10, to what extent do you agree with the following statement: “**The one-cent coin (i.e. the penny) should be removed from circulation**”?

Strongly Disagree

Neither Agree nor Disagree

Strongly Agree

0 1 2 3 4 5 6 7 8 9 10

On a scale from 0 to 10, to what extent do you agree with the following statement: “**The members states of the European Union should cede more powers to the E.U.**”?

Strongly Disagree

Neither Agree nor Disagree

Strongly Agree

0 1 2 3 4 5 6 7 8 9 10

I Decoding Experiment: Instructions

Decoding Experiment

Hi! Thank you for agreeing to participate in this study!

Today, you will be asked to **predict the results of a study that we ran a in May 2023 at UCSD**. You have a chance to win a **$10 bonus payment based on the accuracy of your predictions**.

The questions in this survey will either be multiple choice or will require a numerical answer.

For each multiple-choice question, you will earn points for choosing the correct answer. For the questions requiring a numerical answer, you will earn points for making an accurate guess. Specifically, the closer your guess is to the actual numerical answer, the more points you will earn.

At the end of the study, we will sum all the points you earned and we will assign you the $10 bonus with a probability that depends on the sum total of your points. Specifically, the more points you earned during the study, the higher the probability that you will win the $10 bonus.

The details of the point system used to determine your chance of winning the $10 bonus are a bit complicated (they are explained below if you are interested). **What is important to know is that the procedure by which the bonus is awarded ensures that it is in your best interest to give your most accurate guess to each question.**

You will receive the $7 baseline payment at the end of the study and the $10 bonus, if applicable, at the end of the data collection process.

Here's how the point system works (in case you are interested).

- You will receive between 0 and 100 points for each guess.
- For multiple-choice questions, you will earn 100 points if you select the correct answer and 0 points otherwise.
- For questions where you are asked to make a numeric guess, the formula that determines how many points you earn is as follows: $y = max\{100 - (x - g)^2, 0\}$, where y is the number of points you get for the question, x is the correct answer, and g is your guess. Therefore, if your guess is exactly correct (i.e. $g = x$), you will earn 100 points (the maximum) on that question. If your guess is more than 10 units away from the correct answer, you will earn 0 points (the minimum) on that question.
- Overall, the further your guess is from the correct answer, the fewer points you will earn.
- We will then average your points across all the questions and pay you with a probability equal to the average number of points divided by 300. For example, if you earned 90 points on average across all the questions, you will have a 90 out of 300 chance of winning the bonus.

Private-First Treatment Only

In May 2023, we ran study in which we asked around **500 undergraduate students at UCSD** sampled from the same subject pool as you to tell us the extent to which they agreed with a set of socio-political statements.

After completing a brief pre-screen survey asking them about their demographic characteristics, participants were given the following instructions:

“We will start by asking you a set of questions about socio-political topics that have sparked public debates in recent years.

IMPORTANT: We will share **aggregate-level answers** to the questionnaire on the next screen with approximately 200 UCSD 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.**”

Public-First Treatment Only

In May 2023, we ran study in which we asked around **500 undergraduate students at UCSD** sampled from the same subject pool as you to tell us the extent to which they agreed with a set of socio-political statements.

After completing a brief pre-screen survey asking them about their demographic characteristics, participants were given the following instructions:

“We will start by asking you a set of questions about socio-political topics that have sparked public debates in recent years.

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 UCSD 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 UCSD laboratory recruitment system.**

According to the UCSD laboratory recruitment system, your first and last name are:

First Name: [Participant's First Name]
Last Name: [Participant's Last Name]"

Common to Both Treatments

In case you find it useful, you can find the distribution of demographic characteristics of the participants in the previous study below:

Male	55%
White	27%
Junior/Senior	66%
Science/Engineering Major	38%

We will now ask you to forecast the answers of participants in the previous study as a function of their self-reported political affiliation.

In particular students in the previous study were asked the following question: "Do you consider yourself a **Republican**, a **Democrat**, or an **Independent**?" They had to choose an answer among the following:

1. Democrat (Strongly Democratic)
2. Democrat (Weakly Democratic)
3. Independent (leaning towards the Democratic Party)
4. Independent
5. Independent (leaning the Republican Party)
6. Republican (Weakly Republican)
7. Republican (Strongly Republican)

We say that participants who chose options 1 or 2 self-identify as **Democrats** and all other participants self-identify as **Independents** or **Republicans**.

You may find it useful to know that, according to our classification, around **40%** of participants in the previous study self-identified as **Democrats** and the remaining **60%** self-identified as **Independents** or **Republicans**.

In the question below, we will talk about participants in the previous study **agreeing** with a certain statement.

By agreeing with a statement, we mean that the participant answered a number **strictly larger than 5** on a 0-10 scale where 0 corresponded to "Strongly Disagree with the Statement", 5 corresponded to "Neither Agree nor Disagree with the Statement" and 10 corresponded to "Strongly Agree with the Statement".

Private-First Treatment Only

Before showing you the statement, we will ask you a quick question to make sure you understood the instructions given to participants in the previous study.

What did the instructions that we gave participants in the previous study say?

○ The instructions mentioned that the participants' aggregate-level answers would be shared with **students from a different college**, but that the participants' identities would remain completely confidential.

○ The instructions mentioned that the participants' aggregate-level answers would be shared with **other students at UCSD**, but that the participants' identities would remain completely confidential.

Public-First Treatment Only

Before showing you the statement, we will ask you a quick question to make sure you understood the instructions given to participants in the previous study.

What did the instructions that we gave participants in the previous study say?

○ The instructions mentioned that the participants' individual-level answers would be shared with **students from a different college** and that the participants did not need to provide their first and last name because it was already in the UCSD laboratory recruitment system.

- The instructions mentioned that the participants' individual-level answers would be shared with **other students at UCSD** and and that the participants did not need to provide their first and last name because it was already in the UCSD laboratory recruitment system.

- -

Common to Both Treatments

The statement we showed subjects in the previous study was:

"**People who immigrated to the U.S. illegally, when caught, should be deported and sent back to their countries of origin**".

Among students in the previous study who self-identify as **Independents** or **Republicans**, what **percentage** reported **agreeing** with the statement above?

Please enter a number between 0 and 100 and do not include the % sign.

____

Among students in the previous study who self-identify as **Democrats**, what **percentage** reported **agreeing** with the statement above?

Please enter a number between 0 and 100 and do not include the % sign.

____

- -

Private-First Treatment Only

In the previous study, we actually randomized participants into **two groups**: a **Private-Answer Group** and a **Public-Answer Group**. Participants in the two groups were given **different** instructions that corresponded to different experimental conditions.

So far, we have shown you the instructions given to participants in the **Private-Answer Group**. As a reminder, the instructions for participants in the **Private-Answer Group** were as follows:

"We will start by asking you a set of questions about socio-political topics that have sparked public debates in recent years.

IMPORTANT: We will share aggregate-level answers to the questionnaire on the next screen with approximately 200 UCSD 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.**”

Conversely, the instructions given to participants in the **Public-Answer Group** were as follows:

“We will start by asking you a set of questions about socio-political topics that have sparked public debates in recent years.

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 UCSD 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 UCSD laboratory recruitment system.**

According to the UCSD laboratory recruitment system, your first and last name are:

First Name: [*Participant's First Name*]
Last Name: [*Participant's Last Name*]"

- -

Public-First Treatment Only

In the previous study, we actually randomized participants into **two groups**: a **Private-Answer Group** and a **Public-Answer Group**. Participants in the two groups were given **different** instructions that corresponded to different experimental conditions.

So far, we have shown you the instructions given to participants in the **Public-Answer Group**. As a reminder, the instructions for participants in the **Public-Answer Group** were as follows:

“We will start by asking you a set of questions about socio-political topics that have sparked public debates in recent years.

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 UCSD 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 UCSD laboratory recruitment system.**

According to the UCSD laboratory recruitment system, your first and last name are:

First Name: [*Participant's First Name*]
Last Name: [*Participant's Last Name*]"

Conversely, the instructions given to participants in the **Private-Answer Group** were as follows:

“We will start by asking you a set of questions about socio-political topics that have sparked public debates in recent years.

IMPORTANT: We will share **aggregate-level answers** to the questionnaire on the next screen with approximately 200 UCSD 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.**"

- -

Common to Both Treatments

As mentioned on the previous screen, the questions you were asked were about the behavior of participants in the **[Private]/[Public]-Answer Group**.

We will now ask you about the behavior of participants in the **[Public]/[Private]-Answer Group**.

Recall that the statement that participants in the previous study had to report their level of agreement with was:

"***People who immigrated to the U.S. illegally, when caught, should be deported and sent back to their countries of origin***".

When we asked you to guess the percentage of self-identified Independents or Republicans in the **[Private]/[Public]-Answer Group** up of the previous study who reported agreeing with the statement above, your answer was **[answer]**.

What do you think that percentage is for self-identified Independents or Republicans in the **[Public]/[Private]-Answer Group** of the previous study?

Please enter a number between 0 and 100 and do not include the % sign.

____

When we asked you to guess the percentage of self-identified Democrats in the **[Private]/[Public]-Answer Group** of the previous study who reported agreeing with the statement above, your answer was **[answer]**.

What do you think that percentage is for self-identified Democrats in the **[Public]/[Private]-Answer Group** of the previous study?

Please enter a number between 0 and 100 and do not include the % sign.

____

Participants in the previous study were asked to report their level of agreement not just with the statement about immigration, but with 14 other statements.

Starting on the next screen, we will show you the actual distribution of answers of participants in the **Public-Answer Group** and ask you to predict the average answers of participants in the **Private-Answer Group**.

Consider the following question:

*"On a scale from 0 to 10, to what extent do you agree with the following statement: '**All statues and memorials of Confederate leaders should be removed'**?"*

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

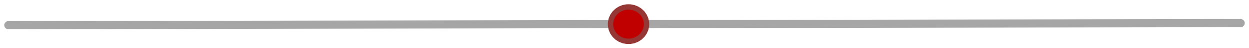
[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **6.2**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0 1 2 3 4 5 6 7 8 9 10



Consider the following question:

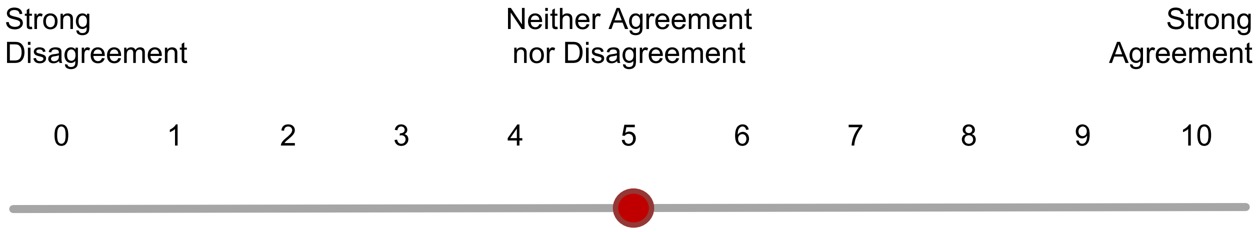
"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**Defunding the police is a bad idea because it will inevitably lead to increased crime rates**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **5.6**.

What do you think the average answer of participants in the **Private-Answer Group** was?



Participants in the previous study were informed that a "**preferred gender pronoun**" is defined as the personal pronoun (he, she, they, zhe) that a person chooses to use to identify that person's gender.

Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**The UCSD administration should require professors to address students according to the students' preferred gender pronouns**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **6.7**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement					Neither Agreement nor Disagreement					Strong Agreement
0	1	2	3	4	5	6	7	8	9	10

Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**Transgender women (i.e., individuals who were male at birth and transitioned later in life) should be allowed to participate in women's sports**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

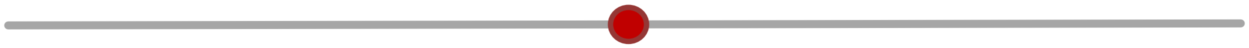
[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **4.2**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0 1 2 3 4 5 6 7 8 9 10



Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**The UCSD administration should require professors to use trigger warnings in their classes**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

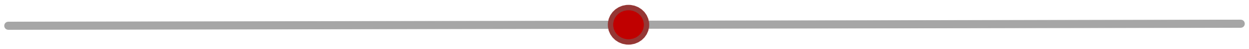
[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **5.8**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0 1 2 3 4 5 6 7 8 9 10



Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**Sexual harassment training should be mandatory for everybody who works or studies at UCSD**?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **8.2**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0 1 2 3 4 5 6 7 8 9 10

●

Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**People who immigrated to the U.S. illegally, when caught, should be deported and sent back to their countries of origin**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

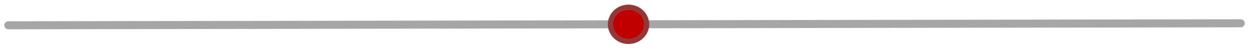
[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **3.8**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0 1 2 3 4 5 6 7 8 9 10



Participants in the previous study were informed that "**reparations for slavery**" refers to the idea that some form of compensatory payment needs to be made to the descendants of Africans trafficked to and enslaved in the Americas as a consequence of the Atlantic slave trade.

Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**The U.S. government should provide reparations for slavery**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **5.6**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0 1 2 3 4 5 6 7 8 9 10

●

Participants in the previous study were informed that "**racial microaggressions**" are defined as brief and commonplace daily verbal, behavioral, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults toward people of color.

Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**Racial microaggressions are an important problem at UCSD'***?"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **6.2**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement					Neither Agreement nor Disagreement					Strong Agreement
0	1	2	3	4	5	6	7	8	9	10
					●					

Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**The UCSD administration should allow students to wear blackface for Halloween**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

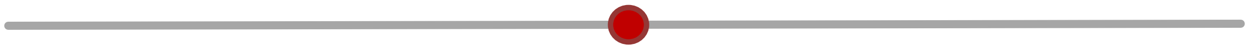
[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **1.7**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0	1	2	3	4	5	6	7	8	9	10



Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**Parents should limit the amount of time their kids spend on their smartphones**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **7.5**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement					Neither Agreement nor Disagreement					Strong Agreement
0	1	2	3	4	5	6	7	8	9	10

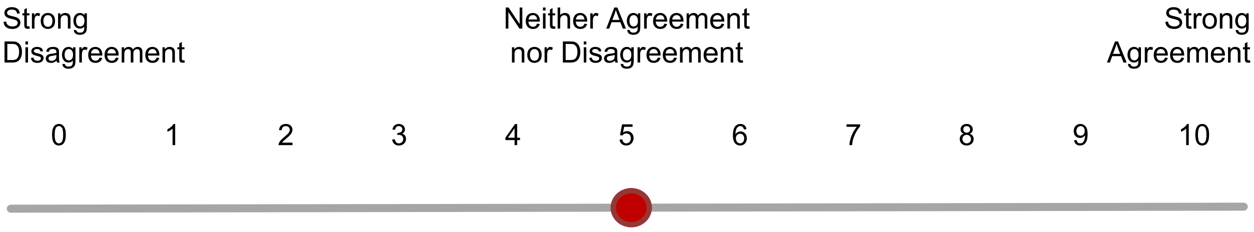
Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**The United States should increase tariffs on foreign imports**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]
The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **4.4**.

What do you think the average answer of participants in the **Private-Answer Group** was?



Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**School uniforms help reduce clothing-related peer pressure**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **5.0**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0 1 2 3 4 5 6 7 8 9 10

Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**The one-cent coin (i.e. the penny) should be removed from circulation'***?"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **5.3**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong
Disagreement | Neither Agreement
nor Disagreement | Strong
Agreement

Strong
Disagreement

Neither Agreement
nor Disagreement

Strong
Agreement

0 1 2 3 4 5 6 7 8 9 10

●

Consider the following question:

"*On a scale from 0 to 10, to what extent do you agree with the following statement: '**The members states of the European Union should cede more powers to the E.U.**'?*"

The distribution of answers of participants in the **Public-Answer Group** is shown in the histogram below (the histogram shows the frequency of each answer on the 0-10 scale).

[Histogram]

The **average** answer of participants in the **Public-Answer Group**, marked by the red dashed line on the graph, was **4.8**.

What do you think the average answer of participants in the **Private-Answer Group** was?

Strong Disagreement	Neither Agreement nor Disagreement	Strong Agreement

0 1 2 3 4 5 6 7 8 9 10

●

In the previous study we hypothesized that, for some of the statements, the average level of agreement of participants assigned to the Private-Answer Group would be significantly different from the average level of agreement of participants assigned to Public-Answer Group.

From now on, we will refer to the **difference** between the **average answers** given by participants in the **Private-Answer Group** and the **average answers** given by participants in the **Public-Answer Group** as the **Private-Public difference**.

Specifically, the statements where we hypothesized that the average level of agreement of participants assigned to the Private-Answer Group and the Public-Answer Group would differ are the following:

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 UCSD 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 UCSD 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 UCSD.
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 UCSD.
10. The UCSD administration should allow students to wear blackface for Halloween.

As you can imagine, the **Private-Public difference** in responses across the **10 statements** above may be **larger for certain demographics than for others**. For instance, it may be the case that the Private-Public difference is larger for students who reported being religious than for students who reported being non-religious.

In this section of the survey we will ask you to **guess whether certain demographic groups exhibited larger Private-Public differences than others**.

If the Private-Public difference between two demographic groups is large enough to be statistically significant (i.e. ~ larger than 0.5 units on the 0-10 scale), you will get points for guessing the group with the larger difference. If the Private-Public difference between two groups is not statistically significant, you will get points if you guess that the difference is about the same for the two groups.

For your information, **there is at least one set of demographics for which the Private-Public difference is statistically significantly different.**

A quick question to make sure you understood what we are asking.

Suppose we consider individuals who reported being religious and individuals who reported being non-religious. Suppose the private- public difference for religious individuals is 1 point on the scale and the private- public difference for non-religious individuals in 0.2 points on the scale. Then:

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **larger** for participants who reported being **religious** than for participants who reported being **non-**religious.

○ The Private-Public difference in average responses across the 10 sensitive statements was **about the same** for participants who reported being **religious** than for participants who reported being **non-**religious.

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **smaller** for participants who reported being **religious** than for participants who reported being **non-**religious.

Now let's move on to the actual questions.

Consider participants who identify as **male** and participants who identify as **female**.

Please select the answer that you think is correct:

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **larger** for participants who identify as **male** than for participants who identify as **female**.

○ The Private-Public difference in average responses across the 10 sensitive statements was **about the same** for participants who identify as **male** as for participants who identify as **female**.

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **smaller** for participants who identify as **male** than for participants who identify as **female**.

Consider participants who identify as **white** and participants who identify as **non-white**.

As a reminder, the composition of participants in the previous study who identify as non-white is as follows: around 3% of them identify as Black or African-American, around 70% of them identify as Asian or Asian-American, around 25% of them identify as Latino or Chicano, and the remaining 2% of them identify as some other race/ethnicity (e.g. Native American).

Please select the answer that you think is correct:

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **larger** for participants who identify as **white** than for participants who identify as **non-white**.

○ The Private-Public difference in average responses across the 10 sensitive statements was **about the same** for participants who identify as **white** as for participants who identify as **non-white**.

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **smaller** for participants who identify as **white** than for participants who identify as **non-white**.

Consider participants who are classified as **Democrats** and participants who are classified as **Independents** or **Republicans**.

Please select the answer that you think is correct:

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **larger** for participants who are classified as **Democrats** than for participants who are classified as **Independents or Republicans**.

○ The Private-Public difference in average responses across the 10 sensitive statements was **about the same** for participants who are classified as **Democrats** as for participants who are classified as **Independents or Republicans**.

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **smaller** for participants who are classified as **Democrats** than for participants who are classified as **Independents or Republicans**.

Consider participants who major in the **humanities** or **social sciences** and participants who major in the **sciences** or in **engineering**.

Please select the answer that you think is correct:

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **larger** for participants who major in the **humanities or social sciences** than for participants who major in the **sciences or in engineering**.

○ The Private-Public difference in average responses across the 10 sensitive statements was **about the same** for participants who major in the **humanities or social sciences** as for participants who major in the **sciences or in engineering**.

○ The Private-Public difference in average responses across the 10 sensitive statements was significantly **smaller** for participants who major in the **humanities or social sciences** than for participants who major in the **sciences or in engineering**.

Of the 5 multiple-choice questions from the previous screen, how many do you think you answered correctly?

○ 0

○ 1

○ 2

○ 3

○ 4

○ 5