Heterogeneous norms: Social image and social pressure when people disagree

Vera L. te Velde*

October 28, 2015

Abstract

People are often divided by what they believe is the right thing to do, such as in partisan politics or when norms evolve. Traditional notions of social norms and social incentives break down in these situations. Using a psychological game theoretic model, I show how social pressure affects behavior in these heterogeneous norm settings through two possible channels: "approval" is conferred when actions are deemed correct and "respect" when actions are motivated by strong personal beliefs. These motives lead to different outcomes in terms of consensus, hypocrisy, compromise, and destructive posturing. Results demonstrate how using social incentives to change behavior, an increasingly popular tactic in domains such as development, may easily backfire if there is hidden heterogeneity in norms.
JEL classification: D03; Z13; C72; D71.
Keywords: Norms, Social Pressure, Social Image, Signaling, Psychological Game Theory, Customs, Morals, Persuasion

1 Introduction

In only the last decade or so, social image and social pressure have emerged as key determinants of prosocial behavior in the economics literature[1]. They have been exploited to

*te Velde: University of Queensland School of Economics, Colin Clark Bldg 39 Level 6, St. Lucia, QLD, 4072, Australia. v.tevelde@uq.edu.au. 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, 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.

[1]For experimental evidence using manipulated anonymity, see 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), and Satow (1975). For evidence on people's desire to avoid being put in pressuring situations, see Dana, Cain and Dawes (2006), Broberg, Ellingsen and Johannesson (2007), Lazear, Malmendier and Weber (2012), and Malmendier, te Velde and Weber (2014).

increase voter participation in a highly cost-effective manner (Gerber, Green and Larimer, 2008), been shown to increase donations to charity by up to 42%, (DellaVigna, List and Malmendier, 2012), and been used to promote safe sex practices more effectively financial incentives (Ashraf, Bandiera and Jack, 2012).[2] But settings like voting and donating to charity, in which there is an essentially unanimously recognized "good" 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.

In these settings with heterogeneous norms, existing notions of social pressure break down. Two types of social image - respect and approval - might now push in opposite directions; for example, voting for a tax hike might be respected as a principled move, especially if personal sacrifice is involved, or it might be condemned as an unpopular choice. In turn, the tradeoffs people are willing to make between self-interest, moral[3] beliefs, and social image may be different depending whether they are more concerned with respect or approval.

In order to understand these situations, I adopt a modeling framework that is portable across settings in which individuals honestly disagree about the correct course of action. Social image in this model might come in the form of either respect or approval (or both). This constitutes a couple of key contributions. First, I introduce the concept of social norms as population aggregates of potentially heterogeneous personal norms. Second, I define the distinct concepts of respect- and approval-based social image. Third, I characterize how social pressure influences behavior when people have these types of social image motivations. Fourth, I demonstrate that efforts to influence behavior critically depend on the type of social image motivations that people have and the heterogeneity in norms in the population and describe scenarios in which not only will social pressure be ineffective at encouraging a desired behavior, it may backfire.

Reconceptualizing norms is necessary to understand moral decisions more complicated than a simple choice between right and wrong. Identifying the right action is rarely straightforward and usually entails choosing an arbitrary point from the universe of opinions as the

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

[3] I will use the term "moral" to describe these beliefs, although many decisions involving heterogeneous norms may have little moral content, such as what to wear to work.

norm, such as a particular acceptable amount to donate to charity. Referring to "heterogeneous norms" is in fact an oxymoron by traditional definitions, which defines norms as unanimously recognized behavioral guidelines (Bicchieri, 2006; Ostrom, 2000). I broaden this definition to allow each individual to hold a distinct idea of the impartial appropriateness of each option, so that the consensus norm reflects an aggregate agreement that a particular action is the most moral.[4] This allows us to talk about settings with heterogeneous norms as nonetheless governed by moral logic, such as wealth redistribution, civil rights, lifestyle choices, etc.

This individualized notion of social norms requires a new understanding of social image motivations. Extensive empirical and experimental evidence points to an intimate link between social norms and social pressure, but this relationship clearly breaks down when people disagree about norms: whose opinion determines the social image resulting from a choice? Individuals now differ in two dimensions, their dedication to their norms and their norms themselves. Approval and respect are the two types of social image that correspond to these two dimensions of individuality and are clearly defined even when norms are heterogeneous.

"Respect" is the social image that accrues to those who always do what they personally believe to be right. "Respect seekers" want to signal their integrity, i.e. the strength of their norms. When (passively) acting as part of the audience, they infer others' integrity and pass judgment on them accordingly. Since integrity is unobservable, observers must infer integrity from choices. Individuals, therefore, would like to make choices that are most strongly attributable to high integrity. They might be vegan, for example, in order to show that they are willing to make personal sacrifices to stay true to their beliefs about animal cruelty.

"Approval", on the other hand, is the social image that results when an observer thinks that the decision maker is doing the right thing. Approval seekers want to make their peers happy by going along with everyone else's beliefs, and they are tempted to abandon their own norms in order to do so. Approval is based on actions, rather than inferred types, so approval emerges in a more straightforward non-signaling game in which individuals simply choose the best trade-off between consumption, guilt, and image, without hoping to

[4]Cialdini, Kallgren and Reno (1991) and Bicchieri (2006) distinguish between injunctive and descriptive norms, respectively the absolute morally appropriate action and the action that is typically chosen. Krupka and Weber (2013) broaden this definition to a set of commonly recognized "moral appropriateness" ratings of *every* option. My notion of personal norms are essentially an individualized version of the latter; descriptive norms, describing what people in fact do choose in that situation, are irrelevant. In particular, and importantly, this rules out norms of the form "do what everyone else is doing".

convincingly imitate other types. This approach 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.

An analysis of respect- and approval-seeking behavior shows that while traditional models of social image motivations are powerful in homogeneous norm domains, those resulting predictions constitute a knife-edge case within the broader universe of heterogeneous norms.

The analysis focuses on the change in population behavior when social incentives change. Rising social pressure increases the weight that people place on social image motivations, but strengthening the approval motive changes behavior very differently than strengthening the respect motive. As vegetarianism becomes the hot topic at the water cooler, respect seekers care more about proving that their dietary choices reflect their true beliefs. They respond to this pressure by trying harder to imitate high integrity types. They look for convincing ways to signal high integrity, possibly using costly signals to establish credibility. They might give up meat or purchase pricier cage-free eggs and free range chicken, to signal that they care about animal rights, or they might take up hunting in order to convince others that their meat-eating lifestyle is a moral choice. Choices can become *more* divided along moral lines, and compromise becomes harder to achieve as individuals become less willing to be seen conceding their norms. If individuals are simultaneously motivated by both respect and approval, they may even be more willing to accept disapproval as another way to send a convincing signal of integrity.

Approval seekers, on the other hand, respond to social pressure by trying harder to please the majority. This leads to conformity: the more important image motivations are, the more power the majority has to sway choices. Approval-seeking vegetarians will hide their preferences and go along with friends to the sushi bar or give up on herbivorism altogether. But when possible, approval seekers may prefer to look for ways to avoid offending any one subgroup too much, leading to widespread compromise. Individuals of all beliefs might be able to agree that free range chickens and farm-raised fish are fine to eat but that beef is not, thereby avoiding offending either the paleo dieters or the animal rights activists. In any case, high social pressure leads to deceptively uniform behavior, disguising the disagreement in the population.

I also find that approval seekers and respect seekers have different proclivities towards "sacrificial" norms that many people would unambiguously prefer not to exist at all. A majority of approval seekers can successfully bully the rest of the population into destroying utility for the sake of demonstrating adherence to a norm *they don't believe in.* Respect

seekers, on the other hand, are less likely to be able to sustain a conformist equilibrium of any kind, for better or for worse. This suggests that the approval motivation might be masking substantial heterogeneity in beliefs in situations with an evidently damaging norm (such as female genital mutilation, for one of many examples), so that reducing social pressure, rather than trying to change norms directly, could improve matters.

The contrast between the effects of respect and approval implies difficulty with using social incentives to influence behavior. But if the relevant social incentives are understood in a given context (both the distribution of personal norms and the type(s) of social image that is relevant), the model can straightforwardly predict behavioral change resulting from manipulation of social pressure.

But more importantly, the model illustrates how designing social incentives under a naive understanding of norms can backfire. Respect and approval both unambiguously motivate norm adherence when norms are homogeneous, so it's understandable (and certainly often highly effective) to try to use social incentives to change behavior when norms are seemingly universal. But heterogeneity in norms is not always obvious. For example, surely most teachers would say they recognize the importance of taking their job seriously, but honoring good work could reduce performance if they are actually afraid of being disapproved of by their less dedicated peers. In development economics in particular, heterogeneity in norms can be introduced simply through resistance to outsiders. Social incentives in development initiatives have yielded some notable successes, but this increasingly popular tactic raises a red flag when understood within this model of heterogeneous norms.

These applications are discussed further in section 5. Until then, section 2 will formalize the model of respect and approval, and sections 3 and 4 will explore the different behavior of approval seekers and respect seekers.

2 Model

Consider a setting in which an individuals 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 includes the immediate costs and benefits of the action along with any expected long run change in utility, such as the expected change in tax policy after volunteering to campaign for a particular candidate. This setting is intended

to intuitively capture a wide array of moral and customary decisions, such as whether to eat meat, which church to go to (if any), whether to send one's kids to private school, how to share resources, how to reciprocate kind actions, what to wear to work, what brand of shoes to buy, who to vote for, etc.[5] But because these are diverse examples that each deserve detailed discussion, I will refer to more stylized examples when possible in order to avoid excessive peripheral discussion, and will rely on the reader to imagine how the results apply to their preferred context.

Formally, an individual i is faced with a choice set X. Each option $x \in X$ leads to personal consumption utility $v(x)$, but in addition, i has a personal norm denoting $\rho_i \in X$ as the morally appropriate choice. When making a choice x_i that deviates from this norm, i pays a psychological cost $G(x_i - \rho_i)$, 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.

Note that only the decision maker's utility is directly modeled here; externalities are discussed further in section 3.3. The impact of a choice on others is indirectly captured through G, since that impact may of course be part of what determines the psychological cost of that choice. I will correspondingly refer to G as the *guilt* function, although guilt may not be an accurate word for all psychological costs associated with deviating from normative behavior, such as warm glow, altruism, cognitive dissonance, etc.

When types are specified in this two dimensional way, there is a natural choice in the definition of social image: do people want others to share their moral beliefs and to signal that they share theirs with the group, or do they want to signal that they adhere to their own idea of right and wrong no matter what material consequences are involved and no matter what other people think? This is the intuition driving the formulation of two models of social image: respect seekers signal integrity, and approval seekers signal their norms.

Social image utility (either from approval or respect) is an increasing function given by $H(m(x_i))$, where m is the image resulting from a given choice x_i. 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

[5] While voting and similar choices are themselves private, it's reasonable to interpret many people's voting decisions as visible amongst their social group, as these topics frequently arise in conversation and there is substantial evidence that people dislike lying, in general and about voting in particular, e.g. Gneezy (2005) and DellaVigna et al. (2013). Moreover there is evidence that social image motivations strongly influence even anonymous choices (Lazear, Malmendier and Weber, 2012; Malmendier, te Velde and Weber, 2014), perhaps through self-signaling (Bénabou and Tirole, 2011; Rabin, 1995) or beliefs-based altruism (te Velde, 2014).

image, such as visibility, audience size, and how harshly choices are judged in a particular setting. [6]

Altogether, person i has utility

$$U(x_i|t_i, \rho_i) = v(x_i) - t_i G(x_i - \rho_i) + sH(m(x_i)). \tag{1}$$

Further assumptions may be helpfully motivated with an example. Imagine that the relevant choice is to argue or vote for a redistributive tax schedule. The decisionmaker believes that ρ represents the best trade-off between equality and efficiency but would rather not pay such a high rate, so he might facetiously argue against social safety nets. He may not even have the option of voting for his most preferred tax rate, if only a few options are on the table. Others disagree with his moral beliefs, and some argue in favor of credible alternatives, while some argue for transparently selfish policies that no one believes are fair. Any of these morally contentious decisions can be understood to be contained within the general scenario specified by the following intuitive, non-restrictive assumptions:

Assumption 1. *i) X is indexed by $\mathbb{R}$. (Results below specify $X = \{x_1, x_2\} \in \mathbb{R}^2$ or $X = \{x_1, x_2, x_3\} \in \mathbb{R}^3$.)*

ii) $s \geq 0$.

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

iv) $G(x_i - \rho_i)$ is increasing in $x_i > \rho_i$ and decreasing in $x_i < \rho_i$. Normalize $G(0) = 0$. (Results below take G to be symmetric around 0.)

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

Continuity of ϕ is assumed to guarantee existence of equilibrium; this could be relaxed in specific instances to allow, for example, an atom in the distribution of types. Symmetry of G is not at all crucial but reduces the number of cases to consider and thus eases

[6] "Social pressure" may not be the perfect term to capture all of these various factors, but it is convenient to have a shorthand term for the "weight placed on social image". This definition conflates approval and respect but is consistent with terminology from the literature on image effects in homogeneous norm settings, and since many of the same things (like anonymity and publicity) intuitively increase the weight on both types of image, it isn't unreasonable to keep the single term. This also serves to highlight my point, from the introduction and discussion, that manipulating social pressure can backfire if in fact approval and respect are at odds due to heterogeneous norms.

exposition. Similarly, binary and ternary choice sets analyzed are also not crucial; larger discrete choice sets can be analyzed similarly but the added complexity does not produce valuable returns to intuition or insight.

I begin by considering two distinct populations of people with two different types of social image motivations, the *approval seekers* and the *respect seekers*. They respectively have social image functions m_{as} and m_{rs}, and image utility functions H_{as} and H_{rs}, for which part 5 of assumption 1 holds separately. Approval seekers and respect seekers are first analyzed as disjoint populations for the sake of clearly contrasting the effects of the two kinds of social image, but of course often these two motivations might operate simultaneously; Section 4.3 explores the interaction between approval and respect.

Approval seekers: Approval seekers derive utility from praise for their actions, and observers praise actions that agree with their personal norms. Note that approval seekers are not concerned with actually signaling either their norm ρ or their integrity t; they only seek praise for their 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: a vegetarian would have to approve of an admittedly hypocritical, lapsed vegetarian friend just because they philosophically agree about the merits of vegetarianism. The non-signaling specification is more realistic: it's quite plausible that a vegetarian would be happy to convert an insincere meateater, or that Republican constituents would be happy to continue reelecting a Democrat so long as that representative didn't introduce new taxes. These are scenarios in which individuals might confer approval, but *not* respect (as defined below).

For approval seekers, $m(x)$ is defined as $m_{as}(x) = -\int_{-\infty}^{\infty}\int_{0}^{\infty}\phi(t,\rho)G(x-\rho)dtd\rho$. That is, observers each judge a choice of x according to their personal guilt functions, and the decisionmaker's social image is the negative of the average of these individual judgments over the full population. For example, if half of the population believes in ρ_1 and half the population believes in ρ_2, then $m(\rho_2) = -\frac{1}{2}G(\rho_2-\rho_1)$. The best attainable image, $m_{as} = 0$, only occurs when perfectly adhering to a homogeneous norm.

This approach is based on modeling efforts in homogeneous norm contexts. Akerlof (1980), most similarly, models social customs in which individuals feel social pressure to adhere to a norm and social pressure increases in the fraction of the population that believes in the norm. Akerlof and Kranton (2000) model choices as influenced by the identities we derive from membership in different social categories. Harbaugh (1998*b*), Harbaugh (1998*a*), and Glazer and Konrad (1996) model the prestige motivation for charitable giving

and find empirical evidence from the desire of donors to be listed in elite categories of supporters.

Respect seekers: For respect seekers, social image is based on an observer's estimate m of the decisionmaker's integrity t. Formally, their image function is simply defined as $m_{rs}(x) = E[t|x]$, the rational inference of someone's t conditional on the choice x. Respect-seeking individuals want to be seen as unhypocritical, whatever their personal beliefs. However, since integrity is not directly observable, inferences about integrity must be rational in equilibrium, and optimal choices must anticipate those equilibrium inferences. Note that observers do *not* need to care about the decisionmaker's ρ (or if they do, a pure respect seeker will only care about the observer's inference of t_i). There are intuitively many scenarios in which consistent adherence to beliefs is emphasized over the beliefs themselves (I can admire vegans without believing that such extremes are necessary) but the analysis of section 4.3 allows for both respect and approval to operate in tandem for the situations when complete separation isn't plausible.

This signaling model of respect builds on similar approaches in homogeneous norm contexts as well. Andreoni and Bernheim (2009), Bernheim (1994), Bénabou and Tirole (2006), Seabright (2009), and Grossman (2015) all model altruistic decisions as signals of altruistic type parameters.

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 in the realm of psychological game theory (Geanakoplos, Pearce and Stacchetti, 1989; Battigalli and Dufwenberg, 2009). As long as types are exogenously assigned, however, 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. The respect seekers play a much more complicated signaling game, and analysis relies on that notion of psychological equilibrium.

Homogeneous Norms Before presenting the main results, a short note on the baseline setting with homogeneous norms is called for. Even in these scenarios, the two models

lead to slightly different predictions, fundamentally because respect seekers are playing a signaling game, while approval seekers are being judged directly for their actions rather than inferences based on those actions.

First of all, the respect-seeking model can generate pooling equilibria in situations where the approval-seeking model can't. This is demonstrated by Andreoni and Bernheim (2009) (or Bernheim's (1994) model of conformity) which can be seen as a (continuous choice) 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 at this point. Approval seekers in the same setting[7] 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. The pooling outcomes generated by respect-seeker signaling only occur for approval seekers if we allow incidental structural features in their preferences. See also Seabright (2009).

Second of all, the respect-seeking model produces interdependent preferences more naturally than the approval-seeking model. Interdependent preferences exist when an individual's preferences depend on the preferences of others (Gul and Pesendorfer, 2006; Postlewaite, 2011), such as being altruistic only towards altruistic people (Levine, 1998). This behavior is easily understood in the respect seekers' signaling game. If someone shares \$1 in a dictator game when everyone else shares nothing, that person might also choose to share \$4 if everyone else starts sharing \$3 in order to continue signaling an above average level of generosity. Whether this switch arises because isolated games are played with evolving, rather than equilibrium, beliefs, or because multiple equilibria exist, interdependent preferences appear endogenously for respect seekers but not for approval seekers (without modifying assumptions).

Having completed the model setup, we can now turn to the main results describing how respect seekers and approval seekers behave in settings with heterogeneous norms.

[7]Bernheim (1994) emphasizes this difference explicitly as well.

3 Results

The following five subsections characterize the behavior of approval seekers and respect seekers facing several types of choices.

3.1 Equilibrium

The main intuition for the difference between respect and approval arises clearly in the simplest possible binary choice environment. Assume that each individual has exactly two options, x_1 or x_2, and one of two personal norms, $\rho_1 = x_1$ or $\rho_2 = x_2$. Each option provides consumption utility $v(x_i)$ to the decisionmaker; WLOG assume $v(x_2) > v(x_1)$. Guilt is given by $G(x_1 - \rho_2) = G(x_2 - \rho_1) = G$. Additionally, assume that t is distributed according to ϕ_t, independently from ρ (independence will be relaxed in section 4.1. 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 $\rho = \rho_1$.

For concreteness, imagine the following stylized example. The decisionmaker must choose between two allocations of wealth for himself and a partner. x_2 corresponds to $(3, 0)$; that is, 3 units of utility for the decisionmaker and 1 for the partner. x_1 corresponds to $(1, 1)$. Some subset of the population is utilitarian and believes x_2 is the fairer allocation. Another subset is egalitarian and believes that x_1 is fairer. Individuals who don't care too much about following their own norms would prefer to choose the personally advantageous allocation x_2.

Whether a respect-seeker or an approval-seeker, individuals compare the utility of choosing their personal norm and avoiding guilt to the utility of the guilt-inducing other option. Those with $\rho_i = x_1$ choose x_1 if $v(x_1) + sH(m(x_1)) > v(x_2) - tG(x_2 - x_1) + sH(m(x_2))$, which is true if t is sufficiently large. Likewise for those with $\rho_i = x_2$, but the condition is easier to meet since $v(x_2) > v(x_1)$.

The role of H depends on whether individuals are approval seekers (with H_{as}) or respect seekers (with H_{rs}). Proposition 1 describes equilibrium for approval seekers, who perform a straightforward utility maximization, and for respect seekers, which emerges in a more complicated 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 $P(x_i) = E[t|x_i]$. Equilibrium transfers must be optimal given P and inferences must be consistent with Q. Throughout

this paper, I also restrict attention to equilibria satisfying the D1 criterion of Cho and Kreps (1987), which requires that inferences about types from disequilibrium actions must be reasonable in the sense that, roughly, all weight must be placed on the types who would be tempted to deviate to that action for the widest range of mistaken beliefs.[8]

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

1. *Among approval seekers:*
   - *Sufficiently high-*t *types will choose consistently with their norms. Lower-*t *types 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. Equilibrium is unique if 1)* G *is sufficiently large, 2)* s *is sufficiently small, 3)* p_1 *is sufficiently small, or 4)* $\max \phi(t)$ *is sufficiently small.*
   - *In any equilibrium, sufficiently high-*t *types will choose consistently with their norms and low types 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 norm, then anyone with either lesser t or the opposite ρ will have an even stronger motive to choose the same. And, if imitation is in the direction of choosing x_1, then the average t of people choosing x_1 will be *lower* than those choosing x_2, which would create an unequivocal motivation to defect instead to x_2. So, imitation can only occur as stated in the proposition.

[8] 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 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 cost of norm adherence is a key determining factor in aggregate behavior. 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 ρ_i, they are tempted to choose x_i 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.[9]

Proposition 1 also doesn't rule out multiple equilibria for respect seekers in general, although it provides some conditions that ensure uniqueness. These are quite conservative conditions, and the meaning of "sufficient" is of course jointly determined and contingent on all parameters and the distribution of t. But they are informative: If guilt is sufficiently powerful or social pressure sufficiently weak, the system approaches a non-signaling model in which only self-interest and morality determine decisions and hence a unique equilibrium exists. That is, the special case of zero social pressure is not a knife-edge case; the model behaves similarly in an (often large) region of parameters. Additionally, if p_1 is small enough or ϕ is "flat" enough, the inference function is essentially forced to behave convexly enough to ensure a unique equilibrium; see the Appendix for details.

For approval seekers, this result implicitly states that when *norms* shift slightly in the population such that the majority belief changes, *behavior* of approval seekers can shift much more dramatically and suddenly. This may be apparent, for example, in the shifting tide of public opinion about marriage equality. Meta-surveys indicate that 2010 or 2011 was when a majority of Americans first supported marriage equality, but the shift has been slow and steady (Silver, 2011). Support among senators, however, has changed much more dramatically, and more quickly than can be accounted for by turnover: only 15 senators openly supported marriage equality in 2011, and 51 did as of April 2013 (Matthews, 2013). Since senators derive utility (re-election) exactly from pleasing the largest fraction of the population, approval seeking is a likely explanation for at least part of this phenomenon. Similar forces may be behind sudden changes in taboos, such as political correctness. Behavior can appear to be nearly unanimous, but heterogeneous beliefs are simply hidden due to high social pressure.

[9] Note that respect seekers, but not approval seekers, can display the counterintuitive behavior explored theoretically by Bénabou and Tirole (2006) and demonstrated by Ariely, Bracha and Meier (2009) and Carpenter and Myers (2010): Extrinsic motivations may interact with image motivations in a negative way, because the signaling value of doing good is reduced when doing good is rewarded.

3.2 Changes in social pressure

More than the static equilibrium, however, we are interested in the response of aggregate behavior as s changes so that we might understand how social pressure influences norm adherence. Proposition 2 summarizes the high pressure equilibria for both approval seekers and respect seekers in the same setting as proposition 1:

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

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

2. *For respect seekers, as $s \to \infty$, the fraction of the population choosing $x_i = \rho_i$ approaches 100%, and $m(x_2)$ approaches $m(x_1)$.*

Once again, the intuition for approval seekers is very straightforward: as social pressure increases, it's more likely that the option providing the best combination of consumption, guilt, and image utility is the one with the higher image, since that component is weighted by social pressure in the utility function. For respect seekers, the intuition for the limiting case comes from the fact that if there were any difference in image between the two options (i.e., if $m_1 > m_2$ strictly), high enough social pressure would lead to crowding on m_1. But we know imitation can't occur in this direction, from in Proposition 1. The only possibility satisfying the equilibrium restrictions is that both options lead to the same image. Proposition 1 says that the only way this is possible is if everyone sticks to their personal norms.

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

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

Our stylized example of wealth allocation can make the contrast between respect- and approval-seekers clear. Respect-seeking egalitarians believe that $(1, 1)$ is fair despite selfishly wishing to choose $(3, 0)$. Only the egalitarians who care enough about fairness will

Figure 1: Approval seekers' versus respect seekers' choices

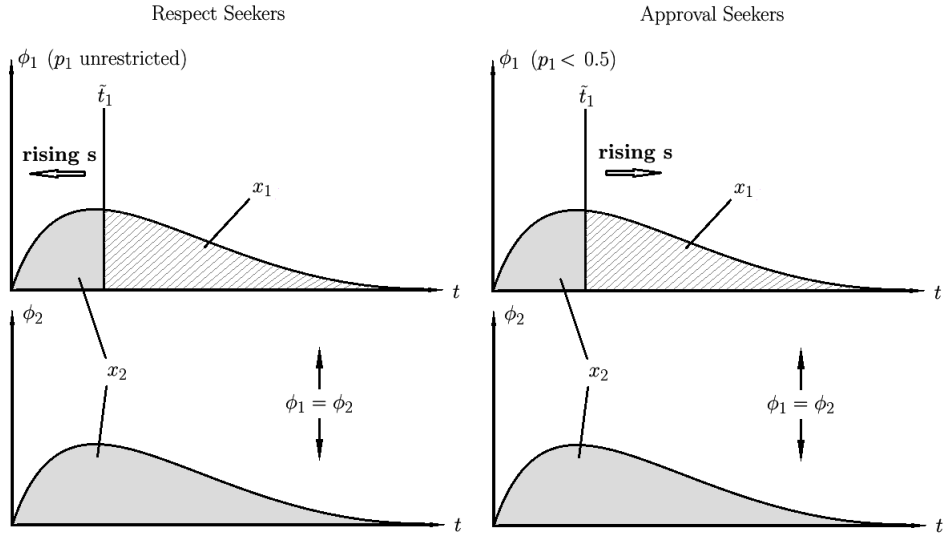


Illustration of choices made by the distributions (of t) of approval seekers (right) and respect seekers (left) with personal norm x_1 (top) and x_2 (bottom). Types in the shaded regions of the distributions choose x_1, and types in the filled regions choose x_2. Increasing social pressure causes more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

choose $(1, 1)$. Realizing this, observers admire the integrity represented by the choice of $(1, 1)$ even though many individuals who choose $(3, 0)$ are equally committed to their utilitarian beliefs. If social pressure is very high, more egalitarians will choose $(1, 1)$ due to the added benefit of a better social image. In the limit, all egalitarians choose $(1, 1)$, all utilitarians choose $(3, 0)$, and remarkably, no one is suspected of hypocrisy.

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

Many tactics (more subtle than attempting to change norms directly) used to encourage certain behaviors are understandable with these results. Shaming the rich for self-interestedly voting for low taxes is clearly an attempt to hurt their credibility (i.e. to confer disrespect) and discourage that kind of hypocrisy. The model predicts limited success with this tactic up until the point when voting is sincere, expectations reflect this, and accusations of hypocrisy are no longer credible. Influencing behavior through ap-

proval relies on having the majority norm. The Human Rights Campaign rather explicitly acknowledged this after beginning a new campaign following the change in the majority opinion of marriage equality in around 2011. Admitting limited success with changing minds directly, they switched to changing perceptions of $p > 0.5$ by "trying to foster the sense that ... history has already ruled in favor of their cause" (Issenberg, 2013).

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

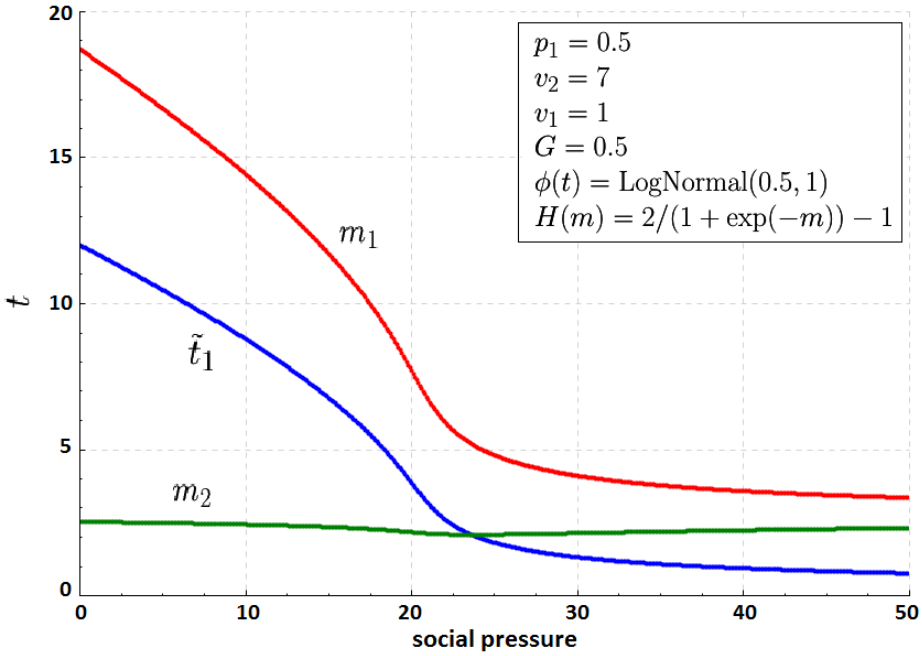
Figure 2 shows an example of these relationships, with parameters chosen so that a unique equilibrium exists at all levels of social pressure.

3.3 Sacrificial equilibria

What are the welfare implications of these behavioral responses? The model does not specify the impact of choices on others' utilities, so without specifying material externalities, or how moral hypocrisy affects others, welfare isn't clearly defined in this model. 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 general patterns in average U_i, as defined above, worth noting. When individuals sacrifice material utility to adhere to their norm, we can defend this choice as utility enhancing because they value the norm itself. And when someone fails to follow their norm in order to obtain greater material utility, we infer they care less about the norm than the material outcome. Neither of these situations is remotely surprising or uncommon in other models of social image and norms. But it's much more surprising if

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_2 and the top line m_1, both equilibrium values as a function of s. Note that as social pressure rises, the gap between m_1 and m_2 shrinks, so that in the limit either action will yield the same level of respect.

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 in order to attain a social image 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 $\rho = x_2$ nonetheless choose a more costly option, x_1. (And an equilibrium is said to be more sacrificial when the fraction of the population who does this rises.)*

These sacrificial equilibria are very 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. Respect seekers, on the other hand, are never able to sustain a sacrificial equilibrium.

Proposition 3. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, $G(x) = G(-x)$, and $\rho \in \{\rho_1, \rho_2\}$ 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 follows directly from the previous section. Respect seekers are disinclined towards conformity but a majority can drive consensus behavior of approval seekers as social pressure rises, even if that consensus is sacrificial. Approval seekers are thus more likely to be able to sustain an equilibria in which, for example, individuals choose allocations of $(0, 3)$ over $(1, 1)$. This would be a welfare increasing result, but a similarly maintained equilibrium in which most individuals choose $(1, 1)$ instead of $(3, 0)$ would be welfare reducing. In either case, consensus behavior may be masking substantial heterogeneity in beliefs.

The following section examines the robustness of these basic results, and in the process identifies several additional surprising phenomena.

4 Robustness and Extensions

Before exploring how the model operates in more complex settings, let's recall the main results from the simple setting of propositions 1, 2, and 3:

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 a choice of multiple equilibria.

3. Respect seekers, but not approval seekers, always respect the higher cost option more than lower cost option.

4. As social pressure rises, respect seekers become strictly divided along norm lines, while approval seekers form a perfect consensus 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 disappears.

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

4.1 Correlated type parameters

The initial analysis was simplified by the assumption that t and ρ are independent, which I now relax. Assume the same setting, but distinguish between the distributions of types t among those with ρ_1 and ρ_2: $t|\rho_1 \sim \phi_1$, $t|\rho_2 \sim \phi_2$. 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 high integrity.

The qualitative behavior of approval seekers does not change, since part 1 of Proposition 1 does not depend on independence of ρ and t. Proposition 4 describes the behavior of respect seekers:

Proposition 4. *If $X = \{x_1, x_2\}$, $v(x_2) > v(x_1)$, $G(x) = G(-x)$, and t and $\rho \in \{\rho_1, \rho_2\}$ are correlated, then for respect seekers:*

1. *There exists an equilibrium in which sufficiently high-t types choose according to their norm, and lower types all choose either x_1 or x_2.*
2. *As $s \to \infty$, if $E[\phi_1] > E[\phi_2]$ ($E[\phi_2] > E[\phi_1]$), only those with $\rho_i = x_1$ ($\rho_i = x_2$) and sufficiently high t will choose x_1 (x_2).*
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 allows for imitation to occur in either direction, since there are now two mechanisms with which to signal integrity: acting in according with the elite norm or choosing the costly action. This modifies our understanding of low-t behavior, result 1 above, to take into account their desire to be associated with the elite norm.

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. In the degenerate case, doing something that is *no* one's norm, such as burning money, can't possibly signal integrity. This is a very 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 approval seekers' sacrificial equilibria (see section 3.3), 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 almost an entire population of approval seekers to choose the sacrificial action, but rational expectations limit sacrificial behavior among respect seekers to only a few 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 operation of approval and respect is similar. Theoretically, its importance lies in allowing us to think about heterogeneity in consumption utility within this model; see section 4.4.

4.2 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; in particular, continuous choice settings with homogeneous norms that have been analyzed in work such as Andreoni and Bernheim (2009) and Bernheim (1994) can be embedded in this model via a fine discretization of the choice set.

Consider a setting in which x_i is chosen from $\{x_1, x_2, x_3\}$. ρ_i is either $\rho_1 = x_1$ or $\rho_3 = x_3$, and x_2 is a middle ground option: $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$. As before, fraction $p_1 \in (0, 1)$ have ρ_1 and the remainder have ρ_3. Without loss of generality, assume that $v_3 > v_1$. As in section 1, ρ and t are independent, which allows us to focus exclusively on the potential for compromise.

Proposition 5 describes the equilibrium:

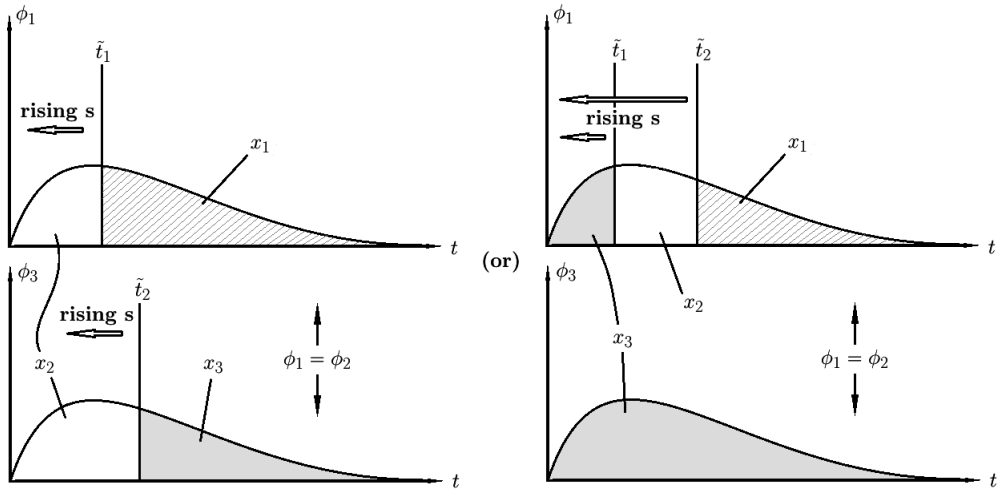
Proposition 5. *If $X = \{x_1, x_2, x_3\}$ with $v_3 > v_1$, $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$, and $\rho \in \{\rho_1, \rho_3\}$ is uncorrelated with t, then the following hold:*

1. *For respect seekers, at least one equilibrium exists, in which sufficiently high-t types adhere to their norms. Among lower types, 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 the remainder choose x_3.*

2. *For approval seekers, a unique equilibrium exists in which sufficiently high-t types adhere to their norms and lower types either 1) all choose x_1, 2) all choose x_3, or 3) mid-level-types with ρ_3 (ρ_1) choose x_2 and the remainder choose x_1 (x_3).*

Figure 3 illustrates two possibilities for respect seekers, one in which all low-t types defect to the profitable 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 takes the same form for approval seekers, but the arrows may point in any direction.

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.

The intuition behind these results is quite similar to that of Proposition 1, but the addition of a third option, when v_2 is at least as large as v_1, provides some low types with one or either of the norms with a defection option that is profitable without inducing too much extra guilt.

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

Proposition 6. *If $X = \{x_1, x_2, x_3\}$ with $v_3 > v_1$, $G(x_2 - x_1) = G(x_2 - x_3) = G_1$ and $G(x_1 - x_3) = G(x_3 - x_1) = G_2 > G_1$, and $\rho \in \{\rho_1, \rho_3\}$ 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 $s \to \infty$, all approval seekers pool on the same choice (x_1, x_3, or the compromise option x_2).*

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

4.3 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. Like the compromise option of the previous subsection, this turns out not to substantially change the basic results, but does lead to new insights.

I model simultaneous respect- and approval-seeking motivations in the natural manner:

$$U(x_i) = v(x_i) - tG(x_i - \rho_i) + s_{as}H_{as}(m_{as}(x_i)) + s_{rs}H_{rs}(m_{rs}m(x_i))$$

Additionally, assume the same binary setting as in section 3.

At higher levels of s_{as} the utility from approval dwarfs consumption utility, and imitation must occur in the direction of higher approval. At the same time, increasing s_{rs} reduces imitation, as people try to convincingly signal their integrity. The net effect may or 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 $v_2 > v_1$, $G(x) = G(-x)$, and ρ and t are independent, the following hold:*

1. *Low-t types choose x_2 for sufficiently small s_{as} but chooses $\arg\max H_{as}(m_{as}(x_i))$ for larger s_{as}.*

2. *At a fixed level of s_{as}, increasing s_{rs} in the limit leads to perfect adherence to personal norms.*

3. *At a fixed level of s_{rs}, increasing s_{as} in the limit leads to perfect conformity with majority opinion.*

4. *As s_{as} and s_{rs} simultaneously increase, in the limit, equilibrium can take* either *of the two forms above.*

One key difference arises from section 3.1: accepting social disapproval can *itself* be a signal of integrity.

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

Combining both types of social image into a single model clarifies the prescriptive discussion of the previous subsections. "Social pressure" is a general concept combining many things into a single parameter, and these components don't necessarily increase or decrease unilaterally. It may be possible to target approval-based social pressure separately from respect-based social pressure. On the other hand, those components of social pressure (anonymity, publicity) that cause people to care more about both approval and respect indicate that complete separation is probably impossible. Attempting to manipulate behavior with those aspects of social pressure may then be difficult process when respect and approval are countervailing forces. See section 5 for further discussion.

4.4 Additional notes

A few points on particular modeling details are collected here.

- Symmetric G: As noted in the model setup, symmetry in G is used purely to ease exposition. Many situations clearly violate this assumption (being overly generous likely induces less guilt than being selfish) but the analysis presented here straightforwardly extends to those scenarios.

- Also as noted in the model setup, focusing on the binary and ternary choice settings merely eases exposition. The same analytical approach can handle larger discrete sets of choices and norms (albeit tediously).

- Infinite support for t: For both approval seekers and respect seekers, since types with arbitrarily high integrity are assumed to exist in the population, no pooling equilibrium exists. This assumption slightly simplifies the analysis but isn't strictly necessary; in particular, appealing to the D1 criterion preserves the exact statement of Proposition 1 with $\operatorname{supp} t = [0, T]$ (see the Appendix).

- v is the same for all individuals: Note that if v is heterogeneous, the model still applies unproblematically so long as v is observable. Someone arguing in favor of vegetarianism will simply be interpreted differently if it's known that they don't like the taste of meat.

  Invisible heterogeneity in v complicates matters, for the model of respect seekers. If rank orderings of material outcomes are at least consistent, this simply renders individuals differentially responsive to social pressure, but limit results will still hold. If heterogeneity in v is additionally correlated with (t, ρ), these individuals will also be differentially responsive to their own guilt functions, effectively introducing correlation between t and ρ. But we can then appeal to the results with correlated parameters from section 4.1 to claim that limit case results will still hold. If not even rank orderings are at least "mostly" consistent across the population, inference becomes very difficult and the signaling model of respect likely pales in importance compared to straightforward approval.

- Other social preferences: As briefly alluded to, this analysis abstracts from other aspects of social preferences, such as altruism or warm glow. This is primarily because the model is intended to be portable across many situations which may or may not involve prosocial considerations. In particular cases, of course, these factors may be understood to influence the "consumption utility" of a particular choice, or the set of impartial norms that are actually held, or the guilt induced by a given choice.

- Identical audiences: The analysis considers a unified population in which each individual considers the same group of people the "relevant" audience. This is of course not always the case, but the mechanics of the model apply whatever the audience.

Further exploration of the very interesting alternative possibilities must be deferred to future work.

5 Discussion

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

Political Economy: Most obviously, the results are applicable to partisan politics and persuasion, which have been occasionally alluded to throughout the paper. Additionally, propositions 5 and 6 may specifically be relevant to political polarization. Polarization has been shown to be increasing in the United States, at least among political elites (McCarty et al., 2006), over at least the last half century. While attention has focused on potential demographic reasons for increasing polarization (for a review, see Layman, Carsey and Horowitz (2006)), this model points to social pressure as another possible explanation. If legislators are respect seekers and are appealing to the same constituency (such as federal or local level politicians), increasing polarization results from increasing social pressure. Or, if legislators are approval seekers, increasing polarization can result from *decreasing* social pressure.

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

explanations for polarization, but can be understood within these models of social pressure, perhaps with a similar mechanism as that proposed to explain the dramatic shift in support for marriage equality discussed in section 3.1.

Similarly, propositions 5 and 6 are clearly relevant to group decision making. In terms of voter mobilization, while it's almost universally agreed that you *should* vote, abstention represents a potential "compromise" option for those who don't want to take sides. A respect seeker might be tempted to abstain to avoid the hassle, but this action would never be socially rewarded. On the other hand, an approval seeker might be tempted to abstain if he doesn't want to be shunned by his friends of either political party. On aggregate, voter turnout could plummet. In fact, the HRC has admitted to trying to exploit this phenomenon strategically as they promote the idea that history has already decided in favor of marriage equality: "It just deflates them. People who may disagree with [gay marriage] but believe it may happen anyway are hard people to mobilize" (Issenberg, 2013).

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

While many individuals may agree with the alternative norms that are promoted by a development initiative, they are still beholden to the social image motivations that enforce the local norm. Understanding the mechanics of these motivations is critically important to designing effective, persuasive interventions. The above results show that if individuals are respect seekers and the desired action is costly, increasing visibility may encourage those

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

who agree to comply. This is the approach taken, for example, by Cameron, Gertler and Shah (2013), which attempts to reduce open defecation by making usage of the sanitation system visible. On the other hand, if individuals are approval seekers and the desirable norm has yet to take significant hold, or if individuals are respect seekers and the *un*desired action is costly, peer visibility and social pressure should be minimized. An initiative relying on social pressure to encourage good behavior may *backfire*.

The case of Kremer, Miguel and Thornton (2009) contains a convenient illustration of the relative effectiveness of an education initiative when there is more or less heterogeneity in norms. They analyze the effect of NGO scholarships offered to girls in two districts in Kenya. In the Busia district, 90% of teachers claimed that parents of students were positive towards the NGO, while in the Teso district this number was only 58%. There was therefore more pressure in the Teso district to drop out of the program or refuse the scholarship (as several schools, and one scholarship winner, did). In the end, not only did this high attrition rate jeopardize many students' chances at a scholarship, the impact on education attainment from the scholarship opportunity was much smaller in the Teso district. Part of this gap might be explained by respect-seeking students who were refusing a valuable program in order to signal dedication to their belief that outsiders are not to be trusted, or by approval-seeking students who caved to community pressure to resist outsiders.

If a beneficial norm has yet to take hold to *any* significant degree, despite seemingly significant costs to the individual, the development economist might also look to the results on sacrificial equilibria in section 3. Social pressure might be maintaining a destructive norm of bad sanitation, or violent civil conflict, or expensive religious sacrifice, or refusing contraception. And if social pressure is too high, it might be difficult to break the cycle. Take, for example, the norm of having expensive funerals. In several African countries, this has clearly reached the status of a destructive, welfare reducing norm; funeral costs are a major cause of families falling into poverty (Case et al., 2008; Krishna et al., 2004). If heterogeneous beliefs are being masked by high social pressure, reducing social pressure is unambiguously predicted to improve the situation.

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

Specific image targeting: It's unlikely that only respect-seeking or approval-seeking motivations are in effect in any particular setting, although the relative strengths surely vary. Political activists then may wish to refer to the model of section 4.3 by *separately*

targeting respect or approval. Take, for example, the efforts to enroll young people in health insurance under the new Affordable Care Act. The partisan reputation of the ACA may discourage young Republicans from enrolling, on principle, despite the fact that taking advantage of the new subsidies is clearly the profitable option for most of them. Democrats, on the other hand, may see political reputation and following their own beliefs as additional advantages to enrolling, on top of the actual subsidies. They are therefore in no danger of not enrolling. Since the concrete incentives already push young people in the direction of enrolling, Propositions 2 and 7 show that the respect-based social image of the choice should be as downplayed as possible, by downplaying its partisan nature or keeping enrollment decisions as low pressure and low visibility as possible. We can see this tactic in action in enrollment campaigns that emphasize state-specific names for ACA implementations and entirely fail to mention the association with "Obamacare".

Marketing: One can take a different view of the results in sections 3.1 and 2 as indicating to companies or groups how they should price products that are used to express identity or beliefs or to support moral beliefs, such as group membership fees, "Save the rainforest" products, brand name clothing, etc. Respect seekers seek opportunities to prove their commitment to their beliefs and are thus willing to pay for those products as a credible signal. Approval seekers, on the other hand, wish to follow their beliefs without being shunned and so will pay less for these products the higher is social pressure. A marketer may wish to appeal to social image differently to these two groups, or to adjust the price for a fixed number of items or membership slots in response to social image/pressure fluctuations.

6 Conclusion

In this paper, I developed a model of social image motivations that influence moral choices even 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 4.3. A likely possibility is that approval seeking is a more salient motivation when externalities of choices are large. On the other hand, Thomas Jefferson seems to prescribe approval-seeking and respect-seeking motivations to different classes of decisions when he said "In matters of taste, swim with the current; in matters of principle, stand like a rock." Characterizing the domains in which each model is applicable is an open empirical question and left for future work, but these models form an analytical foundation for beginning this research agenda.

References

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

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

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

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

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.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review*, 99(1): 544–555.

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

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

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

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2010. "Social Incentives in the Workplace." *Review of Economic Studies*, 77(2): 417–458.

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

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

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

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.

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.

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

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

Cameron, Lisa, Paul Gertler, and Manisha Shah. 2013. "The dirty business of open defecation : Lessons from a sanitation intervention."

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

Carpenter, Jeffrey Paul, and Caitlin Knowles Myers. 2010. "Why volunteer? Evidence on the role of altruism, image, and incentives." *Journal of Public Economics*, 94(11-12): 911–920.

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

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.

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

Dana, Jason, Daylian M. Cain, and Robyn M. Dawes. 2006. "What you dont know wont 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.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2013. "Voting to Tell Others." Working Paper.

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

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

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

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

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.

Glazer, Amihai, and Kai A. Konrad. 1996. "A signaling explanation for charity." *American Economic Review*, 86(4): 1019–1028.

Gneezy, Uri. 2005. "Deception: The role of consequences." *American Economic Review*, 95(1): 384–394.

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

Gul, Faruk, and Wolfgang Pesendorfer. 2006. "The canonical type space for interdependent preferences." Princeton University Working Paper.

Harbaugh, William T. 1998*a*. "The Prestige Motive for Making Charitable Transfers." *American Economic Review*, 88(2): 277–282.

Harbaugh, William T. 1998*b*. "What do donations buy? A model of philanthropy based on prestige and warm glow." *Journal of Public Economics*, 67(2): 269–284.

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/`.

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.

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

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

Krupka, Erin L., and Roberto A. Weber. 2013. "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11(3): 495–524.

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

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.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1(3): 593–622.

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

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

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

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/.

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

Ostrom, Elinor. 2000. "Collective action and the evolution of social norms." *The Journal of Economic Perspectives*, 14(3): 137–158.

Postlewaite, Andrew. 2011. "Social Norms and Social Assets." *Annual Review of Economics*, 3(1): 239–259.

Rabin, Matthew. 1995. "Moral preferences, moral constraints, and self-serving biases." Mimeo.

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

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.

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

te Velde, Vera L. 2014. "Beliefs-based altruism and motivated reasoning." Working Paper.

A Proofs

Throughout these proofs, for notational convenience, define $m_i = m(x_i)$ (or $m_{i,as}, m_{i,rs}$), $H_i = H(m_i)$ (or $H_{i,rs}, H_{i,as}$) and $v_i = v(x_i)$.

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 norm will defect to the other action, and all types with the other norm will adhere to their norm. For approval seekers, $H_1 = H_{as}(p_1)$ and $H_2 = H_{as}(1-p_1)$ 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, $H_{1,rs}$ and $H_{2,rs}$ 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 $H_{1,rs} > H_{2,rs}$. 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:

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

and

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

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 $\tilde{t}_1 > 0$.

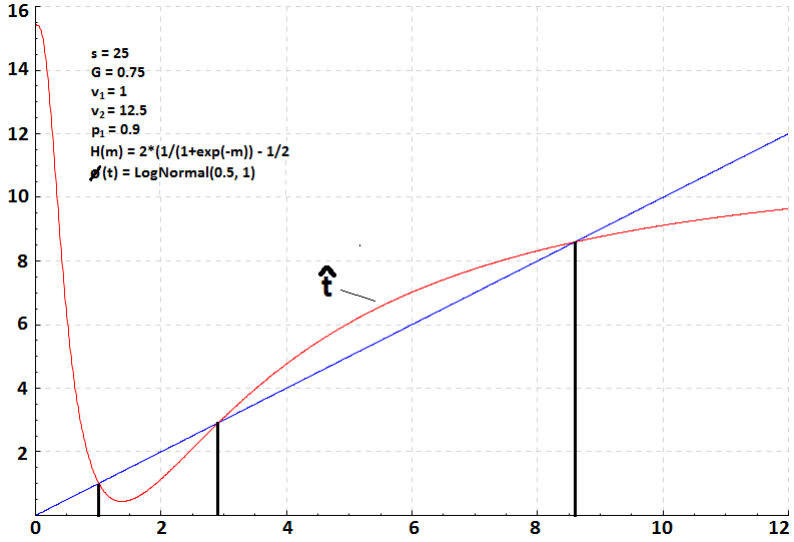
Define

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

This is a continuous and finite valued function, by assumption 1. At $t = 0$, $\hat{t} = (s(H_{rs}(\bar{t}) - H_{rs}(\bar{t})) + v_2 - v_1)/G > 0 = t$. 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.

Number of equilibria: Looking at Figure 4 again for reference, note that multiple equilibria can only exist if $\hat{t}$ achieves a slope of larger than 1. (This is of course a necessary, not sufficient, condition.) Applying Liebniz's rule, we find that $\frac{\delta m_1(t)}{\delta t} = \frac{\phi(t)}{1-\Phi(t)}(m_1(t) - t)$, $\frac{\delta m_2(t)}{\delta t} = \frac{p\phi(t)}{1-p+p\Phi(t)}(t - m_2(t))$, and therefore we can (conservatively)guarantee that there is

a unique equilibrium if $\frac{\delta \hat{t}}{\delta t} < 1$, or equivalently

$$H'(m_2(t)) \left(\frac{\phi(t)}{1 - \frac{1}{p} - \Phi(t)} \right) (m_2(t) - t) - H'(m_1(t)) \left(\frac{\phi(t)}{1 - \Phi(t)} \right) (m_1(t) - t) < G/s.$$

Note that $m_1(t) > t$ and H is increasing so the second term is positive. $1/p > 1$ so the first term can also be positive if $m_2(t) < t$. Since this is not always the case, we can't always guarantee uniqueness, as demonstrated by the example in Figure 4. But, the right side will be always larger than the left if 1) G is sufficiently large, 2) $\max \phi(t)$ is sufficiently small, 3) p_1 is sufficiently small, guaranteeing that $1 - \frac{1}{p}$ is highly negative, or 4) s is sufficiently small.

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

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

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

First suppose no one chooses x_1 in equilibrium. Type (t, ρ_2) might deviate to x_1 under mistaken beliefs $\tilde{m}_1$, resulting in mistaken beliefs about image utility $\tilde{H}_1$, if $s\tilde{H}_1 \geq v_2 - v_1 + sH_2 + tG$. Type (t, ρ_1) might deviate if $s\tilde{H}_1 > v_2 - v_1 + sH_2 - tG$. Clearly, if any type is willing to deviate for a given $\tilde{H}_1$, then the type (T, ρ_1) strictly prefers to deviate. This is therefore the only type that can be inferred after observing x_1, and $m(x_1)$ is required to be T.

On the other hand, $m(x_2) = \bar{t} < T = m(x_1)$. If $v_2 - v_1$ is large enough to overcome the image benefit of defecting, then, this pooling equilibrium is sustainable. This occurs when $v_2 + sH(\bar{t}) - TG \geq v_1 + sH(T) \iff T \leq \frac{v_2 - v_1 + s(H(\bar{t}) - H(T))}{G}$. But this is exactly the opposite

of the condition that guaranteed a separating equilibrium above. Therefore, if no separating equilibrium exists, there is a pooling equilibrium (and vice versa, guaranteeing existence of *some* equilibrium) in which all types choose x_2, in accordance with the statement of the proposition.

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

Proof of Proposition 2 part 1: 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 $H_{2,rs}$ and $H_{1,rs}$ is not sustainable in equilibrium. By Proposition 1 part 2, the only way for them to be equal is for $\tilde{t}_1 = \tilde{t}_2 = 0$.

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 $m_{1,rs}(\tilde{t}_1)$, $m_{2,rs}(\tilde{t}_1)$, 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 norm. The system of equations is:

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

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

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

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 higher average t) to seek a higher status by choosing against their norm. The costliness of c influences the exact cutoff values but doesn't change which action is chosen by impostors seeking higher status.

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 1: Define $\tilde{t}_{i,j,k}$ to be the cutoff type t above which someone with 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 $t > \tilde{t}_{1,1,3}$ choose x_1, with $\tilde{t}_{1,2,3} < t < \tilde{t}_{1,1,3}$ 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_2 choose x_3 if $t > \tilde{t}_{3,3,2} > 0$ and x_2 otherwise.

Type 4: $\tilde{t}_{3,2,1} > \tilde{t}_{3,3,1} > 0, \tilde{t}_{3,2,3}$. In this type, types with ρ_2 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 $t > \tilde{t}_{3,3,1}$ choose x_3, with $\tilde{t}_{3,2,1} < t < \tilde{t}_{3,3,1}$ 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_{rs}, 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_{rs}, $m_1 > m_3$, 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 $s\hat{H}_2 > v_1 - v_2 + sH_1 + tG$, 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 $\tilde{t}_{1,2,3} > 0 \leftrightarrow v_3 - v_2 + sH_3 - sH_2 > 0$. And by definition of m_{rs}, $m_1, m_3 > m_2$. The latter inequality is therefore always satisfied as $s \to \infty$. The former inequality requires both that $v_2 > v_1$ and $H_1 \to H_2$. But by definition of m_{rs}, 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:

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

$$\hat{t}_{1,1,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{s(H(m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})) - H(m_1(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}))) + v_2 - v_1}{G_2}$$

$$\hat{t}_{1,2,3}(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}) = \frac{s(H(m_3(t_{1,1,2}, t_{1,1,3}, t_{1,2,3})) - H(m_2(t_{1,1,2}, t_{1,1,3}, t_{1,2,3}))) + v_2 - v_1}{G_2 - G_1}$$

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

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

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:

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

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,

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

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

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 notice that

$$\begin{aligned}
& t_{1,1,2} > t_{1,1,3} \\
\iff & \frac{s(H_2 - H_1) + v_2 - v_1}{G_1} > \frac{s(H_3 - H_1) + v_3 - v_1}{G_2} \\
\iff & s\left(\frac{H_2}{G_1} - \frac{H_3}{G_2} - \frac{(G_2 - G_1)H_1}{G_1 G_2} \right) + \frac{v_2}{G_1} - \frac{v_3}{G_2} - \frac{(G_2 - G_1)v_1}{G_1 G_2} \\
\iff & s\left(\frac{H_2}{G_2 - G_1} - \frac{H_3}{G_2(G_2 - G_1)} - \frac{H_1}{G_2} \right) + \frac{v_2}{G_2 - G_1} - \frac{v_3}{G_2(G_2 - G_1)} - \frac{v_1}{G_2} > 0 \\
\iff & \frac{s(H_3 - H_1) + v_3 - v_1}{G_2} > \frac{s(H_3 - H_2) + v_3 - v_2}{G_2 - G_1} \\
\iff & t_{1,1,3} > t_{1,2,3}.
\end{aligned}$$

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 2: Any equilibrium must be of the form described in the proof of Proposition 5 part 1, as that argument does not depend on the definition of H_{rs} compared to H_{as}. 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 1.

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 $v(x_i) + s_{as}H_{as}(m_{as}(x_i))$. At small s_{as}, the first term dominates, and at higher s_{as}, the latter dominates.

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

Following the same intermediate value theorem argument of the proof of Proposition 1 part 2, with an appropriately modified definition of $\hat{t}$, shows that when s_{as} is large enough that $v_2 - v_1 + s_{as}(H_{as}(m_2) - H_{as}(m_1)) > 0$, $\hat{t}$ must fall within $[0, \infty]$. Along with the previous two parts, this proves part 4.