Acceptable Discourse: Social Norms of Beliefs and Opinions

Russell Golman*

April 7, 2021

Abstract

This paper develops a theory of social norms of beliefs and opinions, which provides an account of political correctness and the backlash against it. In the model, social norms about opinion expression emerge as equilibria of a signaling game in which expressing an unpopular opinion leads to bad judgments about one's values, but may also be attributed to one's factual beliefs. We find that multiple equilibria may co-exist, corresponding to norms with more or less conformity and social pressure. Additionally, motivated reasoning and persuasion allow norms to influence privately held opinions and underlying factual beliefs. The theory helps us understand normative social influence on beliefs and identity-protective cognition, for example, when norms against criticizing one's country lead to idealized beliefs about the country or when communities with stronger norms of political correctness keep a lid on racist opinions, yet believe that racism is more prevalent.

KEYWORDS: Beliefs; Opinions; Political Correctness; Signaling; Social Norms

*Carnegie Mellon University, Department of Social & Decision Sciences, Pittsburgh, PA 15217, rgolman@andrew.cmu.edu. Thanks to Gordon Brown, Rachel Kranton, George Loewenstein, John Miller, Peter Schwardmann, Daniel Stone, and Joël van der Weele for helpful comments on earlier drafts, and to Benjamin Williams for help editing the paper.

1 Introduction

Some beliefs and opinions are taboo. Few people are comfortable advocating Nazism, arguing for eugenics, or expressing a belief in racial differences in intelligence. The beliefs and opinions that are socially acceptable, or that are stigmatized, can depend on one's community (Sunstein, 2018). In progressive circles, for instance, questioning affirmative action is seen as politically incorrect, and people are reluctant to do so publicly, even though they may harbor doubts about it in private (van Boven, 2000). The topic is fraught because people do not want to be perceived as a racist, so expressing an opinion about affirmative action requires an understanding of social norms. For someone who believes that affirmative action will help promote equal opportunity, expressing support is both consistent with his personal view and safe from social judgement. But for someone who believes that affirmative action will create inequity, the desire to express an *honest* opinion conflicts with the desire for social approval. Navigating this tradeoff is complex because his social image depends on the extent to which opposition to affirmative action would be attributed to a person's (possibly) honest belief that the policy would create inequity or to a person's (possibly) racist values (Loury, 1995; Kuran, 1997). Thus, a person's comfort with expressing an unpopular opinion about affirmative action depends on other people's willingness to do the same. The social norm in some communities may induce pressure to conform by adopting the politically correct opinion, while in others it may be more tolerant of disagreement. How do these social norms emerge?

Social norms facilitating coordination and cooperation arise naturally in contexts in which people care about each other's behavior, e.g., in the workplace (Akerlof, 1980; Elster, 1989), but it is less clear why people care about each other's beliefs and opinions and why social norms regulating them would emerge. Disagreements about the best way to butter bread do not typically lead to bitter disputes (Seuss, 1984). Yet proponents and opponents of affirmative action view each other with hostility, and dismiss the other side's views as racist or as cowardly, politically correct (Sherman et al., 2003). There is widening recognition that shared beliefs and opinions help people forge social identity (Turner et al., 1987; Akerlof and Kranton, 2000; Bénabou and Tirole, 2011; Kahan, 2015). Accordingly, people express particular beliefs and opinions as badges of identity. Still, the question remains why beliefs and opinions about issues like affirmative action are such an integral part of identity. A person's identity could be determined by personal characteristics or values (e.g., a hard-worker or an intelligent person) or by group membership, without beliefs get-

ting in the way.[1] To account for the passions stirred by beliefs and opinions and for the emergence of social norms regulating them, this paper develops a theory of social norms of belief and opinions in which personal values are fundamental to identity and argues that beliefs and opinions matter because they are informative signals about a person's values. The theory holds that social norms governing opinion expression emerge as equilibria of this signaling game. By characterizing these equilibria and the beliefs that support them, the theory helps us understand how social norms affect opinion expression and may even shape private opinions and factual beliefs that exist only in a person's mind.

The paper proposes that social pressure to express popular opinions arises because opinions are informative signals about a person's values. Values, defined as a set of characteristics that affect opinions and preferences, are judged by others, so people care about how their values are perceived by others. In some cases, a person may receive favorable treatment or direct material benefit from being seen to have the "right" values. In other cases, a person may care about social approval for his values intrinsically or because they help construct a desired social identity. In any case, people derive utility from esteem for their perceived values (i.e, "reputational utility"). Concomitantly, we assume that people also care about authenticity and thus derive utility from expressing opinions that are consistent with their private views (i.e., "expressive utility"). However, we assume no direct effect of holding or expressing particular factual beliefs on utility. Instead, preferences about factual beliefs arise endogenously because expressing an opinion sends a signal about a person's values and the informativeness of this signal depends on the person's beliefs. As stable equilibria of this signaling game, social norms about the expression of opinions and beliefs and their implications about a person's values emerge endogenously.

In the model we develop, bad judgements about one's values result from expressing an unpopular opinion – i.e., violating the social norm. There is thus social pressure to espouse particular opinions. Moreover, particular beliefs can be offered as a rationale for holding an unpopular opinion, to mitigate the social stigma, but these beliefs themselves become associated with social stigma. For example, nativists claim that undocumented immigrants commit crimes to justify their anti-immigrant opinions. Indeed, Bursztyn et al. (2020b) find that people are more likely to publicly contribute to a campaign to "Fund the Wall" when evidence that they believe that undocumented immigrants commit crimes is made public, and this implicit justification does in fact mitigate social stigma. Recognizing

[1]Think of beliefs as judgments (of probability) about an objective state of the world, and opinions as value-laden judgments that are not objectively right or wrong.

social pressure to adopt and express particular beliefs and opinions, our theory is consistent with self-categorization theory and the theory of identity-protective cognition, according to which, people report opinions and beliefs that will help them fit in with the groups with which they want to identify (Turner et al., 1987; Abrams et al., 1990; Cohen, 2003; Kahan, 2013; Kahan, 2015; Kahan, 2017). However, instead of assuming that particular opinions and beliefs are necessary to establish an identity, our theory predicts that the desirable beliefs and opinions are specifically those that allow a person to signal desirable values. This is why belief in racial differences in intelligence is stigmatized, whereas differences in beliefs about, say, the safety of nuclear power plants are more easily tolerated.

The model allows for the existence of multiple equilibria, i.e., multiple possible social norms, in the case that desirable beliefs and desirable values are complements (but not in the case that they are substitutes). In a pooling equilibrium, social pressure to conform to the norm of acceptable discourse is sufficiently strong for everybody to do so. In a semi-separating equilibrium, people who are sufficiently confident in their beliefs are willing to express unpopular opinions, and it may not be possible to determine whether an unpopular opinion should be attributed to a person's values or underlying factual beliefs. If we think of equilibrium selection as subject to exogenous variation, then the social norm is not merely *descriptive*, but also *determinative* of social pressure to conform to politically correct standards. Thus, by attending to multiple equilibria, the theory accounts for empirical findings that manipulation of social norms of acceptable discourse affects willingness to express unpopular opinions (Higgins and McCann, 1984; Wood et al., 1996; Cohen, 2003; Masser and Phillips, 2003; Bursztyn et al., 2020a), as does manipulation of the relative importance of reputational utility and expressive utility (Ybarra and Trafimow, 1998). Additionally, the theory predicts that conformity pressure arises when desirable beliefs and desirable values are complements, but not when they are substitutes.

The primary contribution of the theory is the characterization of norms of opinion expression for a population with heterogenous values and subjective beliefs. The basic signaling model giving rise to these norms has two distinctive features. First, an individual's type consists of two attributes (his values and his beliefs) and the incentives prevent everyone from fully revealing their types, so the signaling game involves a signal extraction problem. This signal extraction problem is the reason why expressed beliefs become associated with a person's values. According to the theory, in situations in which a person's values are already common knowledge, the person would feel comfortable expressing unpopular beliefs and opinions. A second distinctive feature of the basic model is that there are multiple

people signaling rather than just a single informed expert. That is, the entire society is participating in the signaling game. This means that other people's signaling strategies affect audience attributions and thus affect one's own signaling strategy. Thus, this feature of the model is critical to account for social pressure to conform to politically correct discourse.

The basic signaling model presented in Section 2 relies on fixed, subjective beliefs, which are taken as given. In Section 3 we then consider the belief formation process, closing the feedback loop between beliefs and social norms. To account for the divided beliefs that are a necessary part of semi-separating equilibria and to understand how social norms might influence private beliefs, motivated reasoning and persuasion are incorporated into the theory here. This allows preferences to affect beliefs. However, people who engage in motivated reasoning cannot simply choose to believe anything they want to; rather, motivated beliefs are constrained by a need to construct justification for the desired beliefs (Kunda, 1990; Epley and Gilovich, 2016). Similarly, one view of persuasion is that people adopt models of how the world works from each other but are constrained by the need for the adopted model to account for the observed data and by the use of Bayes' rule to form beliefs about states of the world (Schwartzstein and Sunderam, 2019). Following this literature, we consider persuasion about model parameters (i.e., about the prevalence of desirable values) with assumptions that it must account for the observed pattern of opinion expression to be persuasive and that belief about the state of the world must be rationalized by Bayes' rule. Accordingly, we consider motivated reasoning to be a process of persuading oneself (Mercier and Sperber, 2011).

The assumption that people engage in motivated reasoning and persuasion implies that social pressure affects privately held opinions and factual beliefs, not just publicly expressed opinions. The motive to engage in persuasion arises for a person expressing an unpopular opinion. He will want the audience to attribute his deviant opinions to his factual beliefs, not his values, and he will try to persuade his audience that desirable values are held universally and thus his opinion is illustrative only of his beliefs. Preferences about one's own beliefs, and thus motivated reasoning, arise because people want to hold the beliefs that will permit them to express socially desirable opinions, i.e., opinions that reveal desirable values (as well as because they want higher confidence in the opinions they choose to express). When others' beliefs conflict with the beliefs one is promoting, they are a threat (Bénabou and Tirole, 2011; Golman et al., 2016). As a result, a persuasive narrative constructs a reason to dismiss opposition to it, which is consistent with empirical literature on naive realism and group polarization (e.g., Ross and Ward, 1996; Taber

and Lodge, 2006; Iyengar et al., 2012; Iyengar and Westwood, 2015; Marks et al., 2019). Beliefs about model parameters may become correlated with beliefs about whose opinions are informative and about the state of the world, consistent with the empirical finding that people adopt motivated assumptions about the informativeness of others' beliefs (Oprea and Yuksel, 2020). We thus account for polarization of identity-relevant factual beliefs.

By considering social norms of opinion expression to be equilibria of a signaling game, the model captures a variety of familiar and important phenomena including political correctness (i.e., social pressure to express socially desirable opinions and the dismissal of alternative opinions) along with the contrarian backlash against it. In addition, the model offers unique insights, in the form of new testable predictions. For example, in the case that desirable values and desirable beliefs are complements (but not when they are substitutes) the model predicts that when other people express a socially desirable opinion, the social pressure to conform and express the same opinion increases. Also, perhaps counterintuitively, when more people have the socially desirable values, the social pressure to express the same politically correct opinions as almost everyone else actually decreases. That is, we should expect to observe less tolerance for dissenting opinions when a larger segment of the population is perceived to have questionable values. By incorporating motivated reasoning and persuasion, the model also predicts that beliefs about states of the world that support socially desirable opinions become correlated with negative judgments about other people who do not share those socially desirable opinions.

1.1 Applications

One of the most obvious applications of the theory concerns social norms about opinions that are possibly indicative of racism and the associated backlash against political correctness by those who nevertheless express these opinions. The theory can help us understand the factors that determine whether people will feel comfortable expressing such politically incorrect opinions. It also offers an explanation for the persistence of disagreements about factual beliefs which also become tinged by racism and political correctness, e.g., beliefs about whether minorities suffer discrimination, the prevalence of racism, and the degree to which political correctness restricts discourse. To illustrate these insights, let us return to the example introduced earlier concerning opinions about affirmative action. Suppose that some people have racist values (e.g., think that one race should have more power or advantages than another), while others are (racially) egalitarian, but nobody wants to be seen as racist (at least with an audience that is believed to be sufficiently egalitarian). Expressing a

particular opinion on a racially charged topic, such as affirmative action, may send a signal about one's values. Yet opinions about affirmative action may also reflect factual beliefs about the world, such as how likely it is that minorities suffer discrimination and, in turn, whether racial disparities are due to inequity in opportunity. (Rather than defining such beliefs as racist, our theory offers an explanation about why beliefs like these, which are in principle about factual matters, come to be associated with one's values.) A racist would be opposed (privately, at least) to affirmative action, regardless of his beliefs. However, an egalitarian's private opinion would depend on his beliefs – if he believed that minorities do not suffer discrimination, he might oppose affirmative action to avoid creating inequity, whereas if he believed that they do suffer discrimination, he might support affirmative action to mitigate existing inequity (Harrison et al., 2006). That is, belief that minorities suffer discrimination complements egalitarian values in producing support for affirmative action. Thus, opposition to affirmative action could be attributed to racism (values) or doubt about the reality of racial discrimination (belief about the state of the world).

For someone who is opposed to affirmative action, willingness to express that opinion would depend on how much the person cares about appearing to (possibly) be racist compared to how much he cares about expressing his authentic opinion. It also depends on how confident he is in his beliefs (assuming that the authenticity motive is stronger when a person is more confident). And, most interestingly, it depends on the emergent social norm, i.e., how many other people are willing to express their opposition (which affects the attribution of whether opposition to affirmative action is more likely to be due to racist values or honest beliefs about the state of the world). Multiple equilibria may exist: in one, everybody may feel sufficient social pressure to express support for affirmative action, regardless of their private opinion about it; in others, egalitarians who confidently believe that minorities no longer suffer racial discrimination express opposition to affirmative action, as do racists who appear indistinguishable from them. In this semi-separating equilibrium, racists are thus motivated to express the belief that minorities no longer suffer racial discrimination (whether honestly held or not), as a relatively innocuous explanation for their opposition to affirmative action.

Moreover, motivated reasoning and persuasion can preserve disagreement about the truth. An egalitarian who has plausible reasons to believe that minorities suffer discrimination, and who thus can support affirmative action, may try to dismiss any information that would lead him to a politically incorrect view and thus raise questions about his values. He may even convince himself that anybody who disagrees with him must be racist, as a way

of bolstering his own desired belief. On the other hand, egalitarians who remain sufficiently confident that minorities no longer suffer discrimination, and who thus oppose affirmative action, may argue that racism does not even exist (i.e., that the debate is entirely on the substance) and that only the oppressive environment of political correctness stifles agreement with their beliefs about the absence of racial discrimination. Perplexingly, racists will make the same arguments. Their backlash against political correctness lets them dismiss the beliefs of people who support affirmative action, as they seek to ameliorate the stigma attached to their own opinions. Still, these opinions, and the beliefs that support them, are indeed correlated with racism.

The theory can be applied to other beliefs and opinions as well. For example, consider the widespread political support within the United States for the invasion of Iraq in 2003. This opinion was based on the belief that Saddam Hussein had weapons of mass destruction, along with, presumably, the proponent's patriotic values. Opposition to the war could have been attributed to skepticism that there really were such weapons or to being anti-American. Fear of being seen as unpatriotic made it difficult for many people to question the existence of the weapons and to oppose the war, and those who did question the existence of the weapons were indeed characterized as anti-American by supporters of the war. Discounting the beliefs of anybody who opposed the war, supporters of the war believed there was a consensus that the weapons existed, which strengthened their support for the war and thus made it possible to clearly signal their patriotism. The same logic applies broadly to reluctance to criticize one's country and adoption of idealized beliefs in support of nationalistic opinions that signal patriotic values (Herrmann, 2017).

1.2 Related Literature

Social norms of opinion expression, as illustrated by the above examples, can be seen through the broader lens of social norms governing most social and economic behavior (Elster, 1989; Bicchieri, 2006; Akerlof, 2007). In a wide variety of situations, individuals find it in their best interest to conform to social norms (Young, 2015) because of incentives for coordination (Lewis, 1969; Sugden, 1989), social sanctions for noncompliance (Akerlof, 1980; Ostrom, 2000), or personal identity (Akerlof and Kranton, 2000). As it is in people's best interest to conform, norms of behavior are self-perpetuating. That is, social norms emerge as stable equilibria of coordination games (Young, 1993), repeated games with opportunities for punishment (Kandori, 1992), and signaling games (Bernheim, 1994).

The signaling model we develop builds on Bernheim's (1994) model of endogenous

social norms in a signaling game, but here, the interpretation of the signal (in this case, a publicly expressed opinion) depends on a person's beliefs, so there is an additional signal extraction problem. (Austen-Smith and Fryer (2005) analyze a different signaling game involving a signal extraction problem.) Moreover, here, motivated reasoning and persuasion arising from the motive to send a socially desirable signal impact people's beliefs, and thus affect the informational content of the signals people are sending. If social norms shape deep-seated beliefs, and not just public discourse, they are far more consequential.

The coexistence of multiple equilibria and the collective dynamics, in which some people expressing socially desirable opinions increases pressure on others to hold and express them as well, resembles Bénabou's (2008; 2013) models of groupthink and ideology. But, here, the motives are reputational and expressive utility (instead of anticipatory and material utility). Our model thus describes different phenomena, e.g., political correctness.

This paper shares and formally develops Loury's (1994) insight that political correctness arises from people self-censoring their opinions to avoid signaling the wrong values, and offers additional insights about how polarized opinions may persist despite being subject to social influence. It also shares and formally develops Kuran's (1997) insight that "preference falsification," the divergence of public opinions from private opinions, results from a tradeoff between reputational utility and expressive utility. The formal modeling here helps us understand when social norms with strong pressure to be politically correct will arise, and how they may then affect private beliefs and opinions.

Morris (2001) also models political correctness as an equilibrium of a signaling game, but considers signaling by a single informed expert instead of by multiple members of a community, and assumes desire for influence and instrumental reputational concerns, rather than desire for esteem and authenticity. Our theory applies more broadly, describing a society of people seeking social approval, rather than covering only experts seeking to remain influential. Applying in different situations, our model makes different predictions. It offers the new prediction that normative social influence does not shape all beliefs and opinions, but specifically puts pressure on opinions that might reveal undesirable values, as well as beliefs that could be used to explain away such opinions. Further, it predicts that when the social norm is more tolerant of such beliefs and opinions being expressed, more people adopt them (privately), and holding and expressing these views correlates with arguing that the undesirable values are not so prevalent in the population. We thus also account for the backlash against political correctness.

This paper also fits into a wider literature describing social influence on beliefs and

opinions. Ortoleva and Snowberg (2015) and Brown et al. (2020), like this paper, predict that more confident individuals are more willing to express unpopular opinions, but they do not model opinion expression as a signaling game. Social influence arising from a signaling game is distinct from (but not mutually exclusive with) direct influence (Murphy and Shleifer, 2004), informational cascades or herding (Banerjee, 1992; Bikhchandani et al., 1992; Acemoglu et al., 2011), and persuasion bias (DeMarzo et al., 2003).

2 Basic Model

Individuals choose an opinion $x \in \{0, 1\}$ to express publicly. Each person has a private opinion about this issue, which may be held with more or less confidence. This private opinion depends on the person's beliefs about the state of world $\omega \in \Omega = \{\text{True}, \text{False}\}$ and the person's values $v \in V = \{V^+, V^-\}$. The relationship of a person's true opinion to the state of the world and the person's values is described by a function $t : \Omega \times V \to \{0, 1\}$, so that confidence $q = \Pr(t(\omega, v) = 1)$ about one's own private opinion may depend on the probability one assigns to a state of the world $p = \Pr(\omega = \text{True})$. The function t could require both a particular state of the world *and* the particular values $V+$ to map into a particular opinion (in which case we say that the belief complements these values in producing this opinion) or could yield this opinion if the person either believes in a particular state of the world *or* holds the particular values $V+$ (in which case we say that the belief substitutes for these values in producing the opinion).

To illustrate the distinction between confidence in an opinion (which may depend on a person's values) and certainty in a belief about a factual matter (which should only depend on information about the state of the world), recall the example of beliefs and opinions about affirmative action. Whether affirmative action is good policy is an opinion; it depends on whether a person believes it alleviates or institutionalizes racial inequality (a belief about the state of the world) as well as whether the person considers racial equality to be a good thing (a matter of personal values). A person with egalitarian values might have high or low confidence q that it is good policy based on whether he assigns high or low probability p to the state of the world that it alleviates inequality. On the other hand, a racist might be sure of his opposition (with $q = 0$) regardless of his belief p about the state of the world, solely due to his values.[2] In this case, belief that affirmative action alleviates racial inequality complements egalitarian values in producing support for affirmative action.

[2] In Section 3 we will recognize that the belief that affirmative action is a source of racial inequality could itself be racist, but that requires a departure from Bayesianism.

People care about esteem, i.e., how they're perceived by others (possibly a specific audience with high status, whose approval may be either intrinsically desirable or materially desirable simply due to favors that result from ingratiation (Bursztyn and Jensen, 2017)). We assume that esteem depends on being perceived to have the right values (according to the relevant audience). Suppose that there is a universal preference to be seen to have $v = V^+$. (The case that people want to be seen to have their actual values is less interesting because then they could simply reveal their values with credible cheap talk.) Let $b(x)$ denote the perceived probability that someone who expresses x has values V^+, i.e., $b(x) = \Pr(v = V^+|x)$. We assume that reputational utility is $g(b(x))$ for some increasing, continuous function g. (If there is heterogeneity in audience members' perceptions, reputational utility becomes the expected value of $g(b(x))$ across the audience.)

People also care about authenticity, i.e., about expressing opinions that are consistent with their private opinions (Kuran, 1997). This concern is what tethers expressed opinions to individuals' types (i.e., to their actual values and their beliefs) in the model, so that expressed opinions are not simply cheap talk. Based on the empirical finding that people are more comfortable expressing unpopular opinions when they are more confident in these opinions (Lasora, 1991; Matthes et al., 2010), we assume that expressive utility is $h(q)$ if $x = 1$ and $h(1 - q)$ if $x = 0$, for some increasing, continuous function h, for a person with confidence q that his true private opinion is $t = 1$. For compactness, we can rewrite expressive utility as $h(q)x + h(1 - q)(1 - x)$.

Putting reputational and expressive utility together, a person's utility function is

$$U(x) = g(b(x)) + \lambda \left(h(q)x + h(1 - q)(1 - x) \right), \tag{1}$$

where $\lambda > 0$ is a parameter that determines the magnitude of expressive utility relative to reputational utility. The parameter λ could of course be pulled into the function h, but we introduce it here to lay the groundwork for a comparative static analysis about the effect of caring more about expressive utility relative to reputational utility.

Individuals differ in their values v and their belief p about the state of the world. To begin, we will assume common knowledge about the distribution of values and beliefs in the population and take individuals' beliefs to be fixed and given. In Section 3 we will consider disagreement about the population distribution, consistency of beliefs about the state of the world and beliefs about the population, and persistent disagreement about the state. Let α denote the fraction of the population with values $v = V^+$, and let $F^+(p)$ and

$F^-(p)$ be the cumulative distribution functions for beliefs about the state of the world for people with values V^+ and V^- respectively.

The interaction of beliefs and values in determining a person's private opinion, and thus shaping his expressed opinion, causes beliefs to carry utility in equilibrium, even though they do not enter the utility function directly. This feature of the model is thus critical to our account of when and why factual beliefs can be identity relevant.

2.1 Case I: Beliefs and Values Are Complements

Consider first the case that beliefs and values are complements, meaning that $t(\omega, v) = 1$ if and only if $\omega = \text{True}$ and $v = V^+$. This case is the setting for all of the applications we discussed earlier. In this case, a person's confidence about his private opinion depends on his belief about the state of the world only if he has the desirable values, i.e., $q = pI^+(v)$ where I^+ is an indicator for having values V^+. Given the alignment of the private opinion $t = 1$ with the desirable values V^+, we occasionally refer to this opinion as the desirable opinion and $t = 0$ as the undesirable opinion.

With beliefs and values interacting as complements, the signal extraction problem arises when expressing the unpopular opinion. People inclined to express the unpopular opinion because of their values may try to justify it by hiding behind their information and beliefs. The audience's inferences will naturally follow Bayes' rule.

We examine the equilibria of the signaling game in this case. In any equilibrium, beliefs must be consistent with the distribution of strategic choices. However, in a pooling equilibrium, we may need to specify beliefs that would hold in non-equilibrium scenarios. We apply an equilibrium refinement (Cho and Kreps, 1987) to rule out a pooling equilibrium in which everybody chooses $x = 0$ because it seems intuitive that an individual with $p > .5$ and $v = V^+$ could deviate and gain expressive utility with no loss in reputational utility. On the other hand, we do consider the possibility of a pooling equilibrium in which everybody chooses $x = 1$, enforced by the belief that choosing $x = 0$ would reveal oneself to have values $v = V^-$ (because individuals with such values would have the most to gain in expressive utility from this deviation). In this pooling equilibrium, $b(0) = 0$ and $b(1) = \alpha$. The incentive constraint for an individual with $v = V^-$ (the individual with the strongest incentive to deviate) is

$$g(\alpha) - g(0) \geq \lambda \left(h(1) - h(0) \right). \tag{2}$$

Proposition 1 *If Equation (2) holds, then there exists a pooling equilibrium in which everybody chooses* $x = 1$.

Proposition 1 states that when reputational utility is sufficiently strong relative to expressive utility, there can be an equilibrium with a strong social norm about the kind of opinion that is acceptable to express in public. In this case, everybody conforms to the norm, even if they do not privately share this opinion.

There may also be a semi-separating equilibrium in which the opinion an individual chooses to express publicly depends on his private opinion. This equilibrium must take the form of a threshold for sufficient confidence in the undesirable opinion that permits a person to publicly express it.

Proposition 2 *In any semi-separating equilibrium, there exists a* $p^* \leq .5$ *such that* $x = 0$ *if* $q < p^*$ *and* $x = 1$ *if* $q > p^*$.

Proposition 2 establishes that semi-separating equilibria involve a threshold of sufficient confidence to express an opinion because the incentive to express an opinion is increasing in the confidence a person has in it. In a semi-separating equilibrium, anybody who chooses $x = 1$ then reveals themselves to have values V^+, i.e., $b(1) = 1$. However, individuals with beliefs $p < p^*$ or with values V^- pool together with $x = 0$, so the Bayesian inference after observing $x = 0$ is

$$b(0) = \frac{\alpha F^+(p^*)}{1 - \alpha + \alpha F^+(p^*)}. \tag{3}$$

Incentive compatibility implies indifference when $q = p^*$, i.e.,

$$g(1) - g\left(\frac{\alpha F^+(p^*)}{1 - \alpha + \alpha F^+(p^*)}\right) = \lambda \left(h(1 - p^*) - h(p^*)\right). \tag{4}$$

Proposition 3 *Any solution* p^* *to Equation 4 defines an equilibrium in which people who are sufficiently confident in the undesirable opinion (with* $q \leq p^*$*) choose to express it (i.e., choose* $x = 0$*), and people who are less sure of their own opinion as well as those who are confident in the desirable opinion (with* $q > p^*$*) choose to express this socially desirable view (i.e., choose* $x = 1$*).*

In general there may be multiple equilibria for a given specification of the functions and parameters describing the environment. With multiple equilibria, it makes sense to focus just on equilibria that are stable.

Definition 1 *An equilibrium is locally stable if any deviation by people who were indifferent (or nearly indifferent) in the equilibrium is strictly unprofitable when the deviation is*

accurately perceived (and sufficiently widespread to be noticed).[3]

Proposition 4 *A pooling equilibrium is locally stable if and only if Equation 2 holds as a strict inequality,* $g(\alpha) - g(0) > \lambda\left(h(1) - h(0)\right)$*. A semi-separating equilibrium with threshold* p^* *is locally stable if and only if*

$$g(1) - g\left(\frac{\alpha F^+(p)}{1-\alpha+\alpha F^+(p)}\right) - \lambda\left(h(1-p) - h(p)\right) \tag{5}$$

changes from negative to positive at $p = p^*$*.*

Proposition 4 refines the set of equilibria to exclude those that are unstable in the following sense: if a person can deviate at no cost, and this deviation would persist when the audience eventually noticed it, then we expect that eventually the equilibrium would fall apart.

This refinement does not guarantee uniqueness. There can still be multiple stable equilibria. For example, if $g(1)-g(0) > \lambda(h(1)-h(0))$ and if $F^+(p) > 0$ for some $p < \frac{1}{2}$, then for α sufficiently close to (but not equal to) 1, there must be multiple equilibria, including both a pooling equilibrium with strong social pressure to conform and a semi-separating equilibrium with more tolerance for disagreement. When there are multiple equilibria, the social norm is not merely descriptive, but actively reinforces behavior. Other people expressing an opinion increases pressure on you to express it as well.

Restricting our focus to the set of stable equilibria does guarantee that a convenient technical condition holds.

Corollary 1 *In any locally stable semi-separating equilibrium with threshold* p^**, the cumulative distribution function* $F^+(p)$ *is continuous at* p^**.*

In general, the cumulative distribution function F^+ has jump discontinuities whenever a population mass shares the same beliefs, but Corollary 1 assures us that we have continuity right at the threshold in a locally stable semi-separating equilibrium. This facilitates comparative static analysis.

We now consider comparative statics. Let the cumulative distribution function for beliefs be decomposed as $F^+(p) = F_0^+(p) + \epsilon\chi(p)$ for some baseline cumulative distribution function F_0^+ and some non-negative function χ, so that increasing ϵ corresponds to shifting the cumulative distribution function upwards so that more people have lower probabilities p about the state of the world.

[3]The requirement that the deviation be sufficiently widespread is meant only to rule out deviations by a negligible fraction of the population, which may be unprofitable, but not strictly unprofitable.

Theorem 1 *For any locally stable semi-separating equilibrium, the threshold p^* is (weakly) increasing in λ (i.e., in the magnitude of expressive utility relative to reputational utility), in ϵ (i.e., in downward shifts in beliefs about the state of the world, corresponding to upward shifts in the cumulative distribution function over beliefs), and in α (i.e., in the fraction of the population with desirable values) in a neighborhood of the parameter space for which the equilibrium continues to exist.*

Theorem 1 tells us that if people care more about expressive utility (or, equivalently, care less about reputational utility), or people are more skeptical about the state of the world necessary for privately holding the desirable opinion, or more people have the desirable values, then there is less pressure to conform to the norm of expressing the desirable opinion, and people who are a bit less sure of the undesirable opinion will still be comfortable expressing it. It is not surprising that caring about authenticity, not caring about social esteem, or doubting that the facts support a politically correct opinion relieves the pressure to express it. However, the last behavioral pattern predicted in the theorem is less obvious. Although other people expressing a politically correct opinion increases the social pressure to conform with it, more people having the socially desirable values actually decreases the social pressure to express politically correct opinions. When desirable values are universal, a person can safely disagree with prevailing opinions without calling his own values into question. In this case the disagreement can be attributed to his beliefs rather than his values. For example, a war hero may feel more comfortable speaking out against his country's foreign policy, without fear of being seen as unpatriotic, than a foreign-born resident who shares the same opinion but feels more vulnerable to such an accusation.

2.2 Case II: Beliefs and Values are Substitutes

Consider now the case the beliefs and values are substitutes in producing the desirable opinion, i.e., that $t(\omega, v) = 1$ if and only if $\omega = \text{True}$ or $v = V^+$. In this case, a person's confidence about his private opinion depends on his belief about the state of the world only if he has the undesirable values, i.e., $q = I^+(v) + (1 - I^+(v))p$ where I^+ is an indicator for having desirable values V^+. For example, a member of a religious community could endorse a particular religious practice either because it is consistent with his spiritual values or because he believes that God will punish people who do not partake in it. The belief that God rewards or punishes people substitutes for internalized spiritual values in producing affirmation of the religious practice.

In the case that beliefs and values are substitutes, there is still a pooling equilibrium in

which everybody chooses $x = 1$, supported by $b(0) = 0$ and $b(1) = \alpha$.

Proposition 5 *If Equation (2) holds, then there exists a pooling equilibrium in which everybody chooses* $x = 1$.

The sufficient condition for the pooling equilibrium remains the same as in the case that beliefs and values are complements.[4]

A semi-separating equilibrium still takes the form of a threshold of sufficient confidence in the undesirable opinion that permits a person to publicly express it.

Proposition 6 *Proposition 2 still applies: in any semi-separating equilibrium, there exists a* $p^* \leq .5$ *such that* $x = 0$ *if* $q < p^*$ *and* $x = 1$ *if* $q > p^*$.

However, the characterization of the semi-separating equilibrium is different in the case that beliefs and values are substitutes. In this case, choosing $x = 0$ perfectly reveals oneself to have undesirable values V^-, i.e., $b(0) = 0$. However, people with beliefs $p > p^*$ and people with desirable values V^+ pool together, with both groups choosing $x = 1$. The Bayesian inference after observing $x = 1$ is

$$b(1) = \frac{\alpha}{\alpha + (1-\alpha)(1 - F^-(p^*))}. \tag{6}$$

Now the threshold p^* in the semi-separating equilibrium will satisfy:

$$g\left(\frac{\alpha}{\alpha + (1-\alpha)(1 - F^-(p^*))}\right) - g(0) = \lambda\left(h(1-p^*) - h(p^*)\right). \tag{7}$$

Proposition 7 *Any solution* p^* *to Equation (7) defines an equilibrium in which people who are sufficiently confident in the undesirable opinion (with* $q \leq p^*$*) express it (i.e., choose* $x = 0$*), and people who are sufficiently confident in the desirable opinion (with* $q > p^*$*) express that view (i.e., choose* $x = 1$*).*

Whereas there may be multiple equilibria when beliefs and values are complements, in the case that they are substitutes, there can only be a single, unique equilibrium.

Theorem 2 *The equilibrium is unique.*

[4] If nobody holds beliefs p below some minimal level $\underline{p}$, then the sufficient condition can be relaxed to be $g(\alpha) - g(0) \geq \lambda\left(h(1-\underline{p}) - h(\underline{p})\right)$. This is equivalent to a semi-separating equilibrium with nobody below the threshold.

There is either a pooling equilibrium, but no semi-separating equilibrium (when Equation (2) holds) or a semi-separating equilibrium, but no pooling equilibrium (when Equation (2) does not hold). They cannot co-exist. While there may be interesting applications in both domains, contexts in which the pooling equilibrium exists correspond to situations in which we usually think of social norms as operative. Contexts in which only the semi-separating equilibrium exists correspond to situations in which there is less social pressure to not look bad, but a person may still have some incentive to try to stand out and be noticed in a good way.

Because multiple equilibria cannot co-exist in the case that beliefs and values are substitutes, behavior should be more stable in these situations, and learning that others are more frequently expressing an opinion should no longer generate any additional pressure to conform and express the same opinion. While there is always some social pressure to express the desirable opinion (to avoid social disapproval), this pressure would actually decrease if the desirable opinion became more popular. Thus, the model predicts that conformity pressure (i.e., increasing social pressure as society approaches conformity in favor of the popular opinion) arises when beliefs and values are complements, but not when they are substitutes.

3 Motivated Reasoning, Persuasion, and Belief Formation

In a semi-separating equilibrium everybody will observe a fraction π of the population choosing $x = 1$. If belief formation were strictly Bayesian, then in many contexts a no-disagreement theorem (Aumann, 1976) would imply that beliefs should converge, eliminating the heterogeneity of beliefs that is a critical element in the semi-separating equilibrium. Here we sketch a set of alternative assumptions that can preserve heterogeneity of beliefs. We assume that beliefs about α (the fraction of the population with values $v = V^+$) are no longer omniscient, but instead are subject to motivated reasoning and persuasion. Motivated reasoning and persuasion are pathways for preferences about one's own beliefs and preferences about others' beliefs to affect these beliefs. Still, we assume that motivated reasoning and persuasion determine only beliefs about model parameters (i.e., about how the world works), but cannot directly determine beliefs about states of the world, which still must accord with Bayes' rule, given the adopted model parameters.[5] Moreover, for a motivated belief to take hold or a transmitted belief to be found persuasive, it must account

[5]We also assume that no individual has private information completely resolving uncertainty about the state of the world, so everyone remains receptive to additional information contained in others' beliefs.

for the observation that a fraction π of the population is choosing $x = 1$. We assume that a fraction γ of the population engages in motivated reasoning (when there is a motive to hold particular beliefs) and that the remainder $1 - \gamma$ of the population considers only beliefs that they find out about from others. We focus on the case in which beliefs and values are complements in producing the desirable opinion, because in this case, the semi-separating equilibrium may coexist with the pooling equilibrium as alternative social norms.

We first consider the range of possible narratives, or sets of beliefs that are consistent with the observation that a fraction π of the population is choosing $x = 1$, arising out of different beliefs about model parameters. Beliefs about model parameters essentially correspond to interpretations of the observed evidence, which support particular beliefs p about the state of the world ω. Let β_α denote a person's belief about the value of α, the proportion of the population with desirable values.[6] At one extreme, a person could believe $\beta_\alpha = 1$, i.e., that everybody has the desirable values and that expressed opinions are thus diagnostic only of beliefs about the state of the world, but not about a person's values. The other extreme, $\beta_\alpha = 0$, would be inconsistent with any observation $\pi > 0$, because in a semi-separating equilibrium, expressing $x = 1$ perfectly reveals oneself as having desirable values. The most extreme minimal value of β_α consistent with the observation of π is $\beta_\alpha = \pi$, the belief that the fraction of the population with desirable values is precisely the fraction of the population expressing the desirable opinion, and consequently that not expressing this opinion reveals a person to have undesirable values.

The belief $\beta_\alpha = \pi$ implies that expressed opinions are perfectly diagnostic of a person's values. This narrative leads to the assessment that everybody with desirable values agrees that $p > p^*$, where p^* is the threshold in a semi-separating equilibrium that is now defined by $g(1) - g(0) = \lambda \left(h(1 - p^*) - h(p^*) \right)$ (because the individual believes $F^+(p^*) = 0$). The narrative supports the belief $p = p_{\max}$, where $p_{\max}$ denotes the value of p consistent with common knowledge that $p > p^*$. (The precise value of $p_{\max}$ depends on the informativeness of others' beliefs.)[7]

The belief $\beta_\alpha = 1$ implies that expressed opinions are diagnostic of beliefs about the

[6] In principle, we could have assumed that the expected reputational utility directly depends on the belief β_α because esteem for one's own values could depend on the audience's values, but instead we assume that we are in a regime where esteem does not vary with audience values, either because the desirable values are not audience-dependent or because the range of beliefs β_α is sufficiently narrow that the desirable values are stable across this range of beliefs.

[7] To account for the observation of π, this narrative also requires that if the person believes that anybody engages in motivated reasoning $\beta_\gamma > 0$, then it must be optimal for individuals with desirable values to want to believe that $\beta_\alpha = \pi$, which, as we will see, turns out to be the case.

state of the world, so some additional belief is necessary to account for the disagreement about the state of the world that is revealed by observing $0 < \pi < 1$. Let β_γ denote a person's belief about the value of γ, the proportion of the population that engages in motivated reasoning. If motivated reasoning leads an individual to hold beliefs that permit $x = 1$, then the observation of π could be rationalized by the additional belief that $\beta_\gamma = \pi$, which leads to the assessment that everybody else agrees that $p < p^*$. (In this scenario, the threshold p^* may be defined by $\pi\left(g(1) - g(0)\right) = \lambda\left(h(1-p^*) - h(p^*)\right)$, given that (only) motivated reasoners will make the attribution $b(0) = 0$.) This narrative supports the belief $p = p_{\min}$, where $p_{\min}$ denotes the value of p consistent with common knowledge that $p < p^*$. (Once again, the precise value of $p_{\min}$ depends on the informativeness of others' beliefs.)[8]

We now describe preferences about one's own beliefs and the optimal strategy of motivated reasoning, before moving on to analyze preferences about others' beliefs and the optimal strategy of persuasion. A motivated reasoner chooses β_α (along with β_γ), subject to the constraint that this choice rationalizes the observation of π, in order to maximize his own utility. (His utility may depend on these parameters because they shape his inferences about other people's choices, and in turn his belief p about the state of the world and his confidence q in his own private opinion, and thus, possibly his choice x about which opinion to express in public. Additionally, they may affect how the motivated reasoner expects others to perceive and interpret his own expressed opinion.)

Consider an individual with desirable values $v = V^+$. He would like to have high confidence in the desirable opinion $t = 1$, allowing himself to express it (as $x = 1$) to obtain high expressive utility and high reputational utility, so he wants to maximize p. We have assumed that he cannot choose p directly, but he can choose his belief β_α (as long as it is consistent with his observation of π), and potentially convince himself that $p = p_{\max}$. As long as $p_{\max}$ is not too small, his optimal strategy will be to choose $\beta_\alpha = \pi$, which provides the strongest possible support for the desired belief, $p = p_{\max}$.[9] (In this case, he will be indifferent between values of $\beta_\gamma < 1$.)

Lemma 1 *Suppose that* $h(p_{\max}) > h(1-p_{\min}) - \frac{\pi}{\lambda}\left(g(1) - g(0)\right)$. *Then an individual with*

[8] Alternatively, if motivated reasoning leads an individual to hold beliefs that permit $x = 0$, then the observation of π could be rationalized by the belief that $\beta_\gamma = 1-\pi$, with the assessment that there is common knowledge among everyone else that $p > \frac{1}{2}$. However, this narrative will not be part of an equilibrium.

[9] If $p_{\max}$ were very small and people were thus willing to sacrifice reputational utility to maximize expressive utility, there would only be a pooling equilibrium with $x = 0$ and no loss of reputational utility, and motivated reasoners would try to minimize p, displaying a kind of confirmation bias.

desirable values $v = V^+$ who engages in motivated reasoning after observing a fraction π of the population choosing $x = 1$ adopts the belief that $\beta_\alpha = \pi$, convincing himself that $F^+(p^*) = 0$ and that $p = p_{\max}$.

Motivated belief that $\beta_\alpha = \pi$ permits an interpretation that choices of $x = 1$ contain informational content about the state of the world, whereas choices of $x = 0$ contain no informational content (because they're only made by individuals with undesirable values who would make that choice regardless of their beliefs). Motivated reasoners thus form a narrative that also presumes common knowledge about the state of the world, $F^+(p^*) = 0$ and $p = p_{\max}$.

We now describe preferences about others' beliefs and incorporate those preferences into strategies of persuasion. A person who does not engage in motivated reasoning will accept a β_α and β_γ offered by another person if these parameter values would rationalize the observation of π. A person will then engage in persuasion and argue for his preferred values of β_α and β_γ (for another person to accept) if his own utility would increase when this other person adopts these beliefs. (His own utility may depend on another person's adoption of these beliefs because they guide the other person's inference about his values from his expressed opinion.)

Consider an individual with desirable values $v = V^+$ and beliefs that support expressing the desirable opinion $x = 1$, i.e., $p > p^*$. By expressing the desirable opinion, he reveals his desirable values to the audience, regardless of the audience's other beliefs. Consequently, he is indifferent to others' beliefs and need not engage in persuasion. On the other hand, consider an individual with desirable values $v = V^+$ and beliefs that support expressing the undesirable opinion $x = 0$, i.e., $p < p^*$. The audience's attribution for his expression of this opinion as either due to his beliefs $p < p^*$ or his values $v = V^-$ depends on their belief β_α. The individual wants to persuade the audience that $\beta_\alpha = 1$, and that his opinion can be attributed to his beliefs, in order to preserve his reputational utility. The same goes for an individual who actually has undesirable values $v = V^-$ (and thus expresses the undesirable opinion $x = 0$. He too wants to persuade the audience that $\beta_\alpha = 1$ to disguise his own type.

Lemma 2 *People who express the undesirable opinion $x = 0$ will attempt to persuade the audience that $\beta_\alpha = 1$. People who express the desirable opinion $x = 1$ will not engage in persuasion.*

Individuals with desirable values $v = V^+$ who are capable of motivated reasoning will not

be influenced by persuasion. They will have already convinced themselves that $\beta_\alpha = \pi$, and they will believe that the only sources of the narrative that $\beta_\alpha = 1$ are people with undesirable values $v = V^-$. However, individuals with desirable values $v = V^+$ who are not capable of motivated reasoning may be influenced by persuasion. If the argument that $\beta_\alpha = 1$ is packaged with an argument that $\beta_\gamma = \pi$ in a coherent narrative, it accounts for the observation that the proportion π of the population chooses $x = 1$. There are always some individuals with undesirable values $v = V^-$ pushing this narrative, and as individuals with desirable values $v = V^+$ (but no motivated reasoning) are persuaded by this narrative, they too will then want to share it themselves.

Incorporating motivated reasoning and persuasion into the model, we now have a basis for heterogeneous beliefs that support a semi-separating equilibrium.

Proposition 8 *Suppose*

$$\alpha\gamma\left(g(1) - g(0)\right) \geq \lambda\left(h(1 - p_{\min}) - h(p_{\max})\right) \tag{8}$$

and

$$\alpha\gamma\left(g(1) - g(0)\right) \leq \lambda\left(h(1 - p_{\min}) - h(p_{\min})\right). \tag{9}$$

Then there is a semi-separating equilibrium in which:

1. *A fraction $\pi = \alpha\gamma$ of the population expresses the desirable opinion $x = 1$;*
2. *People with undesirable values $v = V^-$ choose to express the undesirable opinion $x = 0$ and to argue that $\beta_\alpha = 1$ and $\beta_\gamma = \pi$;*
3. *People with desirable values $v = V^+$ who do not engage in motivated reasoning are persuaded that $\beta_\alpha = 1$, that $\beta_\gamma = \pi$, and that $p = p_{\min}$, and they thus also choose to express the undesirable opinion $x = 0$ and to argue that $\beta_\alpha = 1$ and $\beta_\gamma = \pi$;*
4. *People with desirable values $v = V^+$ who engage in motivated reasoning persuade themselves that $\beta_\alpha = \pi$ and that $p = p_{\max}$, and they thus choose to express the desirable opinion $x = 1$.*

In this equilibrium, beliefs about the state of the world (and associated private opinions) become correlated with beliefs about other people's types, as people form narratives (i.e., a set of coherent beliefs) to account for the persistent disagreement in public opinions.

The semi-separating equilibrium thus captures a polarized society, with some people conforming to politically correct discourse and convincing themselves that any deviation from this norm conclusively reveals racism, and other people arguing that racism has vanished and that political correctness itself is the problem. These incompatible narratives make it impossible for the polarized groups to agree even about factual beliefs.

Of course, in other situations the social norm may be strong enough to induce full conformity to acceptable discourse. That is, the pooling equilibrium may still exist as well. Observing everybody conforming to the social norm, $\pi = 1$, motivated reasoners would want to convince themselves that everybody has desirable values, $\beta_\alpha = 1$. This belief would not only assure oneself that his own values are beyond reproach, but also allow a person to convince himself that others' choices of $x = 1$ are unaffected by their desire for reputational utility and are fully reflective of unbiased belief that $p > \frac{1}{2}$. This interpretation would bolster one's own belief $p > \frac{1}{2}$ and thus maximize expressive utility.

Proposition 9 *Suppose Equation (2) holds:*

$$g(\alpha) - g(0) \geq \lambda \left(h(1) - h(0) \right).$$

Then there is a pooling equilibrium in which everybody chooses $x = 1$, and people who engage in motivated reasoning persuade themselves that $\beta_\alpha = 1$ and that there is common knowledge that $p > \frac{1}{2}$.

Proposition 9 tells us that motivated reasoning pushes people toward beliefs and privately held opinions that align with the social norm when everybody conforms to this norm. Thus, with motivated reasoning, the social norm actually affects underlying beliefs and privately held opinions, not only the opinions that people are willing to express out loud. For example, a strong social norm against criticizing one's country may lead people to form idealized beliefs about the country.

4 Conclusion

This paper develops a theory of social norms of acceptable discourse that explains how norms emerge and how they shape beliefs and opinions. We propose that they emerge because people care about signaling desirable values, which are an important aspect of a person's identity. They shape opinion expression because what an opinion reveals about a person's values depends on how other people choose the opinions they endorse. Norms (equilibria) with more or less conformity induce more or less social pressure to send a

good signal. Particular beliefs may be necessary in order to have the right incentive to send a good signal (i.e., to express an opinion that conforms to the norm), whereas other beliefs may provide a mitigating explanation for why a person might send a bad signal (i.e., express an unpopular opinion). Thus, motivated reasoning and persuasion allow the social norm about opinion expression to also affect privately held opinions and beliefs.

The model here, for simplicity, focuses on a single binary choice about an opinion to express. In reality, when a person expresses multiple opinions, each one may send a signal about the person's values. This could be modeled as a set of linked signaling games. We should expect the esteem motive to be stronger when a person's other opinions (as well as other behavior) allow for more uncertainty about his true values. In the opposite extreme, if a person's values are already observable from other behavior, the esteem motive vanishes, and the theory would predict that the person would then simply express his honest opinion. If audience attention to a person's opinions is uncertain, then multiple opinions could all be similarly informative. Polarized opinions could be correlated across issues (as commonly observed), and more complex phenomena could emerge, such as people trying to signal their desire for authenticity (and, in turn, the sincerity of their opinions) by occasionally expressing an unpopular (politically incorrect) opinion.

The model assumes a single, fixed audience whose judgments people care about. This describes public discourse as well as private conversation with multiple parties when people are constrained to express the same opinions across all of these conversations, perhaps, for example, due to a desire for consistency. An interesting extension, beyond our scope here, would consider multiple audiences. This might permit customization of opinions to ingratiate oneself with distinct audiences or dog whistling to send distinct signals to different audiences with a single common statement of opinion. If reputational utility were sufficiently concave, so that a good judgment from one audience would not compensate for a bad judgment from another audience, then people might be reluctant to express any opinion about which there might be disagreement, avoiding "conversational minefields" (Sugden, 2005).

The model assumes that play of the signaling game follows a perfect Bayesian equilibrium, in which people have unlimited strategic sophistication and inferences about others' private opinions are rationally based on their behavior. In reality, of course, people find it difficult to make Bayesian inferences from others' behavior (Forsythe et al., 1989; Brown et al., 2012) or statements (Braghieri, 2021). Relaxing the assumed strategic sophistication and instead assuming cursed equilibrium (Eyster and Rabin, 2005) or cognitive hierarchy

theory (Camerer et al., 2004) might account for the phenomenon of pluralistic ignorance, which arises when people privately disagree with an opinion perceived to be socially normative, but mistakenly believe that others mostly support it (Prentice and Miller, 1993; 1996; van Boven, 2000).

The model treats people's values, as well as the esteem they may receive for displaying particular values, as given and fixed, and focuses on how these factors affect social norms of opinion expression. At longer time scales, however, it may be possible for social norms of opinion expression to shape the cultural transmission of values in a society, creating a feedback loop here (see Bisin and Verdier (2000; 2011), Akerlof (2017), and Bernheim et al. (2021) for related work). Changes in the values that people consider socially desirable likely play a role in the evolution of social norms of opinion expression over time.

Conspicuously absent from the model is any concern for influencing behavior. While some people clearly are motivated to bring about social change, this modeling choice reflects the view that many people consider their own impact on collective behavior to be insignificant and care more about their reputation and esteem than about instrumental consequences. This desire for esteem is the reason why people use opinions for signaling socially desirable values, and thus is a critical element in our explanation of the emergence of social norms for beliefs and opinions.

Appendix (For Online Publication)

Proof of Proposition 1 An individual with $v = V^-$ has the strongest incentive to deviate from the pooling equilibrium because he is sure about his private opinion $t = 0$, and thus $q = 0$. If he chooses $x = 1$, his utility is $g(\alpha) + \lambda h(0)$. If he chooses $x = 0$, his utility is $g(0) + \lambda h(1)$. Thus, Equation 2 is his incentive constraint.

Proof of Proposition 2 Existence of a threshold p^* follows from monotonicity of expressed opinions with respect to confidence, i.e., that if a person with confidence q chooses $x = 1$, then anybody with confidence $q' > q$ must also choose $x = 1$. The reputational component of utility does not depend on q, and the gain in expressive utility from choosing $x = 1$ instead of $x = 0$ is $\lambda\left(h(q) - h(1-q)\right)$, which is increasing in q. We know that $p^* \leq .5$ because the gain in expressive utility $\lambda\left(h(q) - h(1-q)\right)$ changes from negative to positive at $.5$, and the gain in reputational utility from $x = 1$ instead of $x = 0$ is always non-negative.

Proof of Proposition 3 Let p^* be a solution to Equation 4, and suppose the audience believes that this is the threshold confidence that separates people who choose $x = 0$ from

$x = 1$. The utility from choosing $x = 1$ is $g(1) + \lambda h(q)$. The utility from choosing $x = 0$ is $g\left(\frac{\alpha F^+(p^*)}{1-\alpha+\alpha F^+(p^*)}\right) + \lambda h(1-q)$. Indeed, $x = 1$ is optimal for $q > p^*$, and $x = 0$ is optimal for $q < p^*$.

Proof of Proposition 4 First consider the pooling equilibrium. If Equation 2 holds as a strict inequality, then nobody is indifferent, so the equilibrium must be locally stable. If it holds as an equality, then a person with $q = 0$ could deviate to $x = 0$ with no loss (and, in fact, a strict gain if the person has values V^+ and beliefs $p = 0$, in which case, accurate perception of the deviation will be $b(0) = 1$).

Next consider the semi-separating equilibrium. Let $\Delta(p)$ be the function in Expression (5). People with $q = p^*$ are indifferent in equilibrium. Switching from $x = 0$ to $x = 1$ is strictly unprofitable when accurately perceived if and only if $\Delta(p)$ is negative for p approaching p^* from below. Switching from $x = 1$ to $x = 0$ is strictly unprofitable when accurately perceived if and only $\Delta(p)$ is positive for p approaching p^* from above.

Proof of Corollary 1 Proposition 4 tells us that in any locally stable semi-separating equilibrium with threshold p^*, the function $\Delta(p)$ (given by Expression (5)) is increasing at $p = p^*$. If the cumulative distribution function F^+ were to jump at $p = p^*$, then $\Delta(p)$ would have a downward jump at $p = p^*$ and could not be increasing.

Proof of Theorem 1 The threshold p^* in a semi-separating equilibrium, characterized by Proposition 3 as a solution to Equation 4, can equivalently be characterized as a zero of the function $\Delta(p)$ (given by Expression (5)). Differentiating $\Delta(p^*) = 0$ gives us

$$\Delta'(p^*)dp^* + \left.\frac{\partial \Delta}{\partial \lambda}\right|_{p=p^*} d\lambda + \left.\frac{\partial \Delta}{\partial \epsilon}\right|_{p=p^*} d\epsilon + \left.\frac{\partial \Delta}{\partial \alpha}\right|_{p=p^*} d\alpha = 0. \tag{10}$$

Proposition 4 tells us that in a locally stable equilibrium $\Delta'(p^*) > 0$ (because $\Delta(p)$ changes from negative to positive here). We can directly compute $\left.\frac{\partial \Delta}{\partial \lambda}\right|_{p=p^*} = -\left(h(1-p^*) - h(p^*)\right)$. Given that $p^* \leq .5$ (by Proposition 2) and h is increasing, $h(1-p^*) - h(p^*) \geq 0$, so $\left.\frac{\partial \Delta}{\partial \lambda}\right|_{p=p^*} \leq 0$. To simplify computation of the other partial derivatives, we let $B(p) = \frac{\alpha F^+(p)}{1-\alpha+\alpha F^+(p)}$, and observe that $B(p)$ is (weakly) increasing in $F^+(p)$ (and thus in ϵ) and in α. We then compute $\frac{\partial \Delta}{\partial \epsilon} = -g'(B(p))\frac{\partial B}{\partial \epsilon}$ and $\frac{\partial \Delta}{\partial \alpha} = -g'(B(p))\frac{\partial B}{\partial \alpha}$. Given that g is increasing, $\frac{\partial \Delta}{\partial \epsilon} \leq 0$ and $\frac{\partial \Delta}{\partial \alpha} \leq 0$. Returning to Equation (10), with $\Delta'(p^*) > 0$ and the other three partial derivatives negative (or zero), we can conclude that $\frac{dp^*}{d\lambda} \geq 0$, $\frac{dp^*}{d\epsilon} \geq 0$, and $\frac{dp^*}{d\alpha} \geq 0$.

Proof of Proposition 5 In this case, the individual with the strongest incentive to deviate from the pooling equilibrium has values V^- and beliefs $p = 0$. His confidence is still $q = 0$, as in Proposition 1. Thus, Equation 2 is still his incentive constraint.

Proof of Proposition 6 The proof of Proposition 2 does not rely on any particular expression for confidence q.

Proof of Proposition 7 Let p^* be a solution to Equation 7, and suppose the audience believes that this is the threshold confidence that separates people who choose $x = 0$ from $x = 1$. The utility from choosing $x = 1$ is $g\left(\frac{\alpha}{\alpha+(1-\alpha)(1-F^-(p^*))}\right) + \lambda h(q)$. The utility from choosing $x = 0$ is $g(0) + \lambda h(1 - q)$. Indeed, $x = 1$ is optimal for $q > p^*$, and $x = 0$ is optimal for $q < p^*$.

Proof of Theorem 2 Let

$$\tilde{\Delta}(p) = g\left(\frac{\alpha}{\alpha + (1-\alpha)(1-F^-(p))}\right) - g(0) - \lambda\left(h(1-p) - h(p)\right).$$

Observe that p^* is the threshold in a semi-separating equilibrium if and only if p^* is a zero of $\tilde{\Delta}$ or a point at which $\tilde{\Delta}$ jumps across zero. (In the latter case, the equilibrium requires people with confidence $q = p^*$ to mix their strategies.) The function $\tilde{\Delta}(p)$ is increasing because $g\left(\frac{\alpha}{\alpha+(1-\alpha)(1-F^-(p))}\right)$ is increasing in p and $h(1-p) - h(p)$ is decreasing in p. Thus, it can cross zero at most once. Plug in $p = 0$ to find $\tilde{\Delta}(0) = g(\alpha) - g(0) - \lambda\left(h(1) - h(0)\right)$. Plug in $p = \frac{1}{2}$ to find $\tilde{\Delta}(\frac{1}{2}) = g\left(\frac{\alpha}{\alpha+(1-\alpha)(1-F^-(\frac{1}{2}))}\right) - g(0)$, and so, $\tilde{\Delta}(\frac{1}{2}) \geq 0$. If Equation (2) fails to hold, $\tilde{\Delta}(p)$ must cross from negative to positive, and there is a semi-separating equilibrium where it does, but there is no pooling equilibrium. If Equation (2) holds, then $\tilde{\Delta}(p)$ is always positive, and there is no semi-separating equilibrium, but there is a pooling equilibrium.

Proof of Lemma 1 The belief that $\beta_\alpha = \pi$ is consistent with the observation that the fraction π of the population chooses $x = 1$ if and only if $F^+(p^*) = 0$. Thus, adopting the belief $\beta_\alpha = \pi$ leads a person to also believe $F^+(p^*) = 0$, i.e., that there is common knowledge that $p > p^*$. Choices of $x = 0$ are attributed exclusively to undesirable values $v = V^-$, and the person can believe that there is no disagreement about the state of the world. The person thus believes $p = p_{\max}$. Believing $p > p^*$, the person will choose $x = 1$. For this belief to be optimal, it must yield higher utility than alternative beliefs. Clearly no belief that entails a smaller value of p can be optimal as long as the person is still choosing $x = 1$ (because

it would not affect reputational utility and would only detract from expressive utility). If a person were considering an alternative belief that led to a choice of $x = 0$, he would want to minimize p through the narrative that $\beta_\alpha = 1$, $\beta_\gamma = \pi$, and $p = p_{\min}$ (i.e., that only those biased by motivated reasoning are choosing $x = 1$). With that alternative narrative, only the fraction of the audience that was believing $\beta_\alpha = \pi$ and choosing $x = 1$ would judge the person's values badly (as the rest of the audience would believe $\beta_\alpha = 1$), so the loss in reputational utility would be $\pi(g(1) - g(0))$. To ensure that this alternative narrative is not preferred, we need to guarantee that the utility from believing $p = p_{\max}$ and choosing $x = 1$ exceeds the utility from believing $p = p_{\min}$ and choosing $x = 0$. The condition that $h(p_{\max}) > h(1 - p_{\min}) - \frac{\pi}{\lambda}(g(1) - g(0))$ guarantees exactly that.

Proof of Lemma 2 If a person chooses $x = 0$, his reputational utility depends on the audience's perception of his values $b(0)$. Equation 3 gives us the Bayesian inference $b(0)$ that an audience will make after observing a person choose $x = 0$. Observe that $b(0)$ is maximized when $\alpha = 1$. A person choosing $x = 0$ thus wants to promote the belief that $\beta_\alpha = 1$. On the other hand, if a person chooses $x = 1$, his utility does not depend on others' beliefs.

Proof of Proposition 8 When $\pi = \alpha\gamma$, Equation 8 is the incentive constraint from Lemma 1 that implies that people with desirable values who engage in motivated reasoning will persuade themselves that $\beta_\alpha = \pi$ and that $p = p_{\max}$. If this constraint is satisfied, then it will also be optimal to choose $x = 1$, given these beliefs. When $\pi = \alpha\gamma$, Equation 9 is the incentive constraint for people who believe that $\beta_\alpha = 1$, $\beta_\gamma = \pi$, and $p = p_{\min}$ to choose $x = 0$. People who do not engage in motivated reasoning will be persuaded to hold these beliefs because it is the only narrative that is being promoted and it accounts for the observation that a fraction π of the population chooses $x = 1$. Lemma 2 then tells us that these people will promote this narrative themselves. People with undesirable values have a stronger incentive to choose $x = 0$ than people with desirable values and the belief $p = p_{\min}$, because for them, $q = 0$, so Equation 9 also implies that they will choose $x = 0$. Lemma 2 then tells us that they too will spread the narrative that $\beta_\alpha = 1$ and $\beta_\gamma = \pi$. The fraction of the population with desirable values and who engage in motivated reasoning is $\alpha\gamma$, and these are precisely the people that choose $x = 1$, so $\pi = \alpha\gamma$.

Proof of Proposition 9 Proposition 1 already established the existence of a pooling equilibrium with everybody choosing $x = 1$. For a person who is choosing $x = 1$, the optimal

beliefs are $\beta_\alpha = 1$, which directly maximize reputational utility, and which allow the person to maximize p, which then also maximizes expressive utility. If a person believes $\beta_\alpha = 1$, he can also believe that others are not affected by the esteem motive, so their choices may be informative about whether p is greater or less than $\frac{1}{2}$. Observing $\pi = 1$, he can then conclude that there is common knowledge that $p > \frac{1}{2}$.

References

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British Journal of Social Psychology*, 29(2), 97-119.

Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, 78(4), 1201-1236.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The Quarterly Journal of Economics*, 94(4), 749-775.

Akerlof, G. A. (2007). The missing motivation in macroeconomics. *American Economic Review*, 97(1), 5-36.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3), 715-753.

Akerlof, R. (2017). Value formation: The role of esteem. *Games and Economic Behavior*, 102, 1-19.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, 1236-1239.

Austen-Smith, D., & Fryer Jr, R. G. (2005). An Economic Analysis of "Acting White". *The Quarterly Journal of Economics*, 120(2), 551-583.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3), 797-817.

Bénabou, R. (2008). Ideology. *Journal of the European Economic Association*, 6, 321–352.

Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. *Review of Economic Studies*, 80, 429-462.

Bénabou, R., & Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2), 805-855.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, 102(5), 841-877.

Bernheim, B. D., Braghieri, L., Martínez-Marquina, A., Zuckerman, D. (2021). A theory of chosen preferences. *American Economic Review*, 111(2), 720-754.

Bicchieri, C. (2006). *The Grammar of Society*. Cambridge University Press.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992-1026.

Bisin, A., Verdier, T. (2000). "Beyond the melting pot": cultural transmission, marriage,

and the evolution of ethnic and religious traits. *The Quarterly Journal of Economics*, 115(3), 955-988.

Bisin, A., Verdier, T. (2011). The economics of cultural transmission and socialization. In *Handbook of Social Economics*, (eds. J. Benhabib, A. Bisin, M. O. Jackson), 1A, 339-416. Amsterdam: Elsevier.

Braghieri, L. (2021). Political correctness, social image, and information transmission. Working paper.

Brown, A. L., Camerer, C. F., & Lovallo, D. (2012). To review or not to review? Limited strategic thinking at the movie box office. *American Economic Journal: Microeconomics*, 4(2), 1-26.

Brown, G. D. A., Lewandowsky, S., & Huang, Z. (2020). Social Sampling Theory: Authenticity Preference and Social Extremeness Aversion Lead to Social Norm Effects and Polarization. Working paper.

Bursztyn, L., Egorov, G., & Fiorin, S. (2020a). From Extreme to Mainstream: The Erosion of Social Norms. *American Economic Review* forthcoming.

Bursztyn, L., Haaland, I. K., Rao, A., & Roth, C. P. (2020b). Disguising Prejudice: Popular Rationales as Excuses for Intolerant Expression. (No. w27288). National Bureau of Economic Research.

Bursztyn, L., & Jensen, R. (2017). Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure. *Annual Review of Economics*, 9, 131-153.

Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3), 861-898.

Cho, I. K., & Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2), 179-221.

Cohen, G. L. (2003). Party over Policy: The Dominating Impact of Group Influence on Political Beliefs. *Journal of Personality and Social Psychology* 85(5), 808–822.

DeMarzo, P. M., Vayanos, D., & Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3), 909-968.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4), 99-117.

Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3), 133-140.

Eyster, E., & Rabin, M. (2005). Cursed equilibrium. *Econometrica*, 73(5), 1623-1672.

Forsythe, R., Isaac, R. M., & Palfrey, T. R. (1989). Theories and tests of "blind bidding" in sealed-bid auctions. *The Rand Journal of Economics*, 20(2), 214-238.

Golman, R., Loewenstein, G., Moene, K. O., & Zarri, L. (2016). The preference for belief consonance. *Journal of Economic Perspectives*, 30(3), 165-188.

Harrison, D. A., Kravitz, D. A., Mayer, D. M., Leslie, L. M., & Lev-Arey, D. (2006). Understanding attitudes toward affirmative action programs in employment: Summary and meta-analysis of 35 years of research. *Journal of Applied Psychology*, 91(5), 1013-1036.

Herrmann, R. K. (2017). How attachments to the nation shape beliefs about the world: A theory of motivated reasoning. *International Organization*, S61-S84.

Higgins, E. T., & McCann, C. D. (1984). Social Encoding and Subsequent Attitudes, Impressions, and Memory: "Context-Driven" and Motivational Aspects of Processing. *Journal of Personality and Social Psychology* 47(1), 26–39.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, 76(3), 405-431.

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3), 690-707.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, 8(4), 407-424.

Kahan, D. M. (2015). The expressive rationality of inaccurate perceptions. *Behavioral & Brain Sciences*, 40, 26-28.

Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition. Working paper 164.

Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies*, 59(1), 63-80.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498.

Kuran, T. (1997). *Private Truths, Public Lies*. Cambridge, MA: Harvard University Press.

Lasorsa, D. L. (1991). Political outspokenness: Factors working against the spiral of silence. *Journalism Quarterly*, 68(1-2), 131-140.

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Loury, G. C. (1994). Self-censorship in public discourse: A theory of "political correctness" and related phenomena. *Rationality and Society*, 6(4), 428-461.

Loury, G. C. (1995). *One by one from the inside out: Essays and reviews on race and responsibility in America*. New York: Free Press.

Marks, J., Copland, E., Loh, E., Sunstein, C. R., & Sharot, T. (2019). Epistemic spillovers: Learning others' political views reduces the ability to assess and use their expertise in nonpolitical domains. *Cognition*, 188, 74-84.

Masser, B., & Phillips, L. (2003). "What do other people think?"—the role of prejudice and social norms in the expression of opinions against gay men. *Australian Journal of Psychology*, 55(3), 184-190.

Matthes, J., Rios Morrison, K., & Schemer, C. (2010). A spiral of silence for some: Attitude certainty and the expression of political minority opinions. *Communication Research*, 37(6), 774-800.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2), 57-111.

Morris, S. (2001). Political correctness. *Journal of Political Economy*, 109(2), 231-265.

Murphy, K. M., & Shleifer, A. (2004). Persuasion in politics. *American Economic Review*, 94(2), 435-439.

Oprea, R., & Yuksel, S. (2020). Social exchange of motivated beliefs. Working paper.

Ortoleva, P., & Snowberg, E. (2015). Overconfidence in political behavior. *American Economic Review*, 105(2), 504-535.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3), 137-158.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, 64(2), 243-256.

Prentice, D. A., & Miller, D. T. (1996). Pluralistic ignorance and the perpetuation of social norms by unwitting actors. *Advances in Experimental Social Psychology*, 28, 161-209.

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *The Jean Piaget symposium series. Values and knowledge*, 103-135. Lawrence Erlbaum Associates, Inc.

Schwartzstein, J., & Sunderam, A. (2019). Using models to persuade. Working paper.

Seuss. (1984). *The Butter Battle Book*. New York: Random House.

Sherman, D. K., Nelson, L. D., & Ross, L. D. (2003). Naïve realism and affirmative action: Adversaries are more similar than they think. *Basic and Applied Social Psychology*,

25(4), 275-289.

Sugden, R. (1989). Spontaneous order. *Journal of Economic Perspectives*, 3(4), 85-97.

Sugden, R. (2005). Fellow Feeling. In B. Gui & R. Sugden (Eds.), *Economics and Social Interaction: Accounting for Interpersonal Relations*. Cambridge University Press.

Sunstein, C. R. (2018). *#Republic: Divided Democracy in the Age of Social Media.* Princeton University Press.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755-769.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the Social Group: A Self-Categorization Theory*. Basil Blackwell.

Van Boven, L. (2000). Pluralistic Ignorance and Political Correctness: The Case of Affirmative Action. *Political Psychology* 21(2), 267–276.

Wood, W., Pool, G. J., Leck, K., & Purvis, D. (1996). Self-Definition, Defensive Processing, and Influence: The Normative Impact of Majority and Minority Groups. *Journal of Personality and Social Psychology* 71(6), 1181–1193.

Ybarra, O., & Trafimow, D. (1998). How priming the private self or collective self affects the relative weights of attitudes and subjective norms. *Personality and Social Psychology Bulletin*, 24(4), 362-370.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, 57-84.

Young, H. P. (2015). The Evolution of Social Norms. *Annual Review of Economics*, 7, 359-387.