Heterogeneous norms: Social image and social pressure when people disagree

Vera L. te Velde*

August 1, 2018

Abstract

People are often divided by what they believe is the right thing to do, in domains from partisan politics to lifestyle choices. Existing models of social pressure break down in these settings because it's unclear whose opinion determines an individual's social image. Some models have addressed this by having individuals seek the approval of the largest contingent in the relevant reference group, but this nullifies the signaling role of moral behavior that has been shown to be very important in homogeneous norm settings. Using a psychological game theoretic model, I show how social pressure affects behavior in these heterogeneous norm settings through two possible channels: "approval" is conferred when actions are deemed correct and "respect" when actions are inferred to be motivated by strong personal beliefs. These motives lead to different outcomes in terms of consensus, hypocrisy, compromise, and destructive posturing. Respect and approval can also interact, which may cause people to deliberately choose disfavored options for the sake of signaling true belief in a minority norm. Results demonstrate how using social incentives to change behavior, an increasingly popular tactic in domains such as development, may easily backfire if there is hidden heterogeneity in norms or if approval and respect motivations are conflated.

JEL classification: D03; Z13; C72; D71.

Keywords: Norms, Social Pressure, Social Image, Signaling

1 Introduction

In the last two decades, social image and social pressure have emerged as key determinants of prosocial behavior in the economics literature. They have been exploited to increase voter

*vtevelde@gmail.com. University of Queensland School of Economics, Colin Clark Bldg 39 Level 6, St. Lucia, QLD, 4072, Australia. I thank Shachar Kariv, Matthew Rabin, Ulrike Malmendier, Stefano DellaVigna, Edward Miguel, Gerard Roland, Richard Thaler, Eugene Caruso, Klaus Schmidt, Ted O'Donoghue, Charles Sprenger, Muriel Niederle, Todd Rogers, Aaron Bodoh-Creed, David Hirschleifer, several anonymous referees, and many seminar participants for helpful comments. All errors are my own. Financial support from the National Science Foundation and the Berkeley Center for the Economics of Demography and Aging is gratefully acknowledged.

participation in a highly cost-effective manner (Gerber et al., 2008), increase donations to charity by up to 42% (DellaVigna et al., 2012), increase blood donation rates (Lacetera and Macis, 2010), and promote safe sex practices more effectively than financial incentives (Ashraf et al., 2012).[2] But settings like voting and donating to charity, in which there is a unanimously prescribed action, are rare: far more common are the decisions about which people disagree, such as *how* to vote, *how* to allocate resources, how to raise children, what dietary and religious habits to keep, or what customs to follow.

These settings exhibit heterogeneous "personal norms". A personal norm is an individual belief in a prescribed ideal behavior. [3] When personal norms are shared, social image is naturally captured by public knowledge of the same deviation that the individual recognizes, so that social image further motivates adherence to personal norms by adding (publicly experienced) shame to (personally felt) guilt. But when personal norms are heterogeneous, social image no longer unambiguously motivates a particular choice. In some situations, one may wish to balance approval from one group against disapproval from another, but in others, one may wish to develop a reputation as someone with strong moral integrity even in the face of disapproval. For example, I might want to vote for a tax cut because my peers agree with that political stance. Or, I might want to vote for a tax *hike*, because when my peers see that I am voting for a policy that will increase my own tax bill, they will infer that I have strong beliefs and am voting non-hypocritically. If the right thing to do is unanimously recognized, these two motivations will always push in the same general direction.[4] But when personal norms are heterogeneous, we can separately define two types of potentially conflicting social image motivations:

"Approval" is the social image that results when an observer thinks that the decision maker is doing the right thing. Approval seekers want to make their peers happy by going along with everyone else's beliefs, and they are tempted to abandon their personal norms

[2]In still other domains, Babcock et al. (2010) show that incentivizing teams can be more effective than individuals because team members don't want to let down their peers, and Bandiera et al. (2005) show that altruism alone cannot account for this. Bandiera et al. (2010), Falk and Ichino (2006), and Mas and Moretti (2009) further show that low-productivity workers improve when monitored by higher productivity workers and that this persists even if they can't observe the higher productivity workers, suggesting that social image must be the explanation.

[3]This terminology is standard in the social psychology norms literature (Schwartz, 1977; Parker et al., 1995; Kallgren et al., 2000; Bicchieri, 2006, 2010; Cialdini and Goldstein, 2004; White et al., 2009, e.g.) and has been adopted by some economists (Erkut et al., 2015; Calabuig et al., 2018; Burks and Krupka, 2012), but personal norms have also been referred to as "moral norms" (Ek, 2018), "private norms" (Elster, 1989), "opinions" (Michaeli and Spiro, 2015), "internal norms" (Acemoglu and Jackson, 2017), "cultural ideals" (Carvalho, 2017) and "codes of behavior" (Akerlof, 1980).

[4]Although they can lead to different distributions of choices, as shown by Andreoni and Bernheim (2009).

in order to do so. Approval is based on actions, rather than inferred types, so approval emerges in a straightforward non-signaling game in which individuals simply choose the best trade-off between consumption, guilt, and approval, without hoping to convincingly imitate other types. This definition captures the intuition that a vegetarian would likely approve of a friend's conversion to vegetarianism even if he knew it was for lack of enjoying meat rather than for heartfelt moral reasons, and that friend would obtain some image utility from such approval.

"Respect", on the other hand, is the social image that accrues to those who place high priority on following their personal norms, i.e. those who have high integrity. Respect seekers want to signal their integrity, but since integrity is unobservable, observers must infer integrity from choices. Individuals, therefore, would like to make choices that are most strongly attributable to high integrity. They might be vegan, for example, because the personal sacrifice involved in adhering to that restriction consistently is a credible signal of their deeply-held beliefs about animal cruelty. This definition captures the intuition that even though most people are not vegan and don't think there's anything morally wrong with consuming milk, they nonetheless recognize and respect those who go to great lengths to avoid it, and vegans obtain some image utility from that respect.

In order to understand the influence of these distinct types of image motivations, I analyze a general model of choice driven by consumption utility, guilt, and social image, in which social image can result from either or both of approval and respect. I first analyze a version of the model in which only one image motivation is in effect, in order to most starkly contrast the two forces. The comparative statics analysis focuses on the role of social pressure, and demonstrates that efforts to influence behavior using social incentives (i.e. social pressure) critically depend on the type of social image motivations that people have and the heterogeneity in personal norms in the population.

In a binary choice environment in which individuals select which of two norms to adhere to (e.g. omnivorism versus vegetarianism), increasing social pressure causes approval seekers to try harder to please the majority, which increases aggregate conformity with the majority norm. Vegetarian approval seekers might therefore give in and accompany friends to the sushi bar. Respect seekers, on the other hand, respond to increasing social pressure by trying harder to prove that their dietary choices reflect their true beliefs. Respect thinkers who think vegetarianism is morally preferable to omnivorism are then more likely to refrain from meat-eating in public, where that sacrifice is an effective signal of integrity, even if their moral beliefs are not strong enough to make them vegetarian in private. In

the limit, as social pressure comes to dominate all other concerns, hypocrisy is completely eliminated and the population is maximally divided along moral lines. Social pressure therefore changes aggregate eating habits in opposite directions for respect seekers and approval seekers; approval seekers tends towards conformity, while respect seekers become entrenched along moral dividing lines.

The polarizing force that respect can be is even more evident when "middle ground" options are available. Mark Bittman's recommendation to be "vegan before dinnertime" is one such compromise that surely no one believes is morally *ideal*, but which accomplishes much of the goal of vegetarianism with minimal effort required. Approval seekers, responding to increasing social pressure, may be inclined towards such options if they don't want to anger either moral population too badly. Respect seekers, however, while they may take up these options to avoid personal guilt without requiring too much sacrifice, become less willing to compromise as social pressure increases. Because compromise options can never be credible signals of non-hypocrisy if no one truly believes those options are ideal, respect seekers become increasingly polarized along moral lines when social pressure increases.

While increasing polarization is an obvious downside to respect seeking in many environments, approval seeking is prone to the opposite problem. If a slight majority (or an ardent minority) believe in a costly norm, social pressure can lead people to destroy utility to follow "sacrificial" norms they don't believe in at all. I show that respect inherently limits this phenomenon, but approval, which encourages conformity for better or for worse, can amplify it. High social pressure will disguise underlying heterogeneity of beliefs in populations of approval seekers, who conform to the least overall offensive option available. It that option happens to be detrimental to some or all (e.g. female circumcision, spending life savings on expensive funerals, avoiding investing in education) the use of social incentives to improve welfare must focus on *reducing* the importance of social image.

Finally, in Section 5.3, I allow individuals to care simultaneously about approval and integrity and show that disapproval can itself act as an effective signal of integrity. This also emphasizes that approval and respect are not *competing* concepts, but it is left an open question what factors determine the relative importance of approval and respect in a given situation. More importantly, the model shows how designing social incentives under a naive understanding of which type of image motivation is most important can backfire. Respect and approval both unambiguously motivate norm adherence when norms are homogeneous, so it's understandable (and certainly often highly effective) to try to use social incentives to change behavior when norms are seemingly universal. But heterogeneity in norms is

not always obvious; in development economics, heterogeneity in norms can be introduced simply through resistance to outsiders. Social incentives in development initiatives have yielded some notable successes, but this increasingly popular tactic raises a red flag when understood within this model of heterogeneous norms.

These applications are discussed further in Section 6. Until then, Section 2 discusses related literature, 3 will formalize the model of respect and approval, and Sections 4 and 5 will explore the different behavior of approval seekers and respect seekers.

2 Related Literature

This model relates to several existing literatures.

Most broadly, this study is motivated by the overwhelming evidence that social image motivations are an important driver of prosocial behavior. In addition to the field evidence mentioned in the introduction, experiments have isolated the role of social image motivations by manipulating anonymity (Sell and Wilson, 1991; Andreoni and Petrie, 2004; Bohnet and Frey, 1999; Carpenter, 2005; Soetevent, 2005; Cason and Khan, 1999; Franzen and Pointner, 2012; Hoffman et al., 1994; Koch and Normann, 2008; Satow, 1975). Additional evidence comes from experiments showing that people often desire to avoid being put in pressuring situations (Dana et al., 2006; Broberg et al., 2007; Lazear et al., 2012; Malmendier et al., 2014; Oberholzer-Gee and Eichenberger, 2008), or to delegate morally contentious choices to biased agents (Hamman et al., 2010; Bartling and Fischbacher, 2011).

Heterogeneity in personal norms has been studied outside of the issue of social incentives. Norm conflict has been extensively studied in cooperative games, in particular. Norm conflict has been introduced in public goods games by varying the endowments of participants (Anderson et al., 2008; Buckley and Croson, 2006; Chan et al., 1996a,b; Tavoni et al., 2011; Weng and Carlsson, 2015), method of assigning endowments (Cherry et al., 2005; Hackett et al., 1994), private returns to the public good (Chan et al., 1999; Dekel and Fischer, 2015; Engel and Rockenbach, 2011, 2014; Engel and Zhurakhovska, 2014; Fischbacher et al., 2014; Fisher et al., 1995; Gangadharan et al., 2017; Güth et al., 2014; Nikiforakis et al., 2012), and preferences (de Oliveira et al., 2015; Reuben and Riedl, 2008). These changes typically make efficient cooperation more challenging, although some have found that communication and/or convergence on (potentially different across groups) some contribution norm restores cooperation (Reuben and Riedl, 2013). Heterogeneity in personal norms has also been studied in coordination games (Neary, 2012; Carvalho, 2017).

The notion of "approval" studied in this paper is a very simple non-signaling model of social influence, and is closely related to several other models of social influence. Akerlof (1980), very similarly to my notion of approval, models reputation as an increasing function of the fraction of the population that believes in the code of behavior that the individual is following. Carvalho (2013), also similarly, models social image as average community member's evaluation of the individual's choice, according to their own preferences/beliefs. Bernheim (1994) incorporates signaling by modeling image as the observers' inferences of how far the decision maker's personal preferences are from the norm. Bursztyn et al. (2017b) models social image utility as the observer's (inferred) belief that he shares a personal norm with the decision maker. This approach also relates to the literature on identity, in which group members obtain utility from adhering to prescribed group behavior (Akerlof and Kranton, 2000, 2002, 2005; Benjamin et al., 2010; Chang et al., 2017).

Along similar lines, several models assume that social image is a function of others' actions rather than their beliefs.[5] Fershtman et al. (2011), for example, model taboos by assuming the that the cost of considering or breaking a taboo is decreasing in the number of other people who do so, and López-Pérez (2008) models guilt in extensive form games as an increasing function of the number of other players who consistently adhere to normative behavior. Michaeli and Spiro (2015) models social pressure as an increasing function of distance to mean behavior, while Michaeli and Spiro (2017) integrates social pressure experienced in pairwise interactions. Traxler (2010) models stigma from deviation as increasing in overall level of adherence, but allows individuals to weight judgments from members of several subgroups differently. Models of social norm evolution often similarly assume that behavior tends towards group averages Bose et al. (2017); Centola et al. (2005); Granovetter (1978); Lindbeck (1997); Manski and Mayshar (2003), and several models of personal norm formation assume beliefs themselves track group behavior (Calabuig et al., 2018; Kincaid, 2004; Kuran and Sandholm, 2008).

The notion of "respect" studied in this paper is inspired by several signaling models of social influence, although none of these have considered heterogeneity in norms. Several earlier models study social image specifically by defining social image utility as the inference observers have about the decision maker's personal weight on doing the "right" thing (Andreoni and Bernheim, 2009; Grossman, 2015; Ellingsen and Johannesson, 2011). Bénabou and Tirole (2006), Ellingsen and Johannesson (2008), and Seabright (2009) study how ma-

[5]Conformity in actions in a population is another notion of social norms, known as "descriptive" norms-Bicchieri (2006); Cialdini et al. (1991). These norms, while independent of the analysis in this paper, are important in the behavioral change literature (Berkowitz, 2004).

terial incentives can crowd out these image incentives for prosocial behavior (and related phenomena). Benabou and Tirole (2011) uses this social signaling mechanism to explain the expressive power of laws, and Bénabou et al. (2018) models the trade-off between espousing moral views that encourage behaviors with positive externalities and making excuses for one's own behavior when reputation is based on signaling.

Lastly, a growing literature on counterintuitive effects of social incentives underscores the relevance of the issues discussed in Section 6. Bursztyn et al. (n.d.) finds that single female MBA students report lower desired salaries and lower willingness to travel or work long hours when their responses can be observed by single male peers. Bursztyn et al. (2015) documents a 24% decline in academic performance after the incorporation of a public performance leaderboard in computerized high school courses. Even acknowledging differences in social motivations between, for example, low-income and high-income students does not guarantee successful design of social incentives: in Bursztyn et al. (2017a), social pressure paradoxically decreases SAT prep take-up in *both* populations, because low-income students want to hide effort and high-income students want to high low ability. Along other lines, Kuran (1995) argues that revolutions are inevitably unpredictable due to the hidden heterogeneity in the population caused by a failure of revolutionaries to vocalize their opinions due to social pressure. Centola et al. (2005) develop a computation model based on social influence similar to approval to show how unpopular norms can be enforced. In terms of other mechanisms by which social incentives can backfire, Austen-Smith and Fryer (2005) and Acemoglu and Jackson (2017) provide alternative models of related phenomena and Charness et al. (2014) find that status competition encourages sabotage along with constructive effort.

3 Model

Consider a setting in which an individual within an (observant) population must make a morally contentious choice. Individuals might disagree about which option is the one they *should* take; that is, they have different personal norms. Each option provides an individual a certain consumption utility, which conceptually includes the immediate costs and benefits of the action along with any expected long run change in utility, such as the expected change in tax policy after volunteering to campaign for a particular candidate. This setting is intended to intuitively capture a wide array of moral and customary decisions, such as whether to eat meat, which church to go to (if any), whether to send one's kids to private school, how to share resources, how to reciprocate kind actions, what to wear to work, what brand of shoes

to buy, which political stances to espouse, etc.

I analyze a characteristic individual in this population who is faced with a choice set X. Each option $x \in X$ leads to personal consumption utility $v(x)$, but in addition, an individual i has a personal norm denoting ρ^i as the morally appropriate choice. When making a choice x that deviates from this norm, he pays a psychological cost ("guilt") $G(x - \rho^i)$, which is additionally weighted by an integrity parameter t^i. That is, each person has a two dimensional type (t^i, ρ^i): a personal norm and an integrity parameter that constitutes a weight on their personal norm.

When types are specified in this two dimensional way, we can naturally distinguish between social image regarding ρ and social image regarding t. A respect seeker wants t to be judged favorably, and an approval seeker wants his action x to be judged favorably by observers with ρ close to x.

Social image utility (from either approval or respect) is an increasing function given by $H(m(x))$, where m is the image resulting from a given choice x. The importance of image is determined by a social pressure parameter s, which enters utility as a weight on H. s summarizes the shared attributes of a situation that contribute to a large emphasis on image, such as visibility, audience size, and how harshly choices are judged in a particular setting.

Altogether, an individual i has utility

$$U(x|t^i, \rho^i) = v(x) - t^i G(x - \rho^i) + sH(m(x)). \tag{1}$$

Further assumptions may be helpfully motivated with an example. Imagine that the relevant choice is to argue or vote for a redistributive tax schedule. The decisionmaker believes that ρ represents the best trade-off between equality and efficiency but would rather not pay such a high rate, so he might facetiously argue against social safety nets. Others disagree with his moral beliefs, and some argue in favor of credible alternatives, while some argue for transparently selfish policies that no one believes are fair. While the model abstracts from other forms of social preferences, certain types, like distributional preferences, can be conceptually rolled into v (and they obviously affect norms themselves and the guilt experienced when violating them).

Assumption 1. *i) X is indexed by $\mathbb{R}$.*

ii) $s \geq 0$.

iii) $(t^i, \rho^i) \sim \phi$, with continuous conditional distribution ϕ_t satisfying $\operatorname{supp} \phi_t = \mathbb{R}^+$, and $\operatorname{supp} \phi_\rho \subset X$. These distributions are commonly known.

iv) $G(x-\rho^i)$ is symmetric around 0 and increasing in $|x-\rho^i|$. Normalize $G(0)=0$.

v) H is increasing in $m(x)$, and $\sup_m H(m) = \overline{H} < \infty$.

Continuity of ϕ_t is assumed to guarantee existence of equilibrium; this could be relaxed in specific instances to allow, for example, an atom in the distribution of types. Symmetry of G is clearly violated in many instances (over-generosity surely induces less guilt than under-generosity) symmetry is not at all a critical assumption; it simply reduces the number of cases to consider and thus eases exposition. Infinite support for t ensures that no pooling equilibria exist because types with sufficiently high integrity will always follow their personal norms. This assumption slightly simplifies the analysis but isn't strictly necessary; in particular, appealing to the D1 criterion preserves the basic equilibrium result (Proposition 1) with $\operatorname{supp} t = [0,T]$ (see the Appendix).

The social image function m can embody either respect or approval. In Section 5.3 I will consider individuals who care about both types of image simultaneously, but until then the analysis will compare approval seeking decision makers, with image function $m = m_a$ and corresponding image utility function $H = H_a$, to respect seeking decision makers, with m=m_r and $H = H_r$. Approval and respect are defined as follows:

Assumption 2. *i) $m_a : X \to \mathbb{R}^-$ is defined as $m_a(x) = -\int_{-\infty}^{\infty}\int_0^{\infty}\phi(t,\rho)G(x-\rho)dtd\rho$, the population average judgment of x.*

ii) $m_r : X \to \mathbb{R}^+$ is defined as $m_r(x) = E[t^i|x]$, the rational expectations inference of t^i given choice x.

Approval seekers: An approval seeking individual derives utility from praise for his action, and observers praise actions that agree with *their* personal norms. Note that the approval seeker is not concerned with actually signaling either his personal norm ρ or his integrity t; image is based on his actions directly. This is superficially similar to wanting to signal that you share your beliefs with someone else, but I opt not to use such a signaling model because it immediately leads to counterintuitive predictions: vegetarians would have to approve of admittedly hypocritical, lapsed vegetarians just because they philosophically agree about the merits of vegetarianism. The non-signaling specification is more realistic: it's quite plausible that a vegetarian would be happy to convert an insincere meateater, or that Republican constituents would be happy to continue reelecting a Democrat so long as that representative didn't introduce new taxes. These are scenarios in which observers might confer approval, but *not* respect.

For an approval seeker, according to Assumption 2, observers each judge the decision-maker's choice of x just as they would judge themselves for choosing x, and the decision-maker's social image is the negative of the average of these individual judgments over the full population.[6] For example, if half of the population believes in ρ_1 and half the population believes in ρ_2, then $m(\rho_2) = -\frac{1}{2}G(\rho_2 - \rho_1)$. The best attainable image, $m_a = 0$, only occurs when perfectly adhering to a homogeneous norm.

Respect seekers: For a respect-seeking individual, social image is based on observers' estimate m of his integrity t^i, as defined formally in Assumption 2. Respect-seeking individuals want to be seen as highly motivated to avoid hypocrisy, whatever their personal beliefs. However, since integrity is not directly observable, inferences about integrity must be rational in equilibrium, and optimal choices must anticipate those equilibrium inferences. Note that while choices signal something about ρ^i as well, this does not affect image utility. There are intuitively many scenarios in which consistent adherence to beliefs is emphasized over the beliefs themselves (I can admire vegans without believing that veganism is morally superior to other dietary choices) but the analysis of Section 5.3 allows for individuals to care about both respect and approval simultaneously in situations when complete separation isn't plausible.

Notice that for both approval seekers and respect seekers, there is a maximum possible utility from image. Approval seekers can't do any better than to perfectly please everyone in the population, and respect seekers can't do any better than to be known to be perfectly impartial. It is intuitive that perfect image can't lead to unboundedly high utility, and this upper bound on H (stated in part 5 of Assumption 1) will also provide mathematical utility by restricting equilibrium parameters to a compact space and guaranteeing existence of an equilibrium.

The dependence of utility on beliefs places the model of respect in the realm of psychological game theory (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009). As long as types are exogenously assigned, approval seekers are not playing a strategic equilibrium in which inferences about types matter, so the tools from psychological game theory are not needed to analyze their outcomes.

While approval and respect both encourage the same behavior in situations with homogeneous norms, the distribution of behavior may nonetheless be different under the two types of image motivations. This is fundamentally because respect seekers are playing a signaling

[6] An alternative approach would be for approval to be based on the fraction of the population choosing in the same way. Both approaches have been explored in the literature previously; see Section 2.

game, while approval seekers are being judged directly for their actions rather than inferences based on those actions. As a result, the model of respect can generate pooling equilibria in situations where the approval-seeking model can't. This is demonstrated by Andreoni and Bernheim (2009) (or similarly Bernheim's (1994) model of conformity) which can be seen as a special case of the model of respect seekers in a dictator game setting. Their setting with a unanimous 50-50 split norm shows that respect seekers playing a signaling game create endogenous discontinuities in their preferences that lead to pooling behavior. In particular, they show that respect seekers exhibit pooling on the 50-50 split, despite the fact that preferences have no discontinuities or kinks at this point. If G is convex, approval seekers merely have an increased motivation to approach a fair split when social pressure increases, and the distribution of choices will smoothly approach that point without discontinuously pooling there. If G is concave, approval seekers will also pool on personal norms, as in Michaeli and Spiro (2015).

4 Results

For the sake of readability, from this point I will suppress the i superscripts and analyze a representative individual with type (t, ρ) drawn from a larger population whose aggregate distribution of behavior will be characterized. This section considers a tractable version of the general model above, in which everyone in the population holds one of two personal norms and x must take one of these two values. The next section considers several variations on this basic setting.

4.1 Equilibrium

The main differences between respect and approval arises clearly in this simple binary choice environment. Assume that our representative individual has exactly two options, x_1 or x_2, and one of two personal norms, $\rho_1 = x_1$ or $\rho_2 = x_2$. Each option provides consumption utility $v(\cdot)$; WLOG assume $v(x_2) > v(x_1)$. Guilt is given by $G(x_1 - \rho_2) = G(x_2 - \rho_1) = G$. Additionally, assume that t is distributed according to ϕ_t, independently from ρ (independence will be relaxed in Section 5.1. ϕ_t has full support on $\mathbb{R}^+$ and is continuous, as required by Assumption 1. A fraction $p_1 \in (0, 1)$ of the population has $\rho = \rho_1$.

This captures the running example from the introduction of the choice between omnivorism and vegetarianism, with x_1 representating vegetarianism under the assumption that this is the more difficult choice. For a more stylized example, consider a choice between

two allocations of wealth for the decision maker and a partner. x_2 corresponds to $(3, 0)$ (i.e. 3 units for the decision maker and none for the partner) and x_1 corresponds to $(1, 1)$. If the decisionmaker is utilitarian he will believe x_2 is the fairer allocation, but if egalitarian, x_1 will seem fairer. If he has a low t, however, and doesn't care too much about following his personal norm, he will prefer to choose the personally advantageous allocation x_2.

Whether a respect seeker or an approval seeker, the decisionmaker compares the utility of choosing his personal norm and avoiding guilt to the utility of the guilt-inducing other option. If $\rho = x_1$ he will choose x_1 if $v(x_1) + sH(m(x_1)) > v(x_2) - tG + sH(m(x_2))$, which is true if t is sufficiently large. A similar condition applies if $\rho = x_2$, but the condition is easier to meet since $v(x_2) > v(x_1)$.

The role of H depends on whether the decision maker is an approval seeker (with H_a) or a respect seeker (with H_r). Proposition 1 describes equilibrium for a population of approval seekers, who perform a straightforward utility maximization, and for respect seekers, which emerges in a signaling game. As mentioned above, the dependence of utility on beliefs for respect seekers means that we need a psychological game theoretic equilibrium concept. A signaling equilibrium consists of an action function of types $Q : [0, \infty] \times \{\rho_1, \rho_2\} \to \{x_1, x_2\}$, along with a perception function $P : \{x_1, x_2\} \to [0, \infty]$ with $P(x) = E[t|x]$. Equilibrium transfers must be optimal given P and inferences must be consistent with Q. Throughout this paper, I also restrict attention to equilibria satisfying the D1 criterion of Cho and Kreps (1987), which requires that inferences about types from disequilibrium actions must be reasonable in the sense that, roughly, all weight must be placed on the types who would be tempted to deviate to that action for the widest range of mistaken beliefs.[7]

Proposition 1. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and $\rho \in \{\rho_1 = x_1, \rho_2 = x_2\}$ is independent from t, the following hold:*

1. *Among approval seekers:*

   - *Individuals with sufficiently high t will choose consistently with their personal norms. Lower t individuals will choose whichever option x yields a better combination $v(x) + sH(m(x))$*
   - $m(x_1) > m(x_2)$ iff $p_1 > .5$.

[7] Since $\operatorname{supp} \phi = \mathbb{R}^+$, there is never an off equilibrium path choice since sufficiently high types will always choose in accordance with their personal norm, so the D1 criterion does not refine the result. However, if t is assumed to have an upper bound, Proposition 1 still holds exactly as stated with only equilibria satisfying D1 considered (see the Appendix). Later results will also be substantively refined by the D1 criterion.

2. Among respect seekers:

- *There exists at least one pure strategy equilibrium.*
- *In any equilibrium, individuals with sufficiently high t will choose consistently with their personal norms, and lower t individuals will choose x_2.*
- $m(x_1) > m(x_2)$

The intuition for approval seekers is straightforward. Consumption and image utility are fixed for each option, so people who don't care very much about guilt choose the option with the higher sum of those factors, and people with high enough t stick to their beliefs. The intuition for respect seekers is subtler, since social image depends on aggregate behavior. But, imitation can only ever occur in one direction: If any person defects from their personal norm, then anyone with either lesser t or the opposite ρ will have an even stronger motive to choose the same. And, if imitation is in the direction of choosing x_1, then the average t of people choosing x_1 will be *lower* than those choosing x_2, which would create an unequivocal motivation to defect instead to x_2. So, imitation can only occur as stated in the proposition.

Proposition 1 says that for respect seekers, the material cost of norm adherence is a key determining factor in aggregate behavior. Relatively costly actions will dissuade those with low integrity, leading to a higher social image associated with that choice, and to an overall tendency to choose the cheaper option. This contrasts with the situation for approval seekers, who care about the population distribution of personal norms. If most of the population has $\rho = x$, they are tempted to choose x in order to please their peers. Costliness has no role in social image; it merely factors into individuals' decisions as they trade off cost, image, and guilt.

Proposition 1 also doesn't rule out multiple equilibria for respect seekers in general. The generation of multiple equilibria via social interdependence is common in similar signaling models, for example Brock and Durlauf (2001).

For approval seekers, this result implicitly states that when personal *norms* shift in the population such that the majority belief changes, *behavior* of approval seekers can shift much more dramatically. This may be apparent, for example, in the shifting tide of public opinion about marriage equality. Meta-surveys indicate that 2010 or 2011 was when a majority of Americans first supported marriage equality, but the shift has been slow and steady (Silver, 2011). Support among senators, however, has changed much more dramatically, and more quickly than can be accounted for by turnover: only 15 senators openly supported marriage equality in 2011, and 51 did as of April 2013 (Matthews, 2013). Since senators derive utility

(re-election) exactly from pleasing the largest fraction of the population, approval seeking is a likely explanation for at least part of this phenomenon. Similar forces may be behind sudden changes in taboos, such as political correctness. Behavior can even appear to be nearly unanimous, but heterogeneous beliefs are simply hidden due to high social pressure (Kuran, 1995, e.g.).[8]

4.2 Changes in social pressure

While this model is static, the response of aggregate behavior as s changes is of interest so that we might understand how social pressure can be employed to influence norm adherence. Proposition 2 summarizes the high pressure equilibria for both approval seekers and respect seekers in the same setting as Proposition 1:

Proposition 2. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$ and $\rho \in \{\rho_1 = x_1, \rho_2 = x_2\}$ is independent from t, the following hold:*

1. *For approval seekers, as $s \to \infty$, the fraction of the population choosing x_1 (x_2) approaches 100% if $p_1 > .5$ ($p_1 < .5$).*

2. *For respect seekers, as $s \to \infty$, the fraction of the population choosing according to their personal norms approaches 100%, and $m(x_2)$ approaches $m(x_1)$.*

Once again, the intuition for approval seekers is very straightforward: as social pressure increases, the image component of utility dominates consumption and guilt more and more thoroughly, until (in the limit) no one is willing to choose anything other than the more approved-of choice. For respect seekers, the intuition builds on the understanding of Proposition 1. This result showed that the costlier option is always associated with a higher image since it is the only option that can serve as a costly signal of integrity. When social pressure rises, then, more and more people are tempted to try to signal in this way. The first to switch to x_1 are the ones who already feel guilty for choosing x_2. The ones who feel most guilty and are quickest to switch are the ones who have t just barely lower than anyone else who is choosing x_1, and so as they switch, the signaling value of choosing x_1 is slightly diluted. As social pressure continues rising and people with even lower t's and ρ_1 switch, the signaling value is eventually diluted to the point that $m(x_1)$ is indistinguishable from $m(x_2)$. At this point there is no remaining reason for those with $\rho = x_2$ to switch.

[8] Of course, conformity masks heterogeneity in opinions even in homogeneous norm models, but heterogeneity in those models is in the degree to which individuals prioritize adherence with the norm, rather than in normative beliefs themselves.

Put another way, as social pressure increases, cost disparities between actions become irrelevant for respect seekers, and somewhat counterintuitively, either action will lead to approximately the same image. On the other hand, approval seekers in the same scenario will become more and more conformist to the modal norm, as defectors become more and more harshly shunned.

Figure 1 illustrate the results of Propositions 1 and 2, showing that social pressure can potentially push behavior of respect seekers and approval seekers in opposite directions and that different sets of equilibria are possible.

Figure 1: Approval seekers' versus respect seekers' choices

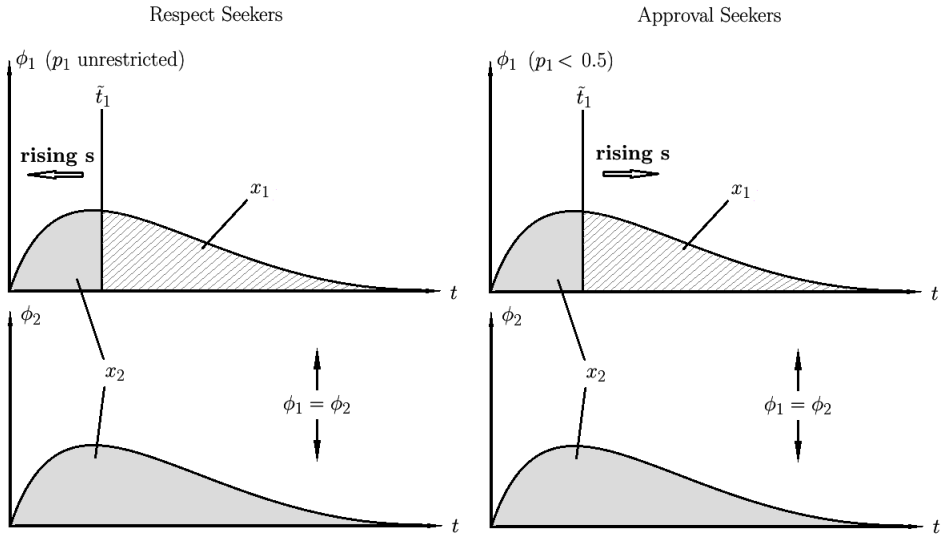


Illustration of choices made by the distributions (of t) of approval seekers (right) and respect seekers (left) with personal norm x_1 (top) and x_2 (bottom). Types in the shaded regions of the distributions choose x_1, and types in the filled regions choose x_2. Increasing social pressure causes more approval seekers with ρ_1 to choose x_2 when $p_1 < 0.5$, but increasing social pressure (overall, although perhaps not for small changes) pushes respect seekers in the opposite direction.

Our stylized example of wealth allocation can make the contrast between respect and approval seekers clearer. Respect-seeking egalitarians believe that $(1, 1)$ is fair despite selfishly wishing to choose $(3, 0)$. Only the egalitarians who care enough about fairness will choose $(1, 1)$. Realizing this, observers admire the integrity represented by the choice of $(1, 1)$ even though utilitarians have equally strong beliefs as egalitarians on average. If social pressure is very high, more egalitarians will choose $(1, 1)$ due to the added benefit of a better social image. In the limit, all egalitarians choose $(1, 1)$, all utilitarians choose $(3, 0)$, and remarkably,

no one is respected more than anyone else or suspected or hypocrisy.

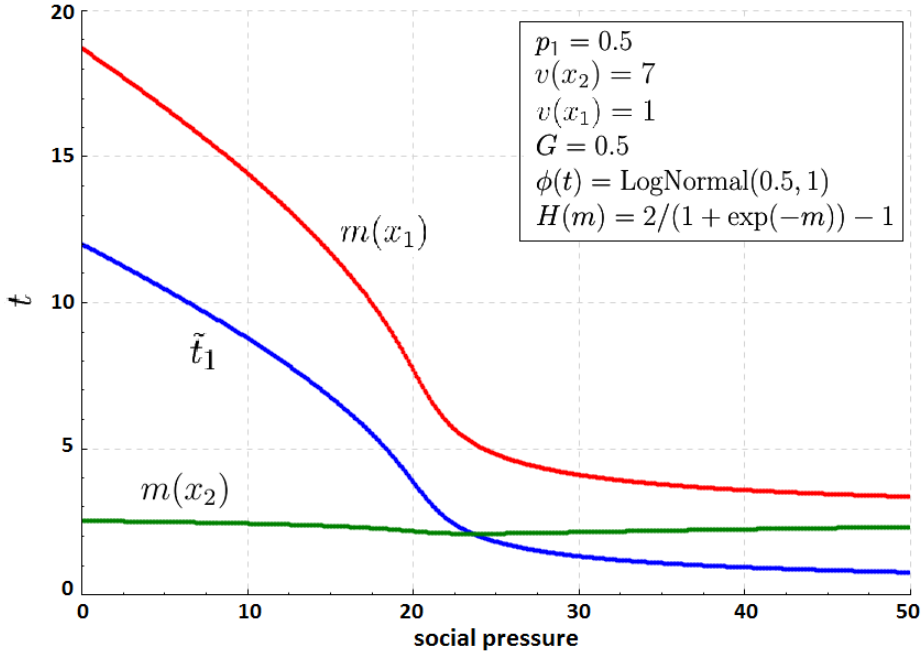
Approval seekers, on the other hand, don't try to discern each others' hypocrisy. If most people are born utilitarian, approval seekers will try to go along with the group, and more so the higher social pressure is. In the limit, only the most extreme egalitarians dare to follow their own moral compass in the face of extreme criticism.

Many tactics (more subtle than attempting to change norms directly) used to encourage certain behaviors are understandable with these results. Shaming the rich for self-interestedly voting for low taxes is clearly an attempt to hurt their credibility (i.e. to confer disrespect) and discourage that kind of hypocrisy. The model predicts limited success with this tactic up until the point when voting is sincere, expectations reflect this, and accusations of hypocrisy are no longer credible. Influencing behavior through approval relies on having the majority norm. The Human Rights Campaign rather explicitly acknowledged this after beginning a new campaign following the change in the majority opinion of marriage equality in around 2011. Admitting limited success with changing minds directly, they switched to changing perceptions of $p > 0.5$ by "trying to foster the sense that ... history has already ruled in favor of their cause" (Issenberg, 2013).

Proposition 2 also states that the relative prestige of the two options changes with social pressure for respect seekers, but not approval seekers. For respect seekers, in settings with extreme social pressure, the image associated with any choice is approximately the same in equilibrium. But when social pressure is lower, costly actions are uniquely admired as true signs of integrity. Religion may be an example of this phenomenon. Religious fakery seems to be judged harshly, indicating high social pressure. Correspondingly, members of reformed denominations are not assumed to be betraying their true orthodox beliefs simply because the rules are too onerous, and atheists are generally assumed to have principled motives. On the other hand, social pressure over dietary habits is not so strong that saying "I admire vegetarians, but I don't want to give up my steak." necessarily attracts horrified looks. In this domain, even though there are plenty of people who honestly think eating meat is the right and natural thing to do, vegetarians project an image of moral integrity more than omnivores.

Figure 2 shows an example of these relationships, with parameters chosen so that a unique equilibrium exists at all levels of social pressure.

Figure 2: The effect of social pressure on respect seekers



An example of the model with the specified parameters. The solid curve shows the equilibrium cutoff value $\tilde{t}_1$ defining the minimum t for types with ρ_1 who choose x_1, as a function of social pressure s. As s rises, a smaller fraction of the population will act hypocritically. The bottom line shows $m(x_2)$ and the top line $m(x_1)$, both equilibrium values as a function of s. Note that as social pressure rises, the gap between $m(x_1)$ and $m(x_2)$ tends to shrink, so that in the limit either action will yield the same level of respect. $v(x_1) = 1$ $v(x_2) = 7$

4.3 Sacrificial equilibria

What are the welfare implications of these behavioral responses? The model does not specify the impact of choices on others' utilities, so without specifying material externalities, or how moral hypocrisy affects others, welfare isn't clearly defined. Settings both with externalities (wealth distribution) or with no or minor externalities (vegetarianism) are possible to understand with the model, but a richer, context specific analysis would be needed for welfare judgments.

But there are still welfare-related effects worth noting. Existing models of social norms and/or social pressure predict that individuals may sacrifice material utility to adhere to a norm, but this unambiguously welfare-destroying because the individual is gaining moral utility. And obviously when someone fails to follow their personal norm in order to obtain greater material utility, this is understandable as potentially welfare-enhancing. But it's

harder to defend an outcome in which an individual defects from his norm *and* sacrifices material utility in order to do so. This individual would clearly prefer a lack of social pressure and is sacrificing utility to adhere to a norm he doesn't believe in. Define a "sacrificial" equilibrium as follows:

Definition 1. *A population's equilibrium choices constitute a* sacrificial equilibrium *when some individuals with ρ nonetheless choose x with $v(x) < v(\rho)$. (And an equilibrium is said to be more sacrificial when the fraction of the population who does this rises.)*

These sacrificial equilibria are surprising from either the perspective of classical economics *or* from models of homogeneous norms. Nonetheless, Proposition 3 states that approval seekers are prone to sacrificial equilibria when the costly action is the majority norm, and moreso when social pressure rises (and similar phenomena have been predicted in models employing an approval-like form of social influence (Michaeli and Spiro, 2015; Centola et al., 2005)). Respect seekers, on the other hand, are never able to sustain a sacrificial equilibrium.

Proposition 3. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, and $\rho \in \{\rho_1 = x_1, \rho_2 = x_2\}$ is independent from t, the following hold:*

1. *For approval seekers, if $p_1 > .5$, equilibrium is increasingly sacrificial when s gets sufficiently high.*

2. *For respect seekers, equilibrium is never sacrificial.*

The intuition for approval seekers is trivial: as long as the majority believes in a more costly action, it can pressure the minority into a sacrificial equilibrium. Respect seekers are disinclined towards conformity and thus immune to sacrificial equilibria: if someone with ρ_2 *were* willing to choose x_1 to signal a high t, this would dilute the value of the signal too much to be worthwhile. Approval seekers are thus more likely to be able to sustain an equilibria in which, for example, individuals choose allocations of $(1, 1)$ over $(3, 0)$ despite the fact aggregate welfare is reduced beyond the point supported by egaliatarians in the population.

5 Robustness and Extensions

Now we can examine the robustness of these basic results by relaxing various restrictions of the simple binary model. To summarize the results so far:

1. Both respect seekers and approval seekers stick to their beliefs if they have a high enough t. But lower-t type approval seekers choose whichever option gives them the best combination of material and image utility, and lower-t type respect seekers choose the option that gives them the best material utility.

2. Respect seekers, but not approval seekers, sometimes have multiple equilibria available.

3. Respect seekers, but not approval seekers, always obtain greater social image utility from the higher cost option.

4. As social pressure rises, respect seekers become more divided along norm lines, while approval seekers tend towards conformity on the majority norm.

5. As social pressure rises, approval associated with either choice remains the same, but the difference in the respect associated with each option shrinks (overall[9]).

6. Approval seekers, but not respect seekers, can sustain sacrificial equilibria in which individuals sacrifice *both* their personal norms and their material outcomes for the sake of image.

5.1 Correlated type parameters

The initial analysis was simplified by the assumption that t and ρ are independent. Instead, now assume the same binary choice setting, but distinguish between the distributions of types t among those with ρ_1 and ρ_2. As before, fraction $p_1 \in (0,1)$ has ρ_1. This leads to one norm being elite in the sense that it is associated with people of higher average integrity.

The statement of Part 1 of Proposition 1 (describing equilibrium for approval seekers) doesn't change with this adjustment. Proposition 4 describes the behavior of respect seekers:

Proposition 4. *If* $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, *and* $\rho \in \{\rho_1 = x_1, \rho_2 = x_2\}$ *with* $t|\rho_1 \sim \phi_1$, $t|\rho_2 \sim \phi_2$ *then for respect seekers:*

[9]It's important to note that results regarding changes in social image are, for respect seekers, true only for large (or limit-case) changes in s. The difference in respect associated with x_1 and x_2 is certainly decreasing overall as s increases from 0 to ∞, but not necessarily monotonically. Figure 2 demonstrates why. When $s = 0$, large numbers of people with ρ_1 defect to x_2 and the respect associated with x_1 is much larger than with x_2. As s rises, $m(x_1)$ decreases monotonically as fewer people with ρ_1 defect. $m(x_2)$, however, initially also falls because the first people to revert to x_1 are those with relatively high t. This is a consequence of modeling respect as the inference of t, and would be different, for example, in a model in which respect is a function of the likelihood that $x = \rho$ given x. The formulation of respect in this paper effectively forgives the choice of a materially beneficial option if the alternative is extremely costly and only extremely high-t types choose it. I thank an anonymous referee for pointing this out.

1. *There exists an equilibrium, and in any equilibrium, sufficiently high-t types choose according to their personal norms, and lower types all choose either x_1 or x_2.*

2. *As $s \to \infty$, if $E[\phi_1] > E[\phi_2]$ ($E[\phi_2] > E[\phi_1]$), an individual will only choose x_2 (x_1) if he has $\rho = x_2$ ($\rho = x_1$) and sufficiently high t.*

3. *If $E[\phi_1] > E[\phi_2]$, social pressure that is sufficiently high can sustain a sacrificial equilibrium.*

Part 1 is very similar to part 2 of Proposition 1, but correlation between type parameters now allows for imitation to occur in either direction. This is because there are now two mechanisms with which to signal integrity: acting in accordance with the elite norm, or choosing the costly action. High integrity individuals will of course still follow their personal norms, but low integrity types (with either personal norm) will now all choose whichever option provides a better combination of image and material utility. This is a slight modification to result 1, in which low types abandon their personal norms only due to material utility differences.

Part 2 states that as social pressure rises, contrary to Section 2, costly actions are only useful signals to the extent that that action is *a priori* associated with high integrity. As social pressure rises, the benefit of mimicking the elite norm will outweigh any difference in consumption utility, so that imitation occurs only in the direction of the elite norm. This is a slight adjustment to result 4; behavior still becomes strictly divided along normative lines, but with a few individuals pretending to hold the elite norm (in order for result 5 to remain exactly the same).

Part 3 establishes that even respect seekers are capable of sustaining sacrificial equilibria. These equilibria are substantially different from the approval seekers' sacrificial equilibria discussed in Section 4.3), however: approval seekers are sacrificial when the majority believe in a costly policy or action; respect seekers are sacrificial when particularly high integrity types disproportionately believe in a costly action. And, high enough social pressure can lead arbitrarily high-t approval seekers with ρ_2 to choose x_1, but rational expectations restricts sacrificial behavior among respect seekers to a limited set of low t types, no matter how high social pressure gets.

The correlated type case requires some tweaks to how the results above are specified but the overall picture of the disparate behavior of approval and respect is similar. Theoretically, however, it also helps us think about heterogeneity in consumption utility within this model. v is assumed to be the same for all people, which is easily so long as v is observable: A

vegetarian who is known to dislike meat will simply have that fact incorporated into the inference of their t. Invisible heterogeneity in v complicates matters. If ordinal rankings are at least consistent across people, we can rescale utility functions such that v is homogeneous, which distorts t and causes people to seem differentially responsive to s, but limit results will still hold. But if this heterogeneity in v is additionally correlated with (t, ρ), the distortion in t effectively introduces correlation between t and ρ. But we can then appeal to the results with correlated parameters from Section 5.1 to claim that limit results will hold. If not even rank orderings are consistent across the population, inference becomes very difficult; this is out of the scope of this paper.

5.2 Unanimously immoral options

Another natural robustness check on the previous results is to allow other options than the ones that correspond to norms. At the very least, a binary choice often admits a third option: abstention. In other cases, opposing sides often have the opportunity to compromise on an option that neither believes in but both can accept. As it turns out, this doesn't substantially change the basic results, but does lead to new insights on the nature of compromise.

I analyze a ternary choice setting, but the intuition of the results would also apply to any richer discrete choice set or set of norms, such as a discretization of a continuous choice set. Consider a setting in which the decision maker chooses x from $\{x_1, x_2, x_3\}$. ρ is either $\rho_1 = x_1$ or $\rho_3 = x_3$, and x_2 is a middle ground option, and to reduce the number of cases, assume $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G_2 > G_1$. As before, fraction $p_1 \in (0, 1)$ have ρ_1 and the remainder have ρ_3. WLOG, let $v(x_3) > v(x_1)$. In particular, $v(x_2)$ can take any value relative to $v(x_1)$ and $v(x_3)$, although this will of course affect which form of equilibrium among the options described in Proposition 5 is possible. As in Section 1, ρ and t are independent.

Proposition 5 describes the equilibrium:

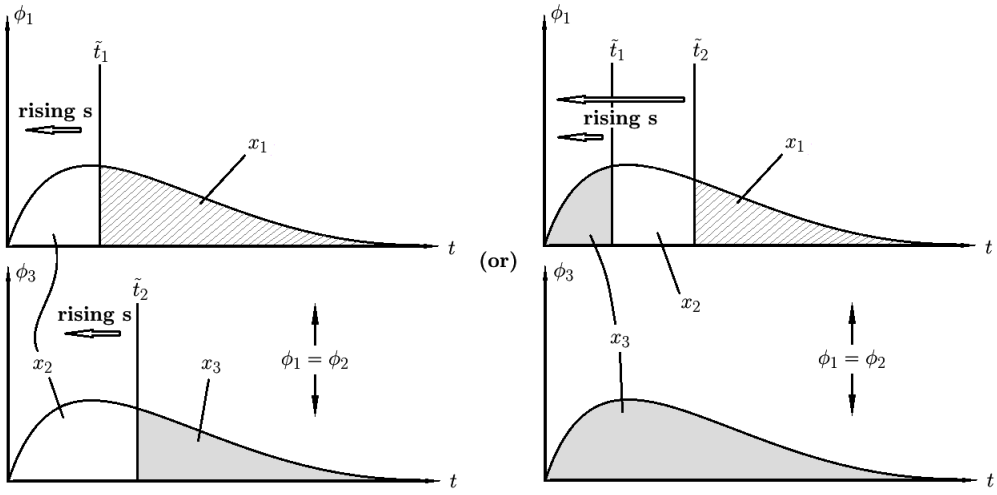
Proposition 5. *If $X = \{x_1, x_2, x_3\}$ with $v(x_3) > v(x_1)$, $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G_2 > G_1$, and $\rho \in \{\rho_1 = x_1, \rho_3 = x_3\}$ is independent of t, then the following hold:*

1. *Among approval seekers, individuals with sufficiently high t adhere to their personal norms and lower types either 1) all choose x_1, 2) all choose x_2, 3) all choose x_3, or 4) mid-level-types with ρ_3 (ρ_1) choose x_2 and all lower types choose x_1 (x_3).*

2. *Among respect seekers, there exists at least one pure strategy equilibrium. In any equilibrium, individuals with sufficiently high t will choose consistently with their personal norms. Among lower t individuals, either: 1) all defect to x_2, 2) all choose x_3, or 3) mid-level-t types with ρ_1 choose x_2 and all lower types choose x_3.*

Figure 3 illustrates two possibilities for respect seekers, one in which all low-t types defect to the middle option, and one in which everyone with ρ_3 always chooses x_3. There is an additional possibility in which neither type chooses x_2. Equilibrium for approval seekers takes a similar form, but a dominant majority norm may drive low-t types to choose x_1 despite its lower material utility. The two cutoff values $\tilde{t}_1$ and $\tilde{t}_2$ shown in figure 3 can therefore, for approval seekers, lie in either the top (ϕ_1) or bottom (ϕ_3) distribution and can move in either direction as social pressure rises.

Figure 3: Respect seekers' compromise



Two (non-comprehensive) possibilities for respect seekers with an opportunity for compromise. The distribution of t among those with $\rho = x_1$ is shown on top, and with $\rho = x_3$ on bottom. The left side possibility occurs when types with either ρ choose x_2 but x_2 disappears as social pressure increases. The right side shows another possibility when ρ_3 is perfectly adhered to. Arrows show the overall movement as s approaches ∞; small changes in s may cause small shifts in either direction.

The intuition behind these results is quite similar to that of Proposition 1, but the addition of a third option, when $v(x_2)$ is at least as large as $v(x_1)$, provides some low types with one or both of the norms with a tempting option that doesn't induce as much guilt as defecting all the way to the other norm.

Proposition 6 provides the high social pressure result analogous to Proposition 2.

Proposition 6. *If $X = \{x_1, x_2, x_3\}$ with $v(x_3) > v(x_1)$, $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G_2 > G_1$, and $\rho \in \{\rho_1 = x_1, \rho_3 = x_3\}$ is uncorrelated with t, then the following hold:*

1. *As s increases, first x_2 will cease to be chosen by any respect seeker, and in the limit as $s \to \infty$ all choices follow personal norms.*

2. *As $s \to \infty$, approval seekers approach perfect conformity with one of the three options.*

Rather than modifying any of the basic results listed above, we can add to the list. As before, approval seekers are more prone to conformity than respect seekers, even conforming to the compromise option that no one believes in in some cases if it's better to partly appease everyone than to perfectly please one group and displease the other (even if compromise is costly). Respect seekers, however, will *only* choose to compromise if it offers a large enough monetary reward and there isn't too much social pressure.

5.3 Simultaneous respect and approval motives

A final natural robustness check on the basic results is to allow respect-seeking and approval-seeking motivations to interact. Respect and approval are not intended to be competing notions of social image, but likely are relevant to varying relative strengths in different contexts/individuals/populations. Like the compromise option of the previous subsection, this turns out not to substantially change the basic results, but does lead to new insights.

The model as defined above includes a single term for social image, which takes on either an approval or respect functional form. I adjust this in the natural way to model individuals who care about respect and approval simultaneously.

$$U(x|t^i, \rho^i) = v(x) - t^i G(x - \rho^i) + s_a H_a(m_a(x)) + s_r H_r(m_r(x))$$

In the binary choice setting of Section 4, at higher levels of s_a the utility from approval dwarfs consumption utility, and imitation must occur in the direction of higher approval. At the same time, increasing s_r reduces imitation, as people try to convincingly signal their integrity. The net effect may look like *either* of the equilibria described in Proposition 1:

Proposition 7. *If individuals are motivated by both approval and respect, $X = \{x_1, x_2\}$ with $v(x_2) > v(x_1)$, and ρ and t are independent, the following hold:*

1. *Low-t types choose x_2 for sufficiently small s_a but choose* $\arg\max H_a(m_a(x))$ *for larger s_a.*

2. *At a fixed level of s_a, increasing s_r in the limit leads to perfect adherence to personal norms.*

3. *At a fixed level of s_r, increasing s_a in the limit leads to perfect conformity with majority opinion.*

The intuition of this result has one key difference from Section 4.1: accepting social disapproval can *itself* be a signal of integrity. Without approval motivations, despite material motivations that favor x_2, those with personal norms ρ_1 and high t choose x_1, and as social pressure increases, those with low t do as well. In Proposition 7, the same happens but now x_1 may additionally be disadvantaged in terms of approval, which dissuades more low-t types from choosing it. Observers therefore infer a higher t when someone chooses x_1. Individuals who choose it as respect-based social pressure rises are therefore doing so because its costliness, in terms of material utility *and* approval, make it an effective signal of integrity.

This interaction between image motivations, perhaps pushing in opposite directions, may explain the seemingly arbitrary costly signals that individuals take to declare their identity convincingly. For example, teenagers might wear goth clothing in order to be shunned by the majority, thereby convincingly signaling their devotion to their social group. This rhetoric is also frequently used to promote evangelism, all the way back to the Jesus' Sermon on the Mount: "Blessed are those who are *persecuted because of righteousness*, for theirs is the kingdom of heaven." (Matthew 5:11, emphasis mine). Accepting disapproval proves the depth of your faith which is the ticket to heaven.

Combining both types of social image into a single model clarifies the prescriptive discussion of the previous subsections. "Social pressure" is a general concept combining many things into a single parameter, and these components don't necessarily increase or decrease unilaterally. It may be possible to target approval-based social pressure separately from respect-based social pressure. On the other hand, those components of social pressure (anonymity, publicity) that cause people to care more about both approval and respect indicate that complete separation is probably impossible. Attempting to manipulate behavior with those aspects of social pressure may then be difficult process when respect and approval are countervailing forces. See Section 6 for further discussion.

6 Discussion

This model is intended to be applicable to a wide range of domains. Most decisions governed by moral guidelines, customs and traditions, or fads and fashions (a very wide range!) fall into an ambiguous domain in which people disagree about appropriate actions. In addition to the various examples referred to throughout the discussion, these models can be used to understand behavior such as middle school fads, child-rearing practices and taboos, religious practice, etiquette, local customs, and so on ad infinitum. Specific contexts of course require care to adapt the approach to the details of a setting, and this is beyond the scope of the discussion, but I briefly highlight some possibilities of particular importance or interest to economists here.

Political Economy: Most obviously, the results are applicable to partisan politics and persuasion, which have been occasionally alluded to throughout the paper. Additionally, Propositions 5 and 6 may specifically be relevant to political polarization. Polarization has been shown to be increasing in the United States, at least among political elites (McCarty et al., 2006), over at least the last half century. While attention has focused on potential demographic reasons for increasing polarization (for a review, see Layman et al. (2006)), this model points to social pressure as another possible explanation.

Either respect or approval motivations can be related to polarization through the channel of changing social pressure, which may come from any number of sources, such as media penetration, political literacy, changing formal or informal institutions that incentivize party loyalty, information technology, etc. The theory points in this direction as a possibility to be checked empirically, or at minimum forms a theoretical basis for existing theories of polarization that can be thought of as based on social pressure. In fact, it could also explain a situation in which polarization increases among political elites despite fairly constant polarization within the electorate, which some authors claim is the case in the United States (Fiorina and Abrams, 2008) although the evidence is somewhat mixed (Abramowitz and Saunders, 2008). This poses a puzzle for many potential explanations for polarization, but can be understood within these models of social pressure because politicians and citizens are naturally subject to different types of social pressure.

Similarly, Propositions 5 and 6 are clearly relevant to group decision making. In terms of voter mobilization, while it's almost universally agreed that you *should* vote, abstention represents a potential "compromise" option for those who don't want to take sides. A respect seeker might be tempted to abstain to avoid the hassle, but this action would never be socially

rewarded. On the other hand, an approval seeker might be tempted to abstain if he doesn't want to be shunned by his friends of either political party. On aggregate, voter turnout could plummet. In fact, the HRC has admitted to trying to exploit this phenomenon strategically as they promote the idea that history has already decided in favor of marriage equality: "It just deflates them. People who may disagree with [gay marriage] but believe it may happen anyway are hard people to mobilize" (Issenberg, 2013).

Development: As mentioned in the introduction, a particularly important domain that can potentially put the models of respect and approval to good use is development economics. Many development initiatives are beholden to or stalled by local norms which either interfere with incentives or are directly the cause of behavior that an initiative aims to change. For example, interventions designed to reduce HIV transmission must deal with different norms for safe sex practices (Macintyre et al., 2001), pregnancy prevention relies on strong norms allowing women to demand the use of contraception (Caldwell and Caldwell, 1987), women can only advance in society if their parents are willing to send them to school (Fuller et al., 1995), and wealth accumulation and capital investment are only possible if the property rights of the culture allow it (Svensson, 1998). The list of such examples is endless.[10] And even when norms are superficially unanimous, pushback against outside intervention can itself induce heterogeneity in the context of a development initiative.

While many individuals may agree with the alternative norms that are promoted by a development initiative, they are still beholden to the social image motivations that enforce the local norm. Understanding the mechanics of these motivations is critically important to designing effective, persuasive interventions. The above results show that if individuals are respect seekers and the desired action is costly, increasing visibility may encourage those who agree to comply. This is the approach taken, for example, by Cameron et al. (2013), which attempts to reduce open defecation by making usage of the sanitation system visible. On the other hand, if individuals are approval seekers and the desirable norm has yet to take significant hold, or if individuals are respect seekers and the *un*desired action is costly, peer visibility and social pressure should be minimized. An initiative relying on social pressure to encourage good behavior may *backfire*, as it did in the education setting of Bursztyn et al. (2017a).

The case of Kremer et al. (2009) contains a convenient illustration of the relative effectiveness of an education initiative when there is more or less heterogeneity in norms. They

[10] Aldashev et al. (2011) also discusses some of these examples through the lens of wanting to change norms, but models norm adjustment through direct institutional changes.

analyze the effect of NGO scholarships offered to girls in two districts in Kenya. In the Busia district, 90% of teachers claimed that parents of students were positive towards the NGO, while in the Teso district this number was only 58%. There was therefore more pressure in the Teso district to drop out of the program or refuse the scholarship (as several schools, and one scholarship winner, did). In the end, not only did this high attrition rate jeopardize many students' chances at a scholarship, the impact on education attainment from the scholarship opportunity was much smaller in the Teso district. Part of this gap might be explained by respect-seeking students who were refusing a valuable program in order to signal dedication to their belief that outsiders are not to be trusted, or by approval-seeking students who caved to community pressure to resist outsiders.

If a beneficial norm has yet to take hold to *any* significant degree, despite apparent benefits even on an individual level, the development economist might also look to the results on sacrificial equilibria in Section 4.3. Social pressure might masking a degree of heterogeneity in beliefs by enforcing a destructive norm of bad sanitation, or violent civil conflict, or expensive religious sacrifice, or refusing contraception. And if social pressure is too high, it might be difficult to break the cycle. Take, for example, the norm of having expensive funerals. In several African countries, this has clearly reached the status of a destructive, welfare reducing norm; funeral costs are a major cause of families falling into poverty (Case et al., 2008; Krishna et al., 2004). If heterogeneous beliefs are being masked by high social pressure to the extent that this situation represents a sacrificial equilibrium, reducing social pressure is unambiguously predicted to improve the situation.

In sum, understanding local norms is always an important part of program design, but this model of heterogeneous norms clarifies important specific pitfalls to ensure against.

Specific image targeting: It's unlikely that only respect-seeking or approval-seeking motivations are in effect in any particular setting, although the relative strengths surely vary. Political activists then may wish to refer to the model of Section 5.3 by *separately* targeting respect or approval. Take, for example, the efforts to enroll young people in health insurance under the Affordable Care Act. The partisan reputation of the ACA may discourage young Republicans from enrolling, on principle, despite the fact that taking advantage of the new subsidies is clearly the profitable option for most of them. Democrats, on the other hand, may see political reputation and following their own beliefs as additional advantages to enrolling, on top of the actual subsidies. They are therefore in no danger of not enrolling. Since the concrete incentives already push young people in the direction of enrolling, Propositions 2 and 7 show that the respect-based social image of the choice should be minimized,

by downplaying its partisan nature or keeping enrollment decisions as low pressure and low visibility as possible. We saw this tactic in action in enrollment campaigns that emphasized state-specific names for ACA implementations and avoided mentioning the association with "Obamacare".

Marketing: One can take a different view of the results in Sections 4.1 and 2 as indicating to companies or groups how they should price and market products that may be used to express identity or beliefs or to support moral beliefs, such as group membership fees, "Save the rainforest" products, brand name clothing, etc. A product that is associated with a norm that is judged according to respect (exactly how this association arises is unfortunately decidedly out of the scope of this paper) could be successfully marketed as a way to demonstrate one's commitment to that norm. The wildly popular fad of "What Would Jesus Do" bracelets in the late 90's is a clearly successful example of this: rather than, for example, a subtle necklace that could be easily hidden from view, or any other equally effective reminder, gaudy bracelets in every neon color were paid for and proudly worn by masses of Christian teens. On the other hand, graphing calculators are a plausible approval-related example: "nerd" is a label teens usually want to avoid except among their likeminded friends, and while calculators look nearly identical from the outside, and come only in a subtle grey or black, programmable TI-89s or Reverse Polish Notation HP models still serve as status symbols within these cliques (so I hear). Texas Instruments is smart not to sell their fanciest models in bright yellow.

7 Conclusion

In this paper, I developed a model of social image motivations that influence moral choices when the population is divided as to what is right. When people disagree about the appropriate action, two natural possibilities arise for the meaning of social image: people may wish to signal their adherence to their personal norm, or they may wish for others to admire their choices. These alternatives lead to substantially different predictions. This work provides a platform for future work on social image in the presence of disagreement over norms in general settings and provides a foundation for rigorously understanding social image motivations in many real world contexts that have previously been out of reach of the social preferences literature, such as partisan politics, contentious moral choices, customs and taboos. Prescriptively speaking, it provides a theoretical basis for wisely designing institutions and/or interventions that anticipate the effect on the social pressure dynamic and

result in the desired behavioral response. It immediately reveals the risks in ignoring the distinction between types of social image or heterogeneity in norms, and points to better alternatives when an initial approach fails due to targeting the wrong motivation.

Rigorously studying how these models play out in practice will also require empirically determining the contexts in which each model is applicable. Surely, people are motivated both by approval and by respect in different relative amounts in different scenarios, as touched on in Section 5.3. A likely possibility is that approval seeking is a more salient motivation when externalities of choices are large. On the other hand, Thomas Jefferson seems to prescribe approval-seeking and respect-seeking motivations to different classes of decisions when he said "In matters of taste, swim with the current; in matters of principle, stand like a rock." Characterizing the domains in which each model is applicable is an open empirical question and left for future work, but these models form an analytical foundation for beginning this research agenda.

References

Abramowitz, Alan I. and Kyle L. Saunders, "Is Polarization a Myth?," *The Journal of Politics*, March 2008, *70* (02), 542–555.

Acemoglu, Daron and Matthew O. Jackson, "Social Norms and the Enforcement of Laws," *Journal of the European Economic Association*, 2017, *15* (2), 245–295.

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

_ **and Rachel E. Kranton**, "Economics and Identity," *Quarterly Journal of Economics*, 2000, *CXV* (3), 715–753.

_ **and** _ , "Identity and schooling: Some lessons for the economics of education," *Journal of Economic Literature*, 2002, *40* (4), 1167–1201.

_ **and** _ , "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, January 2005, *19* (1), 9–32.

Aldashev, Gani, Imane Chaara, Jean-Philippe Platteau, and Zaki Wahhaj, "Using the law to change the custom," *Journal of Development Economics*, March 2011, *97* (2), 182–200.

Anderson, Lisa R., Jennifer M. Mellor, and Jeffrey Milyo, "Inequality and public good provision: An experimental analysis," *The Journal of Socio-Economics*, June 2008, *37* (3), 1010–1028.

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

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

Ashraf, Nava, Oriana Bandiera, and Kelsey Jack, "No margin, no mission? A Field Experiment on Incentives for Pro-Social Tasks," 2012. Working Paper.

Austen-Smith, David and Jr. Fryer Roland G., "An Economic Analysis of "Acting White"," *Quarterly Journal of Economics*, 2005, *120* (2), 551–583.

Babcock, Philip, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer, "Letting Down the Team? Evidence of Social Effects of Team Incentives," 2010. NBER Working Paper Series No. 16687.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Social Preferences and the Response to Incentives: Evidence from Personnel Data," *The Quarterly Journal of Economics*, August 2005, *120* (3), 917–962.

_ **,** _ **,** **and** _ , "Social Incentives in the Workplace," *Review of Economic Studies*, April 2010, *77* (2), 417–458.

Bartling, Björn and Urs Fischbacher, "Shifting the Blame: On Delegation and Responsibility," *Review of Economic Studies*, August 2011, *79* (1), 67–87.

Battigalli, Pierpaolo and Martin Dufwenberg, "Dynamic psychological games," *Journal of Economic Theory*, January 2009, *144* (1), 1–35.

Benabou, Roland and Jean Tirole, "Laws and Norms," Working Paper 17579, National Bureau of Economic Research November 2011.

Benjamin, Daniel J., James J. Choi, and A. Joshua Strickland, "Social Identity and Preferences," *American Economic Review*, September 2010, *100* (4), 1913–1928.

Berkowitz, Alan D., *The social norms approach: Theory, research, and annotated bibliography* 2004.

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

Bicchieri, Christina, *The Grammar of Society: The Nature and Dynamics of Social Norms*, New York: Cambridge University Press, 2006.

Bicchieri, Cristina, "Norms, preferences, and conditional behavior," *Politics, Philosophy & Economics*, August 2010, *9* (3), 297–313.

Bohnet, Iris and Bruno S. Frey, "The sound of silence in prisoner's dilemma and dictator games," *Journal of Economic Behavior & Organization*, 1999, *38* (1), 43–57.

Bose, Gautam, Evgenia Dechter, and Gigi Foster, "Behavioral coordination as an individual best-response to punishing role models," *Journal of Economic Behavior and Organization*, 2017.

Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson, "Is generosity involuntary?," *Economics Letters*, January 2007, *94* (1), 32–37.

Brock, William A. and Steven N. Durlauf, "Discrete Choice with Social Interactions," *Review of Economic Studies*, April 2001, *68* (2), 235–260.

Buckley, Edward and Rachel Croson, "Income and wealth heterogeneity in the voluntary provision of linear public goods," *Journal of Public Economics*, May 2006, *90* (4), 935–955.

Burks, Stephen V and Erin L. Krupka, "A Multimethod Approach to Identifying Norms and Normative Expectations Within a Corporate Hierarchy: Evidence from the Financial Services Industry," *Management science*, 2012, *58* (1), 203–217.

Bursztyn, Leonardo, Georgy Egorov, and Robert Jensen, "Cool to be Smart or Smart to be Cool? Understanding Peer Pressure in Education," 2017.

_ , _ , and Stefano Florin, "From extreme to mainstream: How social norms unravel," 2017. NBER Working Paper Series No. 23415.

_ , **Robert Jensen, Leigh Linden, and Aprajit Mahajan**, "How Does Peer Pressure Affect Educational Investments?," *Quarterly Journal of Economics*, 2015, pp. 1329–1367.

_ , **Thomas Fujiwara, and Amanda Pallais**, "'Acting Wife': Marriage Market Incentives and Labor Market Investments."

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

_ , **Armin Falk, and Jean Tirole**, "Narratives, Imperatives, and Moral Reasoning," Working Paper 24798, National Bureau of Economic Research July 2018.

Calabuig, Vicente, Gonzalo Olcina, and Fabrizio Panebianco, "Culture and team production," *Journal of Economic Behavior & Organization*, May 2018, *149*, 32–45.

Caldwell, John C. and Pat Caldwell, "The Cultural Context of High Fertility in Africa sub-Saharan," *Population and Development Review*, 1987, *13* (3), 409–437.

Cameron, Lisa, Paul Gertler, and Manisha Shah, "The dirty business of open defecation : Lessons from a sanitation intervention," in "Pacific Conference for Development Economics" 2013.

Carpenter, Jeffrey Paul, "Endogenous Social Preferences," *Review of Radical Political Economics*, March 2005, *37* (1), 63–84.

Carvalho, Jean-Paul, "Veiling," *Quarterly Journal of Economics*, 2013, pp. 337–370.

_ , "Coordination and culture," *Economic Theory*, 2017, *64* (3), 449–475.

Case, Anne, Anu Garrib, Alicia Menendez, and Analia Olgiati, "Paying the piper: the high cost of funerals in South Africa," 2008. NBER Working Paper Series No. 14456.

Cason, Timothy N. and Feisal U. Khan, "A laboratory study of voluntary public goods provision with imperfect monitoring and communication," *Journal of Development Economics*, April 1999, *58* (2), 533–552.

Centola, Damon, Robb Willer, and Michael W. Macy, "The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms," *American Journal of Sociology*, January 2005, *110* (4), 1009–1040.

Chan, Kenneth S., Rob Godby, Stuart Mestelman, and R. Andrew Muller, "Spite, Guilt and the Voluntary Provision of Public Goods When Income Is Not Distributed Equally," *The Canadian Journal of Economics / Revue canadienne d'Economique*, 1996, *29*, S605–S609.

_ , **Stuart Mestelman, Rob Moir, and R. Andrew Muller**, "The Voluntary Provision of Public Goods under Varying Income Distributions," *The Canadian Journal of Economics / Revue canadienne d'Economique*, 1996, *29* (1), 54–69.

_ , _ , **Robert Moir, and R. Andrew Muller**, "Heterogeneity and the Voluntary Provision of Public Goods," *Experimental Economics*, August 1999, *2* (1), 5–30.

Chang, Daphne, Roy Chen, and Erin L Krupka, "Rhetoric matters: A social identity explanation for the anomaly of framing," 2017. Working Paper.

Charness, Gary B., David Masclet, and Marie Claire Villeval, "The Dark Side of Competition for Status," *Management Science*, 2014, *60* (1), 38–55.

Cherry, Todd L., Stephan Kroll, and Jason F. Shogren, "The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab," *Journal of Economic Behavior & Organization*, July 2005, *57* (3), 357–365.

Cho, In-Koo and David M. Kreps, "Signaling Games and Stable Equilibria," *The Quarterly Journal of Economics*, 1987, *102* (2), 179–221.

Cialdini, Robert B. and Noah J. Goldstein, "Social influence: compliance and conformity," *Annual Review of Psychology*, January 2004, *55*, 591–621.

_ , **Carl A. Kallgren, and Raymond R. Reno**, "A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior," *Advances in Experimental Social Psychology*, 1991, *24*, 201–234.

Dana, Jason, Daylian M. Cain, and Robyn M. Dawes, "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and Human Decision Processes*, July 2006, *100*, 193–201.

de Oliveira, Angela C.M., Rachel T. A. Croson, and Catherine Eckel, "One bad apple? Heterogeneity and information in public good provision," *Experimental Economics*, March 2015, *18* (1), 116–135.

Dekel, Sagi and Sven Fischer, *Punishment and Reward Institutions with Harmed Minorities** 2015.

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

Ek, Claes, "Prosocial behavior and policy spillovers: A multi-activity approach," *Journal of Economic Behavior & Organization*, May 2018, *149*, 356–371.

Ellingsen, Tore and Magnus Johannesson, "Pride and prejudice: The human side of incentive theory," *American Economic Review*, 2008, (1997), 990–1008.

_ **and** _ , "Conspicuous generosity," *Journal of Public Economics*, October 2011, *95* (9-10), 1131–1143.

Elster, Jon, "Social Norms and Economic Theory," *Journal of Economic Perspectives*, December 1989, *3* (4), 99–117.

Engel, Christoph and Bettina Rockenbach, "We are Not Alone: The Impact of Externalities on Public Good Provision," SSRN Scholarly Paper ID 1463259, Social Science Research Network, Rochester, NY December 2011.

_ **and** _ , "Give Everybody a Voice! The Power of Voting in a Public Goods Experiment with Externalities," SSRN Scholarly Paper ID 2519479, Social Science Research Network, Rochester, NY November 2014.

_ **and Lilia Zhurakhovska**, "Conditional cooperation with negative externalities – An experiment," *Journal of Economic Behavior & Organization*, December 2014, *108*, 252–260.

Erkut, Hande, Daniele Nosenzo, and Martin Sefton, "Identifying social norms using coordination games: Spectators vs. stakeholders," *Economics Letters*, 2015, *130*, 28–31.

Falk, Armin and Andrea Ichino, "Clean evidence on peer effects," *Journal of Labor Economics*, 2006, *24* (1), 39–57.

Fershtman, Chaim, Uri Gneezy, and Moshe Hoffman, "Taboos and Identity: Considering the Unthinkable," *American Economic Journal: Microeconomics*, May 2011, *3* (2), 139–164.

Fiorina, Morris P. and Samuel J. Abrams, "Political Polarization in the American Public," *Annual Review of Political Science*, June 2008, *11* (1), 563–588.

Fischbacher, Urs, Simeon Schudy, and Sabrina Teyssier, "Heterogeneous reactions to heterogeneity in returns from public goods," *Social Choice and Welfare*, June 2014, *43* (1), 195–217.

Fisher, Joseph, R. Mark Isaac, Jeffrey W. Schatzberg, and James M. Walker, "Heterogenous demand for public goods: Behavior in the voluntary contributions mechanism," *Public Choice*, December 1995, *85* (3-4), 249–266.

Franzen, Axel and Sonja Pointner, "Anonymity in the dictator game revisited," *Journal of Economic Behavior & Organization*, January 2012, *81* (1), 74–81.

Fuller, Bruce, Judith D. Singer, and Margaret Keiley, "Why do Daughters Leave School in Southern Africa? Family Economy and Mothers' Commitments," *Social Forces*, December 1995, *74* (2), 657.

Gangadharan, Lata, Nikos Nikiforakis, and Marie Claire Villeval, "Normative conflict and the limits of self-governance in heterogeneous populations," *European Economic Review*, November 2017, *100* (Supplement C), 143–156.

Geanakoplos, John, David Pearce, and Ennio Stacchetti, "Psychological games and sequential rationality," *Games and Economic Behavior*, March 1989, *1* (1), 60–79.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer, "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment," *American Political Science Review*, February 2008, *102* (01), 33–48.

Granovetter, Mark, "Threshold Models of Collective Behavior," *American Journal of Sociology*, 1978, *83* (6), 1420–1443.

Grossman, Zachary, "Self-signaling and social-signaling in giving," *Journal of Economic Behavior & Organization*, 2015, *117*, 26–39.

Güth, Werner, Anastasios Koukoumelis, M. Vittoria Levati, and Matteo Ploner, "Providing revenue-generating projects under a fair mechanism: An experimental analysis," *Journal of Economic Behavior & Organization*, December 2014, *108*, 410–419.

Hackett, Steven, Edella Schlager, and James Walker, "The Role of Communication in Resolving Commons Dilemmas: Experimental Evidence with Heterogeneous Appropriators," *Journal of Environmental Economics and Management*, September 1994, *27* (2), 99–126.

Hamman, John R., George F. Loewenstein, and Roberto A. Weber, "Self-interest through delegation: An additional rationale for the principal-agent relationship," *American Economic Review*, 2010, *100* (4), 1826–1846.

Hoffman, Elizabeth, Kevin A. McCabe, Keith Shachat, and Vernon L. Smith, "Preferences, property rights, and anonymity in bargaining games," *Games and Economic Behavior*, 1994, *7* (3), 346–380.

Issenberg, Sasha, "Gay-Marriage Strategists Plot PsyOps: The Inevitability Campaign," Website, 2013. `http://nymag.com/news/intelligencer/gay-marriage-opponents-2013-2/`.

Kallgren, Carl A., Raymond R. Reno, and Robert B. Cialdini, "A Focus Theory of Normative Conduct: When Norms Do and Do not Affect Behavior," *Personality and Social Psychology Bulletin*, October 2000, *26* (8), 1002–1012.

Kincaid, D. Lawrence, "From Innovation to Social Norm: Bounded Normative Influence," *Journal of Health Communication*, January 2004, *9* (sup1), 37–57.

Koch, Alexander K. and Hans Theo Normann, "Giving in dictator games: Regard for others or regard by others?," *Southern Economic Journal*, 2008, *75* (1), 223–231.

Kremer, Michael, Edward Miguel, and Rebecca Thornton, "Incentives to Learn," *Review of Economics and Statistics*, August 2009, *91* (3), 437–456.

Krishna, Anirudh, Patti Kristjanson, Maren Radeny, and Wilson Nindo, "Escaping Poverty and Becoming Poor in 20 Kenyan Villages," *Journal of Human Development*, July 2004, *5* (2), 211–226.

Kuran, Timur, "The Inevitability of Future Revolutionary Surprises," *American Journal of Sociology*, 1995, *100* (6), 1528–1551.

_ **and William H. Sandholm**, "Cultural integration and its discontents," *Review of Economic Studies*, 2008, *75* (1), 201–228.

Lacetera, Nicola and Mario Macis, "Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme," *Journal of Economic Behavior & Organization*, November 2010, *76* (2), 225–237.

Layman, Geoffrey C., Thomas M. Carsey, and Juliana Menasce Horowitz, "Party Polarization in American Politics: Characteristics, Causes, and Consequences," *Annual Review of Political Science*, June 2006, *9* (1), 83–110.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber, "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, 2012, *4* (1), 136–164.

Lindbeck, Assar, "Incentives and Social Norms in Household Behavior," *The American Economic Review*, 1997, *87* (2), 370–377.

López-Pérez, Raúl, "Aversion to norm-breaking: A model," *Games and Economic Behavior*, September 2008, *64* (1), 237–267.

Macintyre, Kate, Lisanne Brown, and Stephen Sosler, "It's not what you know, but who you knew: examining the relationship between behavior change and AIDS mortality in Africa," *AIDS Education and Prevention*, April 2001, *13* (2), 160–74.

Malmendier, Ulrike, Vera L. te Velde, and Roberto A. Weber, "Rethinking Reciprocity," *Annual Review of Economics*, 2014, *6*, 849–874.

Manski, Charles F and Joram Mayshar, "Private Incentives and Social Interactions: Fertility Puzzles in Israel," *Journal of the European Economic Association*, 2003, *1* (1), 181–211.

Mas, Alexandre and Enrico Moretti, "Peers at work," *American Economic Review*, March 2009, *99* (1), 112–145.

Matthews, Dylan, "In 2011, only 15 senators backed same-sex marriage. Now 49 do.," Website, 2013. `http://www.washingtonpost.com/blogs/wonkblog/wp/2013/04/02/in-2011-only-15-senators-backed-same-sex-marriage-now-49-do/`.

McCarty, Nolan M., Keith T. Poole, Howard Rosenthal, and Janet T. Knoedler, *Polarized America: The dance of ideology and unequal riches*, Cambridge, MA: MIT Press, 2006.

Michaeli, Moti and Daniel Spiro, "Norm conformity across societies," *Journal of Public Economics*, 2015, *132*, 51–65.

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

Neary, Philip R., "Competing conventions," *Games and Economic Behavior*, 2012, *76* (1), 301–328.

Nikiforakis, Nikos, Charles N. Noussair, and Tom S. Wilkening, "Normative conflict and feuds: The limits of self-enforcement," *Journal of Public Economics*, 2012, *96* (9-10), 797–807.

Oberholzer-Gee, Felix and Reiner Eichenberger, "Fairness in extended dictator game experiments," *The BE Journal of Economic Analysis & Policy*, 2008, *8* (1), Article 16.

Parker, Dianne, Antony S. R. Manstead, and Stephen G. Stradling, "Extending the theory of planned behaviour: The role of personal norm," *British Journal of Social Psychology*, 1995, *34* (2), 127–138.

Reuben, Ernesto and Arno Riedl, "Public Goods Provision and Sanctioning in Privileged Groups," *Journal of Conflict Resolution*, November 2008, *53* (1), 72–93.

_ **and** _ , "Enforcement of contribution norms in public good games with heterogeneous populations," *Games and Economic Behavior*, 2013, *77* (1), 122–137.

Satow, Kay L., "Social approval and helping," *Journal of Experimental Social Psychology*, 1975, *11*, 501–509.

Schwartz, Shalom H., "Normative Influences on Altruism," in Leonard Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 10, Academic Press, January 1977, pp. 221–279.

Seabright, Paul B., "Continuous preferences and discontinuous choices: How altruists respond to incentives," *The BE Journal of Theoretical Economics*, 2009, *9* (1), 14.

Sell, Jane and Rick K. Wilson, "Levels of information and contributions to public goods," *Social Forces*, 1991, *70* (1), 107–124.

Silver, Nate, "Gay Marriage Opponents Now in Minority," Website, 2011. http://fivethirtyeight.blogs.nytimes.com/2011/04/20/gay-marriage-opponents-now-in-minority/.

Soetevent, Adriaan R., "Anonymity in giving in a natural context: a field experiment in 30 churches," *Journal of Public Economics*, December 2005, *89* (11-12), 2301–2323.

Svensson, Jakob, "Investment, property rights and political instability: Theory and evidence," *European Economic Review*, July 1998, *42* (7), 1317–1341.

Tavoni, Alessandro, Astrid Dannenberg, Giorgos Kallis, and Andreas Löschel, "Inequality, communication, and the avoidance of disastrous climate change in a public goods game," *Proceedings of the National Academy of Sciences*, July 2011, *108* (29), 11825–11829.

Traxler, Christian, "Social norms and conditional cooperative taxpayers," *European Journal of Political Economy*, 2010, *26* (1), 89–103.

Weng, Qian and Fredrik Carlsson, "Cooperation in teams: The role of identity, punishment, and endowment distribution," *Journal of Public Economics*, June 2015, *126*, 25–38.

White, Katherine M., Joanne R. Smith, Deborah J. Terry, Jaimi H. Greenslade, and Blake M. McKimmie, "Social influence in the theory of planned behaviour: The role of descriptive, injunctive, and in-group norms," *British Journal of Social Psychology*, March 2009, *48* (1), 135–158.

A Proofs

Throughout these proofs, for notational convenience, define $m_i = m(x_i)$ (or $m_{i,a}, m_{i,r}$), $H_i = H(m_i)$ (or $H_{i,r}, H_{i,a}$) and $v_i = v(x_i)$. I will suppress superscripts denoting individual i.

Proof of Proposition 1 part 1: Type t with ρ_1 will choose x_1 *iff* $v_1 + sH_1 > v_2 - tG + sH_2 \Leftrightarrow$

$$t > \frac{v_2 - v_1 + s(H_2 - H_1)}{G} \equiv \tilde{t}_1.$$

Likewise, type t with ρ_2 will choose x_2 *iff*

$$t > \frac{v_1 - v_2 + s(H_1 - H_2)}{G} \equiv \tilde{t}_2 = -\tilde{t}_1.$$

Since one of these cutoff values is positive and one is negative, low t types with one personal norm will defect to the other action, and all types with the other norm will adhere to their personal norm. For approval seekers, $H_1 = H_a(p_1 G)$ and $H_2 = H_a((1-p_1)G)$ are exogenous, so all components $\tilde{t}_1$ and $\tilde{t}_2$ are exogenously fixed, so existence and uniqueness is trivial. The last statement is immediate from Assumption 1.

Proof of Proposition 1 part 2: As in the proof of Proposition 1 part 1, the cutoff values $\tilde{t}_1$ and $\tilde{t}_2$ are opposite sign, so there are two possibilities: either all first types choose x_1 while some low t second types also choose x_1, or vice versa. Now, however, $H_{1,r}$ and $H_{2,r}$ are endogenously determined.

Suppose that the former possibility is the case: all types with ρ_1 adhere to x_1 and low t types with ρ_2 defect. Then it must be that $\tilde{t}_1 < 0$ (ignoring knife-edge cases). But then, $s(H_1 - H_2) > v_2 - v_1$, which requires $H_{1,r} > H_{2,r}$. But this cannot be the case because low t types with ρ_2 are also choosing x_1, which makes the conditional expectation of t on choosing x_1 lower than on choosing x_2.

So we must have $\tilde{t}_1 > 0$, $\tilde{t}_2 < 0$. We must now only show that such an equilibrium exists.

Given the inference function and this cutoff value, we can calculate the image associated with each choice:

$$m_{2,r}(\tilde{t}_1) = \frac{(1-p_1)\bar{t} + p_1 \int_0^{\tilde{t}_1} t\phi(t)dt}{1 - p_1 + p_1\Phi(\tilde{t}_1)}$$

and

$$m_{1,r}(\tilde{t}_1) = \frac{\int_{\tilde{t}_1}^{\infty} t\phi(t)dt}{1 - \Phi(\tilde{t}_1)}.$$

These two equations, along with the one defining $\tilde{t}_1$ above, define the equilibria of the model. This system of equations must be shown to have a solution with $\tilde{t}_1 > 0$.

Define

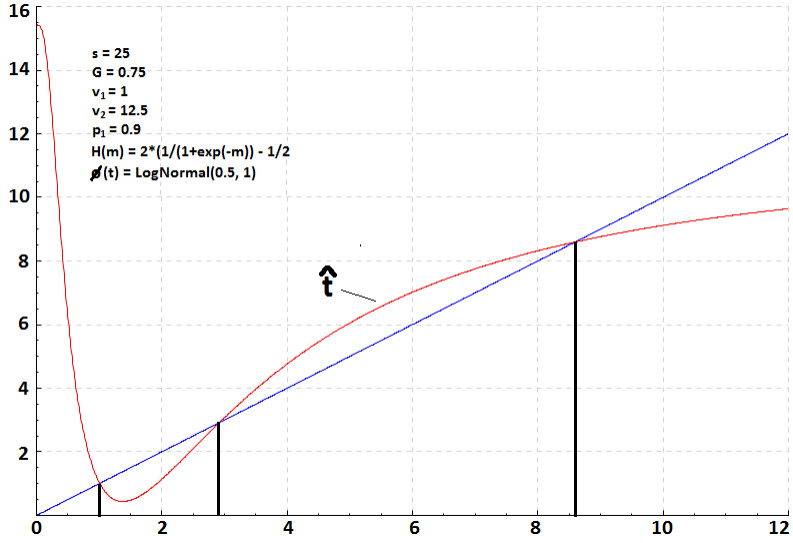
$$\hat{t}(t) = \frac{s(H_r(m_{2,r}(t)) - H_r(m_{1,r}(t))) + v_2 - v_1}{G}.$$

This is a continuous and finite valued function, by Assumption 1. At $t = 0$, $\hat{t} = (s(H_r(\bar{t}) - H_r(\bar{t})) + v_2 - v_1)/G > 0 = t$. As $t \to \infty$, $\hat{t} < t$ necessarily. Therefore by the intermediate value theorem, there is some positive, finite t with $\hat{t}(t) = t$. This provides the desired equilibrium

value of $\tilde{t}_1$ and determines the equilibrium outcome fully.

Figure 4 shows an example graph of t and $\hat{t}$, with three intersections and therefore three possible equilibria.

Figure 4: Example with three equilibria



An example of the model with the specified parameters. The curve shows $\hat{t}$ as defined in the proof of Proposition 1, and any point where it crosses the 45° line marks an equilibrium.

Finite support for $\phi(t)$: As noted in the text, Proposition 1 holds strictly as stated, under the D1 criterion, even if the support of $\phi(t)$ is allowed to be finite. If $\max \operatorname{supp} \phi = T < \infty$, there is a discontinuous drop in m_1 from T to 0 when $\tilde{t}_1$ increases just past T, so the above equilibrium no longer applies. It's now possible that no equilibria exists in which both actions are chosen.

Note that the inference function as described does not apply to actions that are never taken in equilibrium. But, we can use the D1 criterion of Cho and Kreps (1987) to explore these non-separating equilibria.

Under the D1 criterion, in order to rule out type (t, ρ) from the inference function after a disequilibrium choice x is observed, it must be the case that for any mistaken belief about inferences off the equilibrium path that might induce (t, ρ) to deviate to x (that is, with indifference or strict preference), there is another type (t', ρ') (the same type for any potential mistaken belief) who would strictly prefer to deviate with that same mistaken belief.

First suppose no one chooses x_1 in equilibrium. Type (t, ρ_2) might deviate to x_1 under

mistaken beliefs $\tilde{m}_1$, resulting in mistaken beliefs about image utility $\tilde{H}_1$, if $s\tilde{H}_1 \geq v_2 - v_1 + sH_2 + tG$. Type (t, ρ_1) might deviate if $s\tilde{H}_1 > v_2 - v_1 + sH_2 - tG$. Clearly, if any type is willing to deviate for a given $\tilde{H}_1$, then the type (T, ρ_1) strictly prefers to deviate. This is therefore the only type that can be inferred after observing x_1, and $m(x_1)$ is required to be T.

On the other hand, $m(x_2) = \bar{t} < T = m(x_1)$. If $v_2 - v_1$ is large enough to overcome the image benefit of defecting, then, this pooling equilibrium is sustainable. This occurs when $v_2 + sH(\bar{t}) - TG \geq v_1 + sH(T) \iff T \leq \frac{v_2 - v_1 + s(H(\bar{t}) - H(T))}{G}$. But this is exactly the opposite of the condition that guaranteed a separating equilibrium above. Therefore, if no separating equilibrium exists, there is a pooling equilibrium (and vice versa, guaranteeing existence of *some* equilibrium) in which all types choose x_2, in accordance with the statement of the proposition.

Note that if no one chooses x_2 in equilibrium, a similar argument shows that $m(x_2) = T$. But now $m(x_1) = \bar{t} < m(x_2)$, and type (T, x_2) would strictly prefer to deviate, so no pooling equilibrium on x_1 exists. This shows that pooling equilibria also satisfy the conditions of the theorem.

Proof of Proposition 2: Part 1 follows from the proof of Proposition 1 part 1: As $s \to \infty$, one of the cutoff values (corresponding to the norm with the lower image) will approach infinity as well, so that all types with either norm will choose the other action. The relative social image utility is immediate from Assumption 1.

As for part 2, at low levels of s, the relative cost of actions determines the relative numbers that choose those actions and the relative image consequences of them. If $s = 0$ exactly, the signaling game disappears and people simply choose action one unless their guilt from not choosing action two outweighs the cost. As $s \to \infty$, on the other hand, image motivations dominate all other concerns, so any difference between $H_{2,r}$ and $H_{1,r}$ is not sustainable in equilibrium. By the proof of Proposition 1 part 2, the only way for them to be equal is for $\tilde{t}_1 = \tilde{t}_2 = 0$ (recall that pooling equilibria cannot exist, even as s becomes arbitrarily large, because for any fixed finite value of s there are types with sufficiently high t to choose according to their personal norms.)

Proof of Proposition 3: This follows directly from Assumption 1 and Propositions 1 and 2.

Proof of Proposition 4 part 1: Similarly to Proposition 1 part 2, a system of three

equations for $m_{1,r}(\tilde{t}_1)$, $m_{2,r}(\tilde{t}_1)$, and $\tilde{t}_1$ define the equilibria. And as before, if either cutoff value $\tilde{t}_i$ is positive, *all* types with the other norm follow their personal norm. The system of equations is:

$$m_{1,r}(\tilde{t}_1) = \frac{p_1 \int_{\max(0,\tilde{t}_1)}^{\infty} t\phi_1(t)dt + (1-p_1)\int_0^{\max(0,-\tilde{t}_1)} t\phi_2(t)dt}{p_1(1-\Phi_1(\tilde{t}_1)) + (1-p_1)\Phi_2(-\tilde{t}_1)}$$

$$m_{2,r}(\tilde{t}_1) = \frac{p_1 \int_0^{\max(0,\tilde{t}_1)} t\phi_1(t)dt + (1-p_1)\int_{\max(0,-\tilde{t}_1)}^{\infty} t\phi_2(t)dt}{p_1\Phi_1(\tilde{t}_1) + (1-p_1)(1-\Phi_2(-\tilde{t}_1))}$$

$$\tilde{t}_1 = \frac{s(H_{2,r} - H_{1,r}) + v_2 - v_1}{G}$$

The argument for existence of equilibrium follows similarly to Proposition 1 part 2, but is more directly implied by Brouwer's fixed point theorem. $\hat{t}(t)$, as defined above, is finite valued (bounded due to the upper bound on H) and continuous, so it maps a convex, compact subset of $\mathbb{R}^3$ to itself. Therefore $\hat{t} = t$ has a solution, which provides the equilibrium value of $\tilde{t}_1$. But, unlike before, we can't rule out either sign of $\tilde{t}_1$, so imitation in either direction can occur.

Proof of Proposition 4 part 2 and 3: As before, a difference in the image outcome of each choice isn't sustainable in equilibrium as s becomes sufficiently large. Given the operation of the cutoff values $\tilde{t}_1$ and $\tilde{t}_2$, clearly the only way for the image outcome to be the same is for low t types with the less "prestigious" ρ norm (i.e. the norm of the sub-population with the lower average t) to seek a higher status by choosing against their norm.

Part 3 simply points out again what part 1 says when correlation implies this relationship: costliness doesn't prevent imitation, leading to "too much" sacrifice overall.

Proof of Proposition 5 part 2: Define $\tilde{t}_{i,j,k}$ to be the cutoff type t above which someone with personal norm x_i will prefer x_j to x_k. In particular,

$$\tilde{t}_{1,1,2} = \frac{s(H_2 - H_1) + v_2 - v_1}{G_1},$$

$$\tilde{t}_{1,2,3} = \frac{s(H_3 - H_2) + v_3 - v_2}{G_2 - G_1},$$

$$\tilde{t}_{1,1,3} = \frac{s(H_3 - H_1) + v_3 - v_1}{G_2},$$

$$\tilde{t}_{3,3,2} = \frac{s(H_2 - H_3) + v_2 - v_3}{G_1} = -\tilde{t}_{1,2,3}\frac{G_2 - G_1}{G_1},$$
$$\tilde{t}_{3,2,1} = \frac{s(H_1 - H_2) + v_1 - v_2}{G_2 - G1} = -\tilde{t}_{1,1,2}\frac{G_1}{G_2 - G_1},$$

and

$$\tilde{t}_{3,3,1} = \frac{s(H_1 - H_3) + v_1 - v_3}{G_2} = -\tilde{t}_{1,1,3}.$$

Note that $\tilde{t}_{1,1,2}$ and $\tilde{t}_{3,2,1}$, $\tilde{t}_{1,1,3}$ and $\tilde{t}_{3,3,1}$, and $\tilde{t}_{1,2,3}$ and $\tilde{t}_{3,3,2}$, are respectively opposite sign, and that they are pairwise determined. These relationships, along with a requirement of transitivity for all types, restricts the possible relationships between the six cutoff values to one of 5 behaviorally distinct types of equilibria (the reader can check that any relationship not included in this list isn't feasible):

Type 1: $\tilde{t}_{1,1,2} > \tilde{t}_{1,1,3} > \tilde{t}_{1,2,3} > 0$ (while $\tilde{t}_{3,2,1}, \tilde{t}_{3,3,2}, \tilde{t}_{3,3,1} < 0$ necessarily). Types with ρ_1 differentiate between all three options: types with $t > \tilde{t}_{1,1,2}$ choose x_1, with $\tilde{t}_{1,2,3} < t < \tilde{t}_{1,1,2}$ choose x_2, and with $t < \tilde{t}_{1,2,3}$ choose x_3. All types with ρ_3 choose x_3.

Type 2: $\tilde{t}_{1,2,3} > \tilde{t}_{1,1,3} > 0, \tilde{t}_{1,1,2}$ ($\tilde{t}_{1,1,2}$ may have either sign). In this type, types with ρ_1 choose x_1 if $t > \tilde{t}_{1,1,3}$ and x_3 otherwise. All types with ρ_3 choose x_3.

Type 3: $\tilde{t}_{1,1,2} > 0, \tilde{t}_{1,1,3} > \tilde{t}_{1,2,3}$. In this type, types with ρ_1 choose x_1 if $t > \tilde{t}_{1,1,2}$ and choose x_2 otherwise, and types with ρ_3 choose x_3 if $t > \tilde{t}_{3,3,2} > 0$ and x_2 otherwise.

Type 4: $\tilde{t}_{3,2,1} > \tilde{t}_{3,3,1} > 0, \tilde{t}_{3,3,2}$. In this type, types with ρ_3 choose x_3 if $t > \tilde{t}_{3,3,1}$ and x_1 otherwise. All types with ρ_1 choose x_1.

Type 5: $\tilde{t}_{3,3,2} > \tilde{t}_{3,3,1} > \tilde{t}_{3,2,1} > 0$. Types with ρ_3 differentiate between all three options: types with $t > \tilde{t}_{3,3,2}$ choose x_3, with $\tilde{t}_{3,2,1} < t < \tilde{t}_{3,3,2}$ choose x_2, and with $t < \tilde{t}_{3,2,1}$ choose x_1. All types with ρ_1 choose x_1.

Additionally, the assumption that $v_3 > v_1$ eliminates the last two possibilities. In these equilibria, by definition of the image function, $H_1 < H_3$, so since $v_3 > v_1$ as well, $\tilde{t}_{3,3,1} = \frac{s(H_1 - H_3) + v_1 - v_3}{G_2}$ must be negative. But equilibria of type 4 or 5 require that it be positive.

This establishes the described form of all equilibria. Next, I will show that only type 2 equilibria are permitted in the limit when $s \to \infty$.

1. By definition of m_r, in a type 1 equilibrium, $m_1 > m_2, m_3$. A partial requirement for a type 1 equilibrium is that $\tilde{t}_{1,1,3} > \tilde{t}_{1,2,3} > 0 \leftrightarrow \frac{v_3 - v_1 + sH_3 - sH_1}{G_2} > \frac{v_3 - v_2 + sH_3 - sH_2}{G_2 - G_1} > 0$.

Therefore, $\tilde{t}_{1,1,3} > 0$ requires, as $s \to \infty$, that $m_1 \to m_3$ and $\tilde{t}_{1,1,3}$ remains finite. This occurs only when $\tilde{t}_{1,1,3} \to 0$, which implies that $m_2 = 0$, which implies that $\tilde{t}_{1,2,3}$ grows infinite. This contradicts the stated relationship, so no equilibrium of type 1 exists when $s \to \infty$.

2. Type 2 requires, in part, that $\tilde{t}_{1,2,3} > \tilde{t}_{1,1,3} > 0 \leftrightarrow \frac{v_3 - v_2 + sH_3 - sH_2}{G_2 - G_1} > \frac{v_3 - v_1 + sH_3 - sH_1}{G_2} > 0$. By definition of m_r, $m_1 > m_3$, and m_2 is undefined as x_2 is never chosen on the equilibrium path. We must resort to the D1 criterion to evaluate m_2.

   We must consider three types of deviations to x_2: A person with ρ_1 and $t < \tilde{t}_{1,1,3}$ would normally choose x_3, but would prefer x_2 if $sH_2 > v_3 - v_2 + sH_3 - t(G_2 - G_1)$. Since $G_2 > G_1$, then if type $t \in [0, \tilde{t}_{1,1,3})$ is tempted to deviate for some mistaken belief $\hat{H}_2$, then type $t = \tilde{t}_{1,1,3}$ is also tempted to deviate for the same mistaken belief. The D1 criterion therefore says that no weight can be placed on $t \in [0, \tilde{t}_{1,1,3})$ (along with an inferred ρ_1) when inferring a type after observing x_2. By a similar argument, someone with ρ_1 and $t > \tilde{t}_{1,1,3}$ would deviate from their normal choice of x_1 under a mistaken belief satisfying $s\hat{H}_2 > v_1 - v_2 + sH_1 + tG_1$, and similarly no weight can be placed on $t \in (\tilde{t}_{1,1,3}, \infty)$ (along with an inferred ρ_1) when inferring a type from a choice of x_2. Lastly, someone with ρ_3 might wish to deviate for a mistaken belief satisfying $sH_2 > v_3 - v_2 + sH_3 + tG_1$, and no weight may be placed on $t \in (0, \infty)$ (along with an inferred ρ_3) when observing x_2. Altogether, all weight must be placed on $t = 0$ or $t = \tilde{t}_{1,1,3}$, which implies that $m_2 \in [0, \tilde{t}_{1,1,3}]$.

   Referring back to the required relationship above, $\tilde{t}_{1,1,3} > 0$ requires that $H_1 \to H_3$ as $s \to \infty$, which can only occur when $\tilde{t}_{1,1,3} \to 0$. By the D1 criterion, as above, this means that $m_2 \to 0$. Therefore, $\tilde{t}_{1,2,3} \to \infty$, and $\tilde{t}_{1,1,3} \to \frac{v_3 - v_1}{G_2}$, and the relationship is satisfied *iff* $v_3 > v_1$, as we have assumed.

   The final requirement is that $\tilde{t}_{1,1,3} > \tilde{t}_{1,1,2}$, which is also satisfied since $\tilde{t}_{1,1,2} \to -\infty$.

   In sum, there exists an equilibrium of type 2 as $s \to \infty$.

3. Type 3 equilibria require, in part, that $\tilde{t}_{1,1,2} > 0 \leftrightarrow v_2 - v_1 + sH_2 - sH_1 > 0$. And by definition of m_r, $m_1, m_3 > m_2$. This inequality requires both that $v_2 > v_1$ and $H_1 \to H_2$. But by definition of m_r, this can only occur if $\tilde{t}_{3,3,2} \to \infty$. But this can't be true, since $\tilde{t}_{3,3,2} = \frac{v_2 - v_3 + sH_2 - sH_3}{G_1} \to -\infty$ when $H_2 = H_1 = H(\bar{t})$ and $H_3 \to \infty$. So no type 3 equilibrium exists when $s \to \infty$.

It remains to be shown that some equilibrium of one of these three types always exists. I

will again appeal to Brouwer's fixed point theorem, but a continuous function on a compact, convex space that defines equilibrium at its fixed points must be carefully constructed. In the following, the three parameters of interest are $t_{1,1,2}$, $t_{1,1,3}$ and $t_{1,2,3}$, but I will refer to $t_{3,j,k}$ where convenient rather than the equivalent values written in terms of $t_{1,j,k}$.

A type 1 equilibrium is defined by the relationship $\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})$ along with the following six equations that must be satisfied:

$$\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{s(H(m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})) - H(m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}))) + v_2 - v_1}{G_1}$$

$$\hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{s(H(m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})) - H(m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}))) + v_3 - v_1}{G_2}$$

$$\hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{s(H(m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})) - H(m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}))) + v_3 - v_2}{G_2 - G_1}$$

$$m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{1,1,2}}^{\infty} t\phi(t)dt}{1 - \Phi(t_{1,1,2})}$$

$$m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{1,2,3}}^{t_{1,1,2}} t\phi(t)dt}{\Phi(t_{1,1,2}) - \Phi(t_{1,2,3})}$$

and

$$m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{(1-p_1)\bar{t} + p_1 \int_0^{t_{1,2,3}} t\phi(t)dt}{1 - p_1 + p_1\Phi(t_{1,2,3})}.$$

And in a type two equilibrium, the first three equations remain the same, but we must have the relationship $\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})$ and the image functions

$$m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{1,1,3}}^{\infty} t\phi(t)dt}{1 - \Phi(t_{1,1,3})}$$

$$m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = t_{1,1,3}$$

and

$$m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{(1-p_1)\bar{t} + p_1 \int_0^{t_{1,1,3}} t\phi(t)dt}{1 - p_1 + p_1\Phi(t_{1,1,3})}.$$

where m_2 results from restricting attention to a subset of equilibria that satisfy the D1 criterion. As shown above, m_2 must fall in the interval $[0, t_{1,1,3}]$, and imposing $m_2 = t_{1,1,3}$ ensures continuity in $t_{1,1,2}$, $t_{1,2,3}$, and $t_{1,1,3}$.

In a type three equilibrium, the expressions for $\hat{t}_{i,j,k}$ remain the same but we must satisfy $\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})$ and the following image functions:

$$m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{1,1,2}}^{\infty} t\phi(t)dt}{1 - \Phi(t_{1,1,2})}$$

$$m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{(1-p_1)\int_0^{t_{3,3,2}} + p_1 \int_0^{t_{1,1,2}} t\phi(t)dt}{(1-p_1)\Phi(t_{3,3,2}) + p_1\Phi(t_{1,1,2})}$$

and

$$m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{t_{3,3,2}}^{\infty} t\phi(t)dt}{1 - \Phi(t_{3,3,2})}.$$

We can combine the conditions for all three types of equilibria as follows: The equations for $\hat{t}_{i,j,k}$ remain the same, and we must satisfy *either* $\hat{t}_{1,1,2}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})$ *or* $\hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) > \hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,1,2})$. And, we have the following image functions:

$$m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{\int_{\max(t_{1,1,3}, t_{1,1,2})}^{\infty} t\phi(t)dt}{1 - \Phi(\max(t_{1,1,3}, t_{1,1,2}))}$$

$$m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \begin{cases} \frac{p_1 \int_{\max(t_{1,2,3},0)}^{t_{1,1,2}} t\phi(t)dt + (1-p_1)\int_0^{\max(0,t_{3,3,2})} t\phi(t)dt}{p_1(\Phi(t_{1,1,2}) - \Phi(\max(0,t_{1,2,3})) + (1-p_1)\Phi(\max(0,t_{3,3,2})))} & \text{if } t_{1,2,3} < t_{1,1,3} \\ t_{1,1,3} & \text{otherwise} \end{cases}$$

(ensuring continuity again by imposing $m_2 = t_{1,1,3}$ when x_2 is never chosen), and

$$m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{(1-p_1)\int_{\max(0,t_{3,3,2})}^{\infty} t\phi(t)dt + p_1 \int_0^{\max(0,\min(t_{1,2,3},t_{1,1,3}))} t\phi(t)dt}{(1-p_1)(1 - \Phi(\max(0, t_{3,3,2}))) + p_1\Phi(\max(0, \min(t_{1,2,3}, t_{1,1,3})))}.$$

Next define some convenient notation:

$$\underline{\underline{H}} \equiv \min_c \frac{(1-p_1)\bar{t} + p_1 \int_0^c t\phi(t)dt}{1 - p_1 + p_1\Phi(c)}.$$

Then, we can establish that each $\hat{t}_{1,j,k}$ must fall within a finite interval, using the maximum and minimum values of the image functions above. In particular,

$$\hat{t}_{1,1,2} \in \left[\frac{-s\overline{H} + v_2 - v_1}{G_1}, \frac{s(\overline{H} - H(\bar{t})) + v_2 - v_1}{G_1}\right],$$

$$\hat{t}_{1,1,3} \in \left[\frac{s(\overline{H} - H(\bar{t})) + v_3 - v_1}{G_2}, \frac{s(\underline{\underline{H}} - \overline{H}) + v_3 - v_1}{G_2} \right],$$

and

$$\hat{t}_{1,2,3} \in \left[\frac{s\overline{H} + v_3 - v_2}{G_2 - G_1}, \frac{s(\underline{\underline{H}} - \overline{H}) + v_3 - v_2}{G_2 - G_1} \right].$$

Since each of these intervals is finite, the range of $\hat{T} = (\hat{t}_{1,1,2}, \hat{t}_{1,1,3}, \hat{t}_{1,2,3})$ is a compact, convex subset of $\mathbb{R}^3$. Call this set D. Since $\hat{T}$ is also defined to be continuous, by Brouwer's fixed point theorem, we know that $\hat{T}$ has a fixed point within D.

This fixed point will satisfy the six equations necessary for either a type 1, type 2, or type 3 equilibrium, but is not guaranteed to satisfy the inequalities relating $t_{1,1,2}$, $t_{1,1,3}$, and $t_{1,2,3}$ which guarantee that these three parameters describe a state in which preferences are transitive. Restricting attention to the subset of D corresponding to feasible preferences prevents us from appealing to Brouwer's fixed point theorem, as this subset is not convex; for example, while $(\tilde{t}_{1,1,2}, \tilde{t}_{1,1,3}, \tilde{t}_{1,2,3}) = (5, 4, 1)$ falls in the category of type 1 equilibria, and $(-1, 4, 5)$ falls in case 2, the midpoint between these values, $(2, 4, 3)$, leads to intransitive preferences.

However, we can show that the image of *any* point in D under $\hat{T}$ leads to transitive preferences. Transitive preferences arise when either $t_{1,1,2} > t_{1,1,3} > t_{1,2,3}$, or when $t_{1,2,3} > t_{1,1,3} > t_{1,1,2}$. But by construction,

$$t_{1,1,2} > t_{1,1,3} \Rightarrow t_{1,1,3} > t_{1,2,3}$$

and

$$t_{1,2,3} > t_{1,1,3} \Rightarrow t_{1,1,3} > t_{1,1,2}$$

. In the former case, this can be seen by ranking choices for someone with $t = t_{1,1,3}$, and similarly in the latter case. That is, no matter what relation two components of $\hat{T}$ take towards each other, the third is guaranteed to fall in the range required for rational preferences. In other words, while D is a convex, compact subset of $\mathbb{R}^3$, $\hat{T}(D) \subset D$ is the nonconvex subset containing only points that lead to rational preferences. Therefore, whatever the fixed point of $\hat{T}$ is on D, it describes a valid equilibrium of one of the three types described above. This completes the proof.

Proof of Proposition 5 part 1: Any equilibrium must be one of the five forms described in the first part of the proof of Proposition 5 part 2, as that argument does not depend on

the definition of H_r compared to H_a. The unique equilibrium trivially exists as the response of each type to fixed, exogenous factors in their optimization problem.

Proof of Proposition 6: Part 1 is a secondary conclusion of the proof of Proposition 5 part 2.

For part 2, note that the social image of each action is fixed: $m(x_1) = -(1-p_1)G_2$, $m(x_2) = -G_1$, and $m(x_3) = -p_1G_2$. Any of these quantities may be smallest (i.e. most negative), and as s increases, $H(m(x))$ becomes the overwhelming factor in each person's decision. Therefore, in the limit, everyone pools on the action with the least negative image. Note that this is a substantive difference from lower levels of social pressure since, as in Proposition 5, all five types of equilibria exist at low s.

Proof of Proposition 7: Part 1 is true by Propositions 1 and 2, since the quantity $v(x_i)$ in those results is replaced in this setting with $v(x_i) + s_a H_a(m_a(x_i))$. At small s_a, the first term dominates, and at higher s_a, the latter dominates.

Similarly to part 1, parts 2 and 3 follows from Propositions 1 and 2.