Heterogeneous norms: Social image and social pressure when people disagree

Vera L. te Velde*

July 16, 2021

Abstract

Social pressure has been successfully used to encourage prosocial behavior in a diverse range of settings. Some backfiring results, however, have prompted a closer look at the necessary conditions for success. I propose that disagreement about what the right thing is to do is one key underexplored factor that can explain these findings. Existing models of social image break down when personal norms are heterogeneous because it's unclear which choice provides the best image. Some models have addressed this by assuming individuals seek the approval of a relevant reference network, but this approval seeking is qualitatively different from the signaling role of normative behavior that has been shown to be very important in homogeneous norm settings. I distinguish "respect" as the type of social image attained when one's actions are inferred to be motivated by strong personal beliefs, from "approval", which is obtained when one's actions are judged to be normatively correct. Using a psychological game theoretic model, I show how these distinct motives lead to different outcomes in terms of consensus, hypocrisy, compromise, polarization, and destructive posturing. Results demonstrate how using social incentives to change behavior may easily backfire if heterogeneous norms, or approval and respect, are conflated.

JEL classification: D03; Z13; C72; D71.

Keywords: Norms, Social Pressure, Social Image, Signaling

1 Introduction

Social image and social pressure have emerged as key determinants of prosocial behavior in the economics literature. Among many other applications, they have been exploited

*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. Declarations of interest: none.

to improve voter turnout (Gerber, Green and Larimer, 2008), increase charitable giving (DellaVigna, List and Malmendier, 2012), promote blood donation (Lacetera and Macis, 2010), and spread safe sex practices (Ashraf, Bandiera and Jack, 2014).[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, internalized belief about what the right thing to do is in a given situation.[3] When personal norms are shared, social image simply further motivates adherence to personal norms by reinforcing personal moral motivation. 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 eat meat in order to fit in with my peers. Or, I might want to be vegetarian, because when my peers see that I am incurring a personal cost to do so, they will infer that I have strong beliefs and am acting non-hypocritically. If the right thing to do is unanimously recognized, these two motivations will always push in the same general direction. But when personal norms are heterogeneous, we can separately define two types of potentially conflicting social image motivations:

"Approval" 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. 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 the friend would obtain some utility from that approval.

[2]See Bursztyn and Jensen (2017) for a review of the evidence of the importance of social image in other field settings.

[3]This terminology is standard in the social psychology norms literature (Schwartz, 1977; Parker, Manstead and Stradling, 1995; Kallgren, Reno and Cialdini, 2000; Bicchieri, 2006, 2010; Cialdini and Goldstein, 2004; White et al., 2009, e.g.) and has been adopted by some economists (Erkut, Nosenzo and Sefton, 2015; Calabuig, Olcina and Panebianco, 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).

"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 personal norms are 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 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, personal norms, and social image derived from either approval or respect. The comparative statics analysis focuses on the role of social pressure, and demonstrates that efforts to influence behavior using 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 make an exception to 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 seekers 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. If social pressure is strong enough to dominate 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 tend towards conformity, while respect seekers become entrenched along moral dividing lines.

I also explore several extensions to this basic analysis. The polarizing force of respect-seeking is even more evident when middle ground options are available. Approval seekers may be willing to follow Mark Bittman's recommendation to be "vegan before dinnertime" to appease both sides, but respect seekers avoid these options as they do not credibly signal adherance to any norm. On the other hand, approval seekers are prone to entrenching

costly majority norms such as female circumcision or overspending on status goods, so that *reducing* social pressure may be beneficial. If people care about *both* approval and respect, then disapproved actions such as ecoterrorism may act as effective signals of integrity.

The model reveals how designing social incentives under a naive understanding of norm heterogeneity and image concerns may 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 pressure to change behavior when norms are seemingly universal. But this increasingly popular tactic raises a red flag when understood within this model of heterogeneous norms. While experimental work that can pinpoint the source of counterintuitive effects of social pressure nudges is extremely scarce, this model may explain some of those findings and can be used to guide the design and testing of future nudges. This is discussed further in Section 6.

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 reviewed by Bursztyn and Jensen (2017), experiments have isolated the role of social pressure by manipulating observability (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, e.g.). Additional evidence comes from experiments showing that people often desire to avoid being put in pressuring situations (Dana, Cain and Dawes, 2006; Broberg, Ellingsen and Johannesson, 2007; Lazear, Malmendier and Weber, 2012; Oberholzer-Gee and Eichenberger, 2008; Lin, Schaumberg and Reich, 2016; Malmendier, te Velde and Weber, 2014, e.g.), to delegate morally contentious choices to biased agents (Hamman, Loewenstein and Weber, 2010; Bartling and Fischbacher, 2011), or to hide behind ambiguity about the choices they face (Franzen and Pointner, 2012; Andreoni and Bernheim, 2009; te Velde, 2018; Kritikos and Tan, 2014; Grossman, 2015, e.g.).

The notion of "respect" studied in this paper is inspired by several signaling models of social influence in homogeneous norm contexts. This paper builds on this literature by taking seriously the notion of signaling-based image and demonstrating how that leads to very different outcomes in more complex settings with heterogeneous norms. Andreoni and Bernheim (2009), Grossman (2015), and Ellingsen and Johannesson (2011) all model social

image utility as the inference observers have about the decision maker's personal weight on doing the "right" thing. A similar approach has been used to study crowding out by material incentives, expressive powers of laws, and other phenomena (Bénabou and Tirole, 2006; Ellingsen and Johannesson, 2008; Seabright, 2009; Bénabou and Tirole, 2011*b*; Bénabou, Falk and Tirole, 2018; Bénabou and Tirole, 2011*a*).

The notion of "approval" studied in this paper is a simple non-signaling[4] model of social influence designed to capture the basic traits of several similar models of conformity, to contrast with the respect motive. For example, Akerlof (1980) 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) similarly models social image as average community member's evaluation of the individual's choice, according to their own preferences/beliefs.[5]

Akerlof (2017) independently provides a model of the *formation* of heterogeneous norms in a population. He also focuses on the effect of social pressure (via "encouraging interaction"), and can thus be seen as a complementary contribution speaking to similar issues (see next paragraph). Alternatively, using a less literal analogy, if we interpret the choice of values in Akerlof's model as the actions chosen in this model, and the innate abilities in Akerlof's model as the innate personal norms of this model, then it becomes evident that his approach is more similar to the model of approval in the sense that social pressure encourages conformity in his results. The literature on intergenerational transmission of preferences

[4]A few papers model something like approval in a signaling context: Bernheim (1994) models image as the inferred closeness of someone's personal norm to the social norm, and Bursztyn, Egorov and Fiorin (2020) as the observer's inferred belief that he shares a personal norm with the decision maker. These definitions are psychologically distinct from the concept of approval modeled here, and are perhaps more relevant to the literature on identity, in which group members obtain utility from adhering to a shared, prescribed group behavior (Akerlof and Kranton, 2000, 2002, 2005; Benjamin, Choi and Strickland, 2010; Chang, Chen and Krupka, 2019; Akerlof, 2017).

[5]Several models assume that social image is a function of others' actions rather than their beliefs. Fershtman, Gneezy and Hoffman (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, Dechter and Foster (2017); Centola, Willer and Macy (2005); Granovetter (1978); Lindbeck (1997); Manski and Mayshar (2003), and several models of personal norm formation assume beliefs themselves track group behavior (Calabuig, Olcina and Panebianco, 2018; Kincaid, 2004; Kuran and Sandholm, 2008). Modeling injunctive judgement rather than descriptive norms is a simplifying choice that is not ultimately relevant to the focus of this paper.

(Bisin and Verdier, 2001; Tabellini, 2008; Adriani and Sonderegger, 2009; Adriani, Matheson and Sonderegger, 2018, e.g.) is similarly complementary for providing a framework for understanding how heterogeneous norms may arise in the first place.

Finally, it must be noted that social pressure is only one type of social incentive that can form the basis of prosocial nudges. A large literature (see Miller and Prentice (2016)) emphasizes the importance of shared notions of appropriate behavior (injunctive norms), the degree of adherance to these norms (descriptive norms), and the conditional desire to adhere to injunctive norms if enough other people also do so. Providing information about injunctive and descriptive norms can therefore change behavior by informing people what needs to be done to gain approval from their peers, or what they can likely get away with, or by preventing self-serving interpretations of the situation. These types of nudges must also be implemented with care to prevent unintended side effects; see Bicchieri and Dimant (2019) for review and discussion. The contribution of this paper parallels this literature by identifying possible sources of fragility of nudges based purely on social pressure.

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), how much to shirk at work, 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.[6]

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, which I will continue to rely on for the remainder of the model description and results for expositional clarity. Imagine that the relevant choice is meat consumption. The decisionmaker believes that ρ represents the morally sound option but does not want to give up so many of his favorite foods, so he might facetiously argue for the naturalness of omnivorism. Others disagree with his moral beliefs, and some choose different dietary restrictions according to their different moral concerns, or choose to reduce their meat consumption by a certain percentage to balance their environmentalism with concerns about nutrition. Others choose transparently selfish options such as eating everything except goat, simply because they don't like goat.

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 with finite mean 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

[6]Image associated with ρ is not modeled as a signaling game to capture the intuition that I appreciate someone who votes the same way as I do even if I know they are doing it only to stay my friend; independently, I may lose respect for that person since their t is low enough to conform to my preferences.

of G is clearly violated in many instances (over-generosity surely induces less guilt than under-generosity) but 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 focus on the contrast between respect and approval, and so will contrast the predictions of the model of respect on its own to the model of approval. To distinguish them notationally, approval seeking decision makers have image function $m = m_a$ and corresponding image utility function $H = H_a$, while respect seeking decision makers have $m = m_r$ 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. The vegetarian in this scenario would 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.[7] For example, if half of the population believes in ρ_1 and half the population

[7]An alternative approach would be for approval to be based on the fraction of the population choosing

believes in ρ_2, then $m(x_2) = -\frac{1}{2}G(\rho_2 - \rho_1)$. 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 personal norms are 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.

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, Pearce and Stacchetti, 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.[8]

in the same way. Both approaches have been explored in the literature previously and the modeling choice is not ultimately relevant to the focus of this paper; see Section 2.

[8] This distinction between signaling and non-signaling models of image leads to subtle differences even in homogeneous norm contexts, even though both respect and approval encourage the same behavior in those contexts. The signaling equilbrium of 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) 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). In the heterogeneous norm contexts considered here, much more dramatic differences between approval and

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 (vegetarianism or omnivorism) and x must take one of these two same 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 (representing vegetarianism and omnivorism, respectively), and one of two personal norms, $\rho_1 = x_1$ or $\rho_2 = x_2$. Each option provides consumption utility $v(\cdot)$; WLOG assume $v(x_2) > v(x_1)$ - vegetarianism is costly (but see the final paragraph of Section 5.1 for discussion of how individual differences in consumption utility can be accommodated). 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 $\rho = \rho_1$.[9]

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

respect emerge, with social pressure potentially leading to opposing shifts in behavior.

[9] All results hold as stated if p_1 instead represents a common belief about the fraction of the population holding ρ_1.

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.[10]

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:*
   - *In the unique equilibrium, 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 $p_1 > .5$.
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: if vegetarians were less respected than omnivores, then people who believe in omnivorism certainly have no reason to be vegetarian. So, imitation can only occur as stated in the proposition.

[10] 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.

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 more dramatically when social pressure is relatively high. This may be apparent 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.[11] 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.).

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 nudges based on changing social pressure (through ob-

[11] This is not to say that a discontinuous "tipping point" exists, but merely that a particular size shift in beliefs can be amplified by the resulting change in social pressure to cause a larger shift in behavior. More concretely, suppose the costs of the two options are equal (thereby automatically making social pressure high relative to material considerations) and, as assumed in the proposition, $G(x_1 - \rho_2) = G(x_2 - \rho_1)$ so that social pressure changes direction at exactly $p_1 = 50\%$. Then when 49% of the population believes in ρ_1, all 51% who believe in ρ_2 will choose x_2, plus a few more with low t who are swayed by social pressure; conversely if 49% believe in ρ_2. For any distribution of t, then, behavior change will be more significant than belief change, and to a larger degree the higher is s relative to the distribution of t.

servability or anonymity, symbolic rewards, publicity of good behavior, etc) operate when norms are heterogeneous. To understand changes in behavior at aggregate levels, we now compare *populations* of approval seekers to populations of respect seekers, each consisting of individuals simultaneously choosing actions and conferring respect or approval on their peers. 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 a population of approval seekers, as $s \to \infty$, the fraction of the population choosing x_1 (x_2) monotonically approaches 100% if $p_1 > .5$ ($p_1 < .5$).*

2. *For a population of 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 vegetarianism are the ones who feel guilty for not having done so already. The ones who feel most guilty and are quickest to switch are the ones who have t just barely lower than any other vegetarian, and so as they switch, the signaling value of vegetarianism is slightly diluted. As social pressure continues rising and people with even lower t's 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.[12]

Put another way, as social pressure increases, cost disparities between actions become irrelevant for respect seekers, and somewhat counterintuitively, either action will lead to

[12]For respect seekers, the population shift in behavior as social pressure rises may not be perfectly monotonic, hence the emphasis on limit cases in the statement of the result. This is because $E[t|t > A] - E[t|t < A]$ is generally increasing in A for distributions with support $\mathbb{R}^+$ and finite mean, but perhaps not monotonically. That is, as rising social pressure causes people with lower and lower $t's$ to switch to x_1, there may be a point where an influx of low-t types destroys a lot of the signaling advantage of choosing x_1 so that much higher social pressure would be needed to make that slight advantage worthwhile to pursue again. My suspicion is that such extreme situations are rare empirically, but of course the possibility should be borne in mind.

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.

An important corollary of this result is that in certain cirumstances, social pressure will cause respect and approval seekers to shift their behavior in *opposite* directions. In particular, if the majority norm is the cheaper option materially, then increasing pressure will cause conformity to the majority for approval seekers, but will cause a contrarian shift towards the more expensive option for respect seekers.

Corollary 1. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, $\rho \in \{\rho_1 = x_1, \rho_2 = x_2\}$ is independent from t, and $p < 1/2$, the following hold:*

1. *For a population of approval seekers, as $s \to \infty$, the fraction of the population choosing* x_2 increases.

2. *For a population of respect seekers, as $s \to \infty$, the fraction of the population choosing* x_2 decreases.

Figure 1 illustrates these results.

Approval seekers, on the other hand, don't try to discern each others' hypocrisy. If most people believe in omnivorism, 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 animal-rights activists dare to follow their own moral compass in the face of harsh criticism.

In other contexts, 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 possible 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

Figure 1: Approval seekers' versus respect seekers' choices

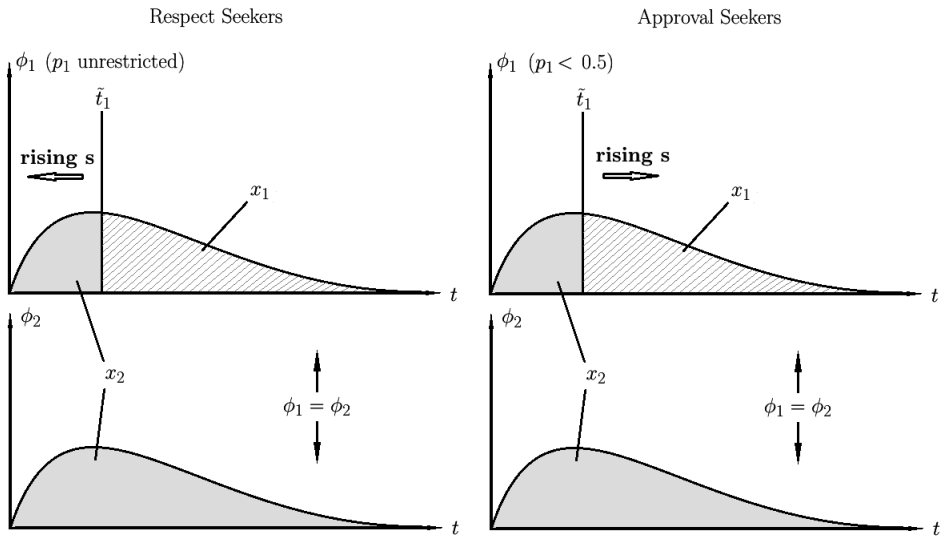
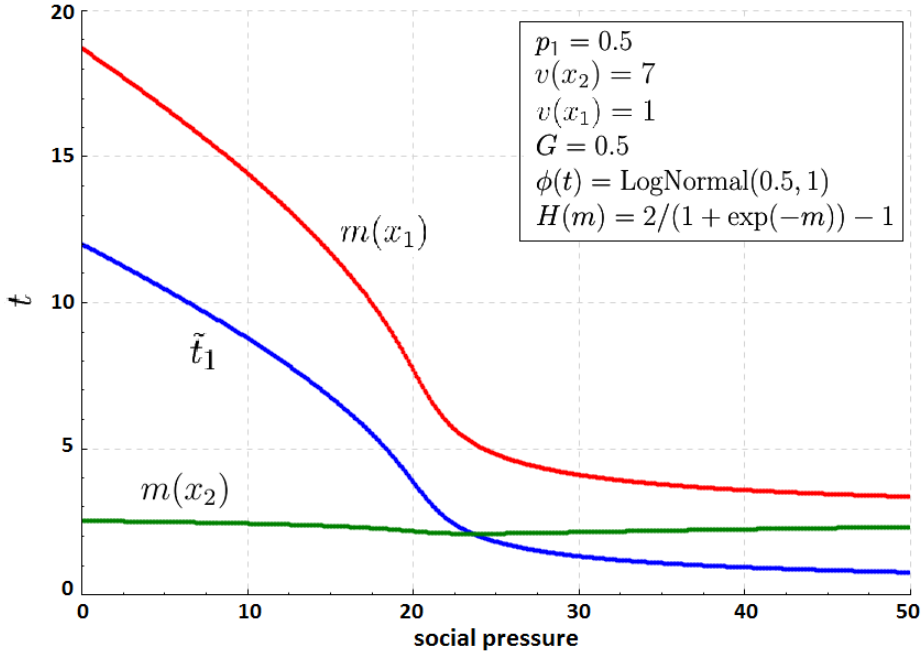


Illustration of choices made by the distributions (of t) of approval seekers (right) and respect seekers (left) with personal norm x_1 (top) and x_2 (bottom). Types in the shaded regions of the distributions choose x_1, and types in the filled regions choose x_2. Increasing social pressure causes more 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.

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 tends 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. 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? 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 is not unambiguously welfare-destroying because the individual gains utility from being more moral. 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, Willer and Macy, 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, entire populations are vegetarian despite the fact that a significant fraction believe there is nothing wrong with eating meat.

5 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[13]).

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:*

[13]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 restrict 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 relaxed 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. Continuing with our dietary example, x_1 now continues to represent vegetarianism, x_3 represents omnivorism, and x_2 represents a compromise such as "vegetarian on Tuesdays".

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:*

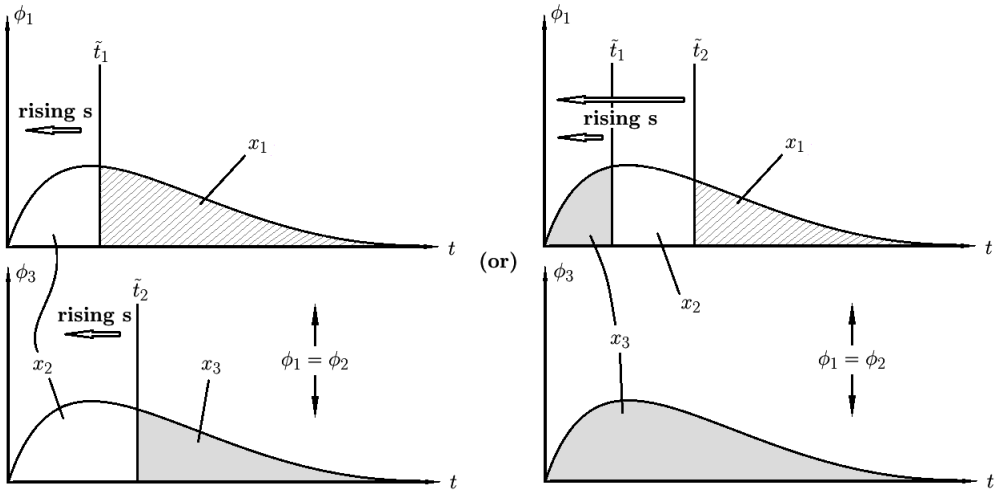
1. *Among approval seekers, individuals with sufficiently high t adhere to their personal norms and lower types either 1) all choose x_1, 2) all choose x_2, 3) all choose x_3, or 4)*

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 material 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. In terms of our dietary example, if most people believe vegetarianism is the morally right choice, people who don't share that belief may act vegetarian simply to avoid disapproval. In this case, the disapproval of meat-eating means that it actually conveys higher integrity than vegetarianism, despite the costliness of vegetarianism, and so increasing levels of respect-based social pressure lead to increasing omnivorism.

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.

5.4 Modeling alternatives

Of course, while examining the robustness of certain simplifying assumptions reveals more about the chosen modeling approach, there are other modeling approaches that could have been used. In this section I will show how a few of these alternatives lead to largely similar behavior to either approval- or respect-seekers.

One modeling choice is that approval seekers are judged according to their actions rather than the norms they are inferred to hold. As mentioned above, this is because judging someone solely according to their norms gives no direct penalty to hypocrisy. But would such an alternative model actually yield such counterintuitive predictions if hypocrisy is indirectly punished by causing observers to doubt that the person truly believes what they say they believe?

Imagine that "camaraderie signalers" derive social image utility according to the following measure:

$$m_c(x) = \int_0^\infty \int_{-\infty}^\infty \phi(t,\rho) P(\rho_i = \rho | x) d\rho dt$$

That is, an observer judges a decision maker more favorably the more likely he thinks it is that they share moral beliefs. (We could also generalize this to reflect that closer norms are better, but such a distinction is immaterial in the binary norm/choice environment which I will stick to for simplicity and brevity.) This is now a signaling game, but the following result shows that camaraderie signalers behave qualitatively similarly to approval seekers, under the same assumptions as Proposition 1:

Proposition 8. *If individuals are motivated by camaraderie, $X = \{x_1, x_2\}$ with $v(x_2) > v(x_1)$, and ρ and t are independent, the following hold:*

1. *There is some value $\overline{p} \in [1/2, 1]$ so that if $p_1 > \overline{p}$ ($p_1 < \overline{p}$) then there exists at least one equilibrium in which all ρ_1 (ρ_2) types choose x_1 (x_2) and ρ_2 (ρ_1) types choose x_2 (x_1) iff they are sufficiently high-t.*

2. *$\lim_{s\to\infty} \overline{p} = 1/2$.*

3. *Increasing s in the limit leads to perfect conformity with the majority norm.*

The intuition of this result is that, while x_2 is attractive for material reasons, majority norm holders are always more likely to follow that norm so in order to signal camaraderie to this larger population, following their norm is attractive. If $p_1 < 1/2$ there is no conflict between these motivations and there is an equilibrium in which low-t types with ρ_1 abandon

their ideals to gain material and image utility from x_2. If p_1 is large enough, however, the larger population encouraging x_1 is enough to offset the material benefit of x_2, so there is an equilibrium in which ρ_2 types choose x_1. And, the more important social pressure is, the easier it is for this image utility to outweigh the material considerations, meaning that only a small majority is needed for x_1 to be conformed to. And of course, the higher social pressure is, the more perfect is conformity on the majority norm.

Aside from the possible existence of multiple equilibria and non-monotonic dynamics, this situation generally resembles the results of Propositions 1 and 2, providing some additional confidence in the simple and intuitive approach used to model approval seekers.

Another possibility is that people signal their norms, rather than the degree to which they believe in their norms. This would allow for a single-parameter model in which signaling a particular ρ yields a higher image utility instead of the two-parameter model of respect. This is essentially the model of Bernheim (1994), and so it deserves consideration at minimum due to precedence.

For precisions sake, suppose that "norm signalers" derive image utility according to the measure $m_n(x) = E[\rho|x]$. There is then the question of H_n: in some contexts higher m may intuitively be more admirable, such as when ρ represents the amount of charitable giving someone believes is appropriate. In other contexts a more centrist ρ may be more admirable, as assumed in Bernheim (1994). Let's start with the simplest situation in which we stick with the binary choice/norm environment of Proposition 1 and specify that H_n is either strictly increasing or strictly decreasing on $[\rho_1, \rho_2]$; as before, we assume $v_2 > v_1$.

This approach produces very different results from respect seekers; in fact equilibrium is qualitatively almost identical to approval seekers. I will omit the technical details for brevity, but briefly, because adhering to the admirable norm it is the most effective way to signal that you believe in it, the admirable choice always yields the highest image utility regardless of material costs. This plays the same role as the fact that the majority choice always yields the highest image utility regardless of material costs in the approval seeking model. At low levels of social pressure people trade off these image considerations with material considerations and imitation may occur in either direction, but as social pressure rises, conformity with the more admirable norm is the dominant concern, just as for approval seekers. The only difference is that social image utility is endogenously determined by the fraction choosing each option rather than exogenously by the fraction who believe in each option.

This was not the approach I used to model respect seekers due to the simple fact that the meaning of "heterogeneous norms" is quite unintuitive when there is nonetheless a particular

norm that is most admired by everyone. If I feel guilt for adhering to a norm that I admire, why is that not my norm and why does it induce guilt? This conceptual issue does not arise for Bernheim (1994) because individual types are said to represent "bliss points", i.e. preferences, rather than normative beliefs.

A final modeling choice that can be explored via an additional model extension is that individuals care about the esteem they receive from everyone equally. Instead, we may care more for respect and/or approval when it comes from those we share norms with. Because respect is accorded via rational inferences about type, the identity of the person making the inference is irrelevant, so no modification to the respect-seeking model is needed. But for approval seekers, image is the population average of individual judgements. If $\alpha \in [0, 1]$ represents the discount factor that approval seekers apply to social image granted by those they do not agree with, we can redefine:

$$m_a(x|\rho_i) = -\int_{-\infty}^{\infty}\int_{0}^{\infty} \phi(t,\rho)\alpha^{\mathbb{1}[\rho_i \neq \rho]} G(x-\rho)dtd\rho$$

The (mathematically trivial) technical details are omitted for brevity, but quite simply, what this means is that if α is large enough, people behave like approval seekers and social pressure promotes the majority norm, but if α is small enough, people behave approximately like respect seekers and social pressure discourages hypocrisy. This is an interesting modeling possibility to consider, however, because the extreme with $\alpha = 0$ captures the notion of the "reference network" which defines the audience whose opinion an individual cares about. Different reference networks may hold different norms, but within a group norms are homogeneous and social pressure acts unambiguously to promote the shared norm. As discussed in the following section, several counterintuitive findings have been attributed to reference networks, and this shows that an alternative explanation is that individuals do in fact care about the opinions of all audience members but that they seek respect, rather than approval.

6 Discussion

Most decisions governed by moral guidelines, customs and traditions, or fads and fashions (a very wide range!) fall into a domain in which people disagree about appropriate actions. The model presented here can be used to understand behavior such as middle school fads, child-rearing practices and taboos, religious practice, etiquette, local customs, and so on,

although the approach will of course need to be carefully adapted to the details of a setting. Particular empirical fields of economics that may be relevant are development (where local norms may differ from the norms of researchers or may differ *because* of the presence of outside researchers), politics and public policy (where people obviously hold diverging opinions about best policies), and marketing (because different demographics have different status incentives that drive their purchases).

A particular literature of recent interest that this model may be usefully applied to is the growing collection of findings that lack of anonymity does not always encourage prosocial behavior. Several authors have already speculated that their findings are due to heterogeneous norms. Dufwenberg and Muren (2006) find a 26.5% decrease in dictator game sharing when economics students must announce their choices on stage in front of their peers, rather than submitting them in anonymous envelopes, and they suggest that "These students want to please their fellow students, especially at the beginning of the course where people do not yet know each other but want to make friends. An aspiring economist may be well advised to conform to the economic stereotype of selfishness." Lambarraa and Riener (2015) similarly claim that Islamic values explain why Moroccan subjects increase their charitable donations when they are made anonymously when the request is made in Arabic (triggering Muslim identity) rather than French.

Several studies in the education sphere also suggest that different types of students care about, and care to signal, different values depending on the audience. Bursztyn, Fujiwara and Pallais (2017) find that single female MBA students report lower desired salaries and lower willingness to travel or work long hours when their responses are made public, especially to single male peers. Bursztyn et al. (2015) find that students reduce performance in remedial test-preparation software by 24% when public leaderboards are introduced; this is driven by top students most likely to be shown on the leaderboard. And Bursztyn, Egorov and Jensen (2019) find that public displays of SAT prep scores decrease SAT prep take-up in *both* low-income/performance and high-income/performance schools but for different reasons, the former because they want to hide effort and the latter because they want to hide low performance.

The notion of "reference networks" is sufficient to explain some (but not all) of these findings. As described in Section 5.4, approval seekers who appeal to their respective reference networks behave similarly to respect seekers, adhering closer to their own norms when social pressure is applied so that social image utility outweighs the material advantage of defecting. This (near-)equivalence implies that respect seeking in a population with hetero-

geneous norms is an alternative explanation for these findings. And more fundamentally, not all results are compatible with reference networks. The Dufwenberg and Muren (2006) results, for example, require that some students are signaling to people they do *not* share a norm with: If economics students who are selfish in public all truly held a selfish norm, they would not need the prompting of publicity to be selfish because that is, after all, the materially advantageous choice.

Because this literature nearly universally assumes that norms are homogeneous, or at the very least that they differ only between reference networks, concrete evidence of heterogeneous norms within reference networks, with which the model can be evaluated in detail, does not yet exist.[14] This is left to future work.

But, considering existing results in light of the possibility of heterogeneous norms points to important issues that demonstrate why attention to underlying norms and image incentives should be paid, not only for the sake of model testing, but for the sake of drawing correct inferences about welfare-enhancing policies from existing studies. Bursztyn, Fujiwara and Pallais's (2017) results from the perspective of homogeneous norms suggest that a harmful norm of discouraging female MBA ambitions should be broken. From the perspective of heterogeneous norms, it could be that single male peers cause single women to hide their true preferences, or it could be that their presence counterbalances a pressure from career counselors (who observe responses even in the anonymous condition) to exaggerate ambition. Similarly, Bursztyn et al.'s (2015) results, assuming that students are driven by approval, suggest that including observers such as teachers and parents could benefit non-honors students who might otherwise avoid effort to fit in with their peers. But if they are instead motivated by respect, a bigger audience may amplify social pressure and thereby cause students to try harder to signal their lack of care about schoolwork.

This point echoes existing discussions about the importance of context for designing prosocial nudges (Hauser, Gino and Norton, 2018). Bicchieri and Dimant (2019) demonstrate the importance of understanding injunctive and descriptive norms in context, but by admitting that in many situations the relevant audience consists of people with heterogeneous personal norms, I argue that this contextual understanding must also extend to the distribution of norms and relevant social image concerns. This is also true for informational nudges, as highlighted by Bicchieri and Mercier (2014), who note that information about compliance may not have any effect on someone who disagrees with the value of compliance

[14] te Velde and Louis (2021) introduce a norm measurement mechanism that is capable of measuring heterogeneous norms.

(i.e. who holds a different personal norm). This model may prove useful for understanding scenarios such as this by relying on the alternative interpretation of p_1 as the common *belief* about the prevalence of a norm, as mentioned in footnote 9; informational nudges then function by correcting these beliefs when pluralistic ignorance, i.e. incorrect common knowledge, prevails.

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

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

Adriani, Fabrizio, and Silvia Sonderegger. 2009. "Why do parents socialize their children to behave pro-socially? An information-based theory." *Journal of Public Economics*, 93(11): 1119–1124.

Adriani, Fabrizio, Jesse A. Matheson, and Silvia Sonderegger. 2018. "Teaching by example and induced beliefs in a model of cultural transmission." *Journal of Economic Behavior & Organization*, 145: 511–529.

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

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

Akerlof, George A., and Rachel E. Kranton. 2002. "Identity and schooling: Some lessons for the economics of education." *Journal of Economic Literature*, 40(4): 1167–1201.

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives*, 19(1): 9–32.

Akerlof, Robert. 2017. "Value Formation: The Role of Esteem." *Games and Economic Behavior*, 102: 1–19.

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

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

Ashraf, Nava, Oriana Bandiera, and Kelsey Jack. 2014. "No margin, no mission? A field experiment on incentives for pro-social tasks." *Journal of Public Economics*, 120: 1–17.

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

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

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

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

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

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

Bicchieri, Cristina, and Eugen Dimant. 2019. "Nudging with care: the risks and benefits of social information." *Public Choice*.

Bicchieri, Cristina, and Hugo Mercier. 2014. "Norms and beliefs: How change occurs." In *The complexity of social norms*. 37–54. Springer.

Bisin, Alberto, and Thierry Verdier. 2001. "The Economics of Cultural Transmission and the Dynamics of Preferences." *Journal of Economic Theory*, 97: 298–319.

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

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

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

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

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

Bursztyn, Leonardo, and Robert Jensen. 2017. "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure." *Annual Review of Economics*, 9(1): 131–153.

Bursztyn, Leonardo, Georgy Egorov, and Robert Jensen. 2019. "Cool to be Smart or Smart to be Cool? Understanding Peer Pressure in Education." *The Review of Economic Studies*, 86(4): 1487–1526.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From Extreme to Mainstream: The Erosion of Social Norms." *American Economic Review*, 110(11): 3522–3548.

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

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

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

Bénabou, Roland, and Jean Tirole. 2011*a*. "Identity, morals, and taboos: Beliefs as assets." *Quarterly Journal of Economics*, 126(2): 805–855.

Bénabou, Roland, and Jean Tirole. 2011*b*. "Laws and Norms." National Bureau of Economic Research Working Paper 17579.

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

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

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

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

Carvalho, Jean-Paul. 2017. "Coordination and culture." *Economic Theory*, 64(3): 449–475.

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

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

Chang, Daphne, Roy Chen, and Erin Krupka. 2019. "Rhetoric matters: A social norms explanation for the anomaly of framing." *Games and Economic Behavior*, 116: 158–178.

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

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

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

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

Dufwenberg, Martin, and Astri Muren. 2006. "Generosity, anonymity, gender." *Journal of Economic Behavior & Organization*, 61(1): 42–49.

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

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

Ellingsen, Tore, and Magnus Johannesson. 2011. "Conspicuous generosity." *Journal of Public Economics*, 95(9-10): 1131–1143.

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

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

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

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

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

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

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

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

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

Hauser, Oliver P., Francesca Gino, and Michael I. Norton. 2018. "Budging beliefs, nudging behaviour." *Mind & Society*, 17(1): 15–26.

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

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

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

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

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

Kritikos, Alexander, and Jonathan H. W. Tan. 2014. "Would I Care If I Knew? Image Concerns and Social Confirmation in Giving." IZA Discussion Paper No. 8739.

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

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

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

Lambarraa, Fatima, and Gerhard Riener. 2015. "On the norms of charitable giving in Islam: Two field experiments in Morocco." *Journal of Economic Behavior & Organization*, 118: 69–84.

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

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

Lin, Stephanie C., Rebecca L. Schaumberg, and Taly Reich. 2016. "Sidestepping the rock and the hard place: The private avoidance of prosocial requests." *Journal of Experimental Social Psychology*, 64: 35–40.

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

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

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

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

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

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

Miller, Dale T., and Deborah A. Prentice. 2016. "Changing Norms to Change Behavior." *Annual Review of Psychology*, 67(1): 339–361.

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

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

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

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

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

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

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

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

Tabellini, Guido. 2008. "The scope of cooperation: Values and incentives." *Quarterly Journal of Economics*, 128(3): 905–950.

te Velde, Vera L. 2018. "Beliefs-based altruism as an alternative explanation for social signaling behaviors." *Journal of Economic Behavior & Organization*, 152: 177–191.

te Velde, Vera L., and Winnifred R. Louis. 2021. "The (lack of) information value of descriptive norms." Working paper.

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

White, Katherine M., Joanne R. Smith, Deborah J. Terry, Jaimi H. Greenslade, and Blake M. McKimmie. 2009. "Social influence in the theory of planned behaviour: The role of descriptive, injunctive, and in-group norms." *British Journal of Social Psychology*, 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 $\tilde{t}_1 > 0$.

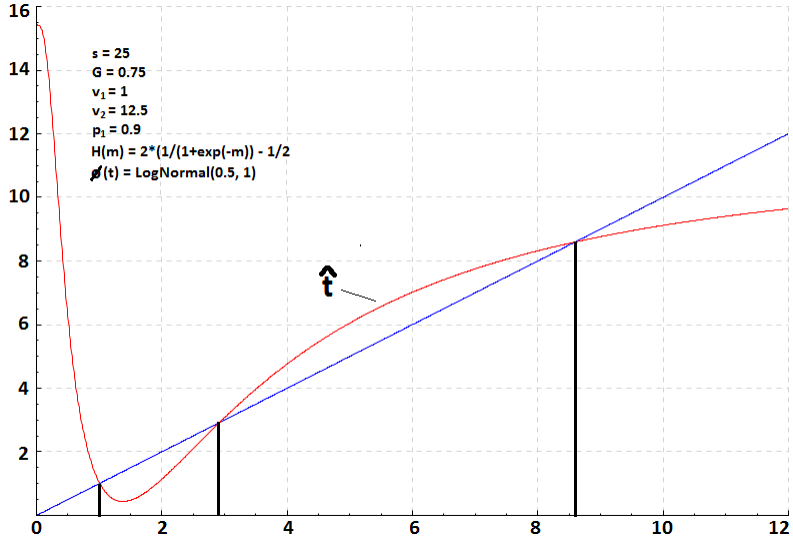
Define

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

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. (Intuitively, if everyone pooled on x_1, this could only be sustained if $E[t|x_2] < \bar{t}$, which doesn't make sense given that the benefit of deviation is increasing in t.) 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 Corollary 1: Follows directly from proof of Proposition 2.

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 $m(x_3) = -p_1 G_2$. 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.

Proof of Proposition 8: Following the same approach as Proposition 1 part 2, suppose $\tilde{t}_1 > 0$. Then $m_c(x_1) = p_1$ because x_1 identifies the actor as definitely holding ρ_1, and

$$m_c(x_2) = \frac{(1-p_1)^2 + p_1^2\Phi(\tilde{t}_1)}{(1-p_1) + p_1\Phi(\tilde{t}_1)}.$$

With $t = 0$, $\hat{t}(t) = \frac{s(H(1-p_1)-H(p_1))+v_2-v_1}{G}$. And with $t \to \infty$, $\hat{t}(t) \to \frac{s(H((1-p_1)^2+p_1^2)-H(p_1))+v_2-v_1}{G} < \infty$. Therefore, the intermediate value theorem guarantees the existence of such an equilibrium as long as $\frac{s(H(1-p_1)-H(p_1))+v_2-v_1}{G} > 0$. This may or may not be true.

On the other hand, suppose $\tilde{t}_1 < 0$. Then $m_c(x_2) = 1 - p_1$ and

$$m_c(x_1) = \frac{p_1^2 + (1-p_1)^2\Phi(\tilde{t}_1)}{p_1 + (1-p_1)\Phi(\tilde{t}_1)}.$$

With $t = 0$, $\hat{t}(t) = \frac{s(H(p_1)-H(1-p_1))+v_1-v_2}{G}$, and with $t \to \infty$, $\hat{t}(t) \to \frac{s(H(p_1)-H((1-p_1)^2+p_1^2))+v_1-v_2}{G} < \infty$. Therefore, the intermediate value theorem guarantees the existence of such an equilibrium as long as $\frac{s(H(p_1)-H(1-p_1))+v_1-v_2}{G} > 0$. This may or may not be true.

However, because these two conditions are negations of each other, one or the other must be true, which guarantees existence of an equilibrium. Also, note that if $p_1 < 1/2$, then $H(1-p_1) > H(p_1)$ unambiguously and likewise the first condition holds unambiguously. But if $p_1 > 1/2$, in particular if it exceeds some threshold $\overline{p}$ (potentially equal to 1), then the latter condition holds instead. This proves the first part of the Proposition. Also, the higher is s, the less $H(p_1)$ needs to exceed $H(1-p_1)$ by in order for the second condition to hold. This proves the second part of the Proposition. Finally, as $s \to \infty$, the relative value of $H_2 - H_1$ dominates $v_2 - v_1$ in determining which condition holds, so that imitation is guaranteed to be in the direction of the majority norm, and $\tilde{t}_1$ (in the first case) or $\tilde{t}_2$ (in the second case) increases arbitrarily. This proves the third part of the proposition.