TI 2018-043/I
Tinbergen Institute Discussion Paper

On Esteem-Based Incentives

Ali Mazyaki[1]
Joel (J.J.) van der Weele[2]

1: Allameh Tabataba'i University
2: Universiteit van Amsterdam; Tinbergen Institute, The Netherlands

On esteem-based incentives

Ali Mazyaki[a][b] Joël van der Weele[c]

May 5, 2018

Abstract

Incentives based on esteem, honor and shame are increasingly popular and easy to use due to modern surveillance techniques. However, the use of shaming is controversial: critics argue that delegating punishment to a crowd can lead to mob justice and a loss of control over the size of the sanction. We use the signaling model of social behavior by Bénabou and Tirole (2011) to explore the effect of esteem-based incentives and their interaction with traditional monetary incentives. We show that esteem-based incentives can indeed lead to a loss of control by generating multiple equilibria, some of which feature high levels of stigma. Monetary and esteem incentives are interdependent. Moreover, if both types of incentives are costly to implement, the optimal incentive mix includes both instruments. In equilibrium, esteem-based incentives will be used relatively more for rare behaviors and in societies that have more heterogenous values.

Keywords: prosocial behavior, signaling, incentives, esteem.
JEL Codes: D02, H41, K42

[a]Department of Economics, Allameh Tabataba'i University, Shahid Beheshti, Tehran, Iran. Tel: +98(0)2188725400. Email: mazyaki@atu.ac.ir.

[b]Department of Economics, Institute for Management and Planning Studies, Niyavaran, Tehran, Iran. Tel: +98(0)2122802707. Email: a.mazyaki@imps.ac.ir.

[c]University of Amsterdam and Tinbergen Institute. Tel: +31(0)205254213. Email: vdweele@uva.nl.

The authors thank Zachary Grossman for useful comments and Ivar Kolvoort for research assistance.

1 Introduction

Esteem and stigma are important drivers of human behavior and provide powerful tools to affect behavior. With the advent of the internet, ever more personal information is available with the associated possibility to dispense shame and praise. When it comes to praise, managers use leaderboards and employee of the month schemes to reward productive employees, while governments use honorary medals to reward virtuous citizens. In the realm of shame, the pillory has been abandoned, but law-makers use reputation and shame by regulating the visibility of violations. One example is the decision when to expunge criminal records or make them available to employers and credit agencies. Another example is the discussion surrounding the public registration of sex offenders, for instance via Megan's law in the United States.[1] Moreover, the use of modern surveillance techniques allow far reaching incentive schemes based on reputation. An example is the social credit score system currently being tested in China, which gives each citizen a public score based on "good behavior".[2]

Legal scholars have vigorously debated the desirability of incentives based on esteem or reputation. Proponents, like Brennan and Pettit (2004) argue for increased used of esteem as it is a cheap and powerful incentive. Similarly, Kahan and Posner (1999: 366) argue that "Shaming [...] may offer a cost-effective and politically acceptable alternative to the short terms of imprisonment that such offenders now typically receive." By contrast, critics argue that shaming is a blunt and unpredictable instrument. By delegating punishment to the public, the effect of esteem sanctions is hard to control.[3] In an influential critique, Whitman (1998) writes:

> "There is no way to predict or control the way in which the public will deal with [an offender], no rhyme or limit to the terms the public may impose. Shame sanctions, in this regard, are very different from prisons or fines." (p. 1090-91).

The increasing availability of electronic surveillance, in combination with the possibility to share information on social platforms only makes these concerns more pertinent. Hess and

[1] Other examples abound. Kahan (1996: 635) document a rise in shaming sanctions in the U.S. for a variety of offenses, taking forms such as visible community service, rituals for disgracing the offender, or forced wearing of symbols publicizing their crime. Other examples include judges' decisions to make offenders place yard signs or ads in local newspapers announcing their crimes, or special license plates for drunk drivers. (See "Crime and Punishment: Shame Gains Popularity", by Jan Hoffman, NYT January 16, 1997.) More recently, Daughety and Reinganum (2010) provide a list of policy examples, such as the UC Berkeley Law Department rule to limit information on class rankings, the shaming of those who waste water during draughts in Georgia, and the shaming of speeders by by public display of license plates.

[2] See `http://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion` (accessed April 3rd, 2018).

[3] In addition, Nussbaum (2009) argues that shaming sanctions are cruel and illiberal. Kahan (2006) similarly retracts his earlier endorsement of shaming sanctions, on the ground that community based punishments have a populist underpinning that may undermine affirmation of individualistic values.

Waller (2014) and Ronson (2015) discuss the unpredictable nature of online trolling and mobbing, leading to outsized punishment for relatively small infractions.

Esteem and shaming have not received much attention from economic theorists. While there is a developing empirical literature on audience effects and image, until recently there has been little formal analysis of esteem incentives and their effects compared to traditional monetary incentives (see our overview in Section 2). Moreover, there have been few attempts to conceptualize the "loss of control" from shaming in economic theory or investigate the conditions under which it occurs.

In this paper, we analyze these questions in the context of a model of social norms, based on the model in Bénabou and Tirole (2011). In the model, an authority interacts with a continuum of agents, who decide whether to engage in an activity that is personally costly but generates positive externalities. In making their decision, agents are assumed to have three motivations. First, they care about the personal (monetary) payoffs from the action. Second, agents have different "values" or "intrinsic motivation" for pro-social behavior, and this preference type is assumed to be private information. Third, people care about their reputation or esteem, modeled as the expectation other people have about their values. This expectation is conditional on the observed action, giving rise to a possible incentive effect of esteem.

We assume that the authority has two policy instruments to increase pro-social behavior: a traditional monetary incentive and an 'esteem incentive'. The latter consists of an increase in the visibility of the prosocial action, thereby increasing the amount of esteem that agents can reap by being prosocial. We analyze the effect and the optimal choice of both incentives in the context of a perfect Bayesian, semi-separating Nash equilibrium, where the highest types behave pro-socially, and the lowest types do not.

We first derive conditions for the existence of a unique and stable equilibrium. We show that in contrast to monetary incentives, esteem incentives can indeed lead to a "loss of control" for the authority: A high level of such incentives can induce multiple equilibria, some of which are associated with high levels of stigma. Since the authority has no way to control which equilibrium will occur, it becomes impossible to predict the effect of high esteem incentives. Thus, the criticism on the unpredictability of shaming sanctions does indeed have a basis in economic theory, and we show under which conditions this loss of control occurs.

When it comes to the optimal use of both policies we show that if both incentives are costly to implement, the optimal policy features positive levels of both types of incentives. Esteem-based incentives are used relatively more for rare or exceptional behaviors: heroic actions that are too costly for all but the most social types, or decent actions that are taken by all but the

most selfish types. Thus, the model explains why medals and awards are typically given out only for heroic acts that few people perform. We also show that esteem incentives should be used relatively more in a society with more heterogenous values. These results increase our understanding of the optimal use of esteem while policy makers struggle to organize the ever increasing amount of personal information that is available online.

2 Esteem incentives in the economics literature

Economists have traditionally focused on monetary incentives as the main tool to influence behavior and reach policy goals. The analysis of esteem has been picked up only recently, both empirically and theoretically. On the empirical side, esteem from peers promote pro-social behavior both in the lab (e.g. Rege and Telle, 2004; Andreoni and Petrie, 2004; Andreoni and Bernheim, 2009; Ariely et al., 2009) and in the field (e.g. Harbaugh, 1998; Lacetera and Macis, 2010; Karlan and McConnell, 2014).

A recent strand of literature investigates the use of symbolic rewards as motivators in workplace settings. Neckermann and Frey (2013) show that providing the prospect of an award has significant effect on stated willingness to contribute to a public good, especially if accompanied by a public ceremony. Kosfeld and Neckermann (2011) show that awards have a considerable impact on work effort in a laboratory environment (see also Neckermann et al. 2011; Bradler et al. 2016). Markham et al. (2002) provides evidence that public recognition boosts attendance in a large manufacturing firm. In the management literature there is ample support for the idea that public recognition is a key motivator of employee performance (e.g. Holton et al. 2009). Ashraf and Bandiera (2018) discuss recent empirical literature showing the importance of the interaction of monetary and social incentives.

On the theoretical side, our paper is based on the model by Bénabou and Tirole (2011), which investigates how mechanisms of reputation or esteem influence optimal monetary incentives, both in the case of symmetric information about the distribution of values in society and in the case the authority has superior information. This fits in a wider literature that analyses the impact of visibility on the effectiveness of monetary incentives (e.g. Bénabou and Tirole 2006; Ariely *et al.* 2009; Bowles and Polania-Reyes 2012).

A few papers analyze the use of esteem incentives or variations in privacy in a signaling context. Bénabou and Tirole (2006) show that higher visibility of prosocial actions increases pro-social behavior by inducing low types to behave better. The policy is partially self-defeating however, since the additional compliance that occurs for reputation reasons weakens the signal of altruism sent by pro-social behavior. Daughety and Reinganum (2010) explore the tradeoffs between incentives provided by visibility and the conformism this induces on agent behavior,

which leads to possible over-investment in the public good. Jann and Schottmüller (2016) show that reduced privacy leads to impaired information aggregation. However, none of these studies address the potential loss of control from esteem-based sanctions, or the optimal joint level of the two incentives.

Like our paper, Ali and Bénabou (2016) also use a signaling model to study the "loss of control". In their model, preferences over moral behavior change over time. The effect of esteem incentives is to induce conformity and obscure changes in preferences, which introduces uncertainty about the optimal level of sanctions and the strength of disapproval generated by a given level of visibility. By contrast, in our setup preferences are fixed, and the unpredictability of the effect of esteem-based incentives arises because they can lead to multiple equilibria.

Finally, in the field of law and economics, there is a literature investigating the interaction between legal rules and informal norms cemented by mechanisms of esteem and reputation (for overviews, see e.g. Kahan 1997, Ellickson 1998, Posner 2000, and Van der Weele 2012a). Like the present study, Harel en Klement (2007) study the relationship between the use and intensity of stigma. They show that liberal use of stigma undermines its value as a deterrent as employers will start hiring stigmatized people. Dur and Van der Weele (2013) show that penalties raising the cost of particular criminal activities will change the signal associated with those activities, causing subtle spillover effects on other criminal activities. Several papers look at the impact of deterrence on stigma in a labor market context (e.g. Rasmusen 1996; Funk 2004), or discuss how reputation loss of convicts should affect awarded damages (e.g. Cooter and Porat 2001). However, none of these papers draws an explicit contrast between the use of different kind of incentives.

Iacobucci (2014) uses the logic of signaling to point out that an increase in legal sanctions may affect the esteem associated with breaking the law, an idea that also underlies the current paper. By using continuous types and continuous policy values, we are able to give a much more detailed analysis of the interaction between the two policies and their effects on aggregate behavior.

3 Model

The model in this section, as well as the notation, is almost identical to that by Bénabou and Tirole (2011, henceforth BT). Like BT, we frame the interpretation of the model in terms of pro-social behavior. Below we will provide an interpretation in terms of crime.

There is a large population of agents of measure 1, each of which decides whether to engage in a pro-social activity ($a_v = 1$) or not ($a_v = 0$). We consider two policy instruments that an authority can use to influence this decision. One instrument is a *monetary incentive* of

size $y \geq 0$, which takes the form of a reward for pro-social behavior. Note that we do not explicitly model any monitoring effort or uncertainty in getting the reward, so y can simply be interpreted as an expected reward. The other incentive is an *esteem incentive* of size $s \geq 0$, which influences the visibility of pro-social actions in the community. Thus, a higher s may reflect ceremonies, awards or honors to virtuous members or the shaming of offenders. A slightly different interpretation is that the government can influence the importance that people attach to their reputation. For example, the government could launch campaigns emphasizing the importance of good character (Kaplow and Shavell 2007).

Each agent has a preference type v, defined by the following utility function, which depends on her own action a_v and the actions of the other types a_{-v}:

$$U_v(a_v, a_{-v}) = \underbrace{[y - c]\, a_v + e\bar{a}}_{\text{Material}} + \underbrace{v a_v}_{\text{Intrinsic}} + \underbrace{s\mu E(v|a_v)}_{\text{Esteem}}. \tag{1}$$

Utility can be decomposed into three components. First, the material or monetary component consists of the reward y, minus some personal cost c associated with the pro-social act, and the fraction of pro-social people in in the population $\bar{a}$. Here, the parameter e measures the importance of the externality that agents impose on others by being pro-social. Second, the intrinsic component depends on the 'type' of the agent v. This type determines the 'intrinsic utility' that an agent gets when she acts pro-socially, and can be interpreted as the degree of 'altruism' of the agent. Agents are distributed over the type space according to the continuous cdf $F(v)$ with full support on $v \in [\underline{v}, \bar{v}]$. The final component of utility is 'reputation' or 'esteem-concern', or the inferred value of her type by other parties. This reputation is determined endogenously in a perfect Bayesian equilibrium, based on the action a_v that she has taken. The parameter $\mu > 0$ measures the importance of such reputation or esteem concern to the agents, which is multiplied by the visibility s. We model μ and s separately, to make the point that the authority does not have perfect control over the importance of esteem.

The model can easily be interpreted as a model of legal sanctions. In this case, one can interpret $a_v = 0$ as a criminal act, $a_v = 1$ as compliance with the law. Incentives are now applied to offenders instead of compliant agents, where y represents a fine that offenders pay to the authority, and s is the strength of a 'shaming sanction'.

In what follows we will first investigate the effect of two exogenously implemented policies in this model. In Section 7, we add a welfare-maximizing government to the model.

4 The esteem premium

In this paper, we will study perfect Bayesian Nash equilibria characterized by threshold type $v^* \in [\underline{v}, \bar{v}]$ such that types $v \geq v^*$ prefer behave prosocially, and types $v < v^*$ do not. In an equilibrium with threshold type v^*, the corresponding compliance level is $a^* = 1 - F(v^*)$. The behavior in this equilibrium is indeed optimal for each type if and only if the following equilibrium condition (EC) is satisfied

$$\begin{aligned} v^* + y - c + s\mu E\left[v \mid v > v^*\right] &= s\mu E\left[v \mid v < v^*\right] \\ v^* &= c - y - s\mu\Delta(v^*). \end{aligned} \tag{EC}$$

Here, $\Delta(v^*) := E\left[v \mid v > v^*\right] - E\left[v \mid v < v^*\right]$ is the difference between the expected type of those who behave prosocially and the expected type of those who do not. In the remainder, we will refer to $\Delta(v^*)$ as the gain in esteem that is associated with behaving pro-socially, or the *esteem premium*. The threshold type v^* is the type who is just indifferent between incurring the net cost of pro-sociality and incurring esteem $\Delta(v^*)$. From the EC it follows that the threshold v^* will be in the interior if $\underline{v} \leq c - y - s\Delta(\underline{v})$ and $c - y - s\Delta(\bar{v}) \leq \bar{v}$. If either of these two conditions is violated, this will result in a pooling equilibrium with either no compliance or full compliance, respectively. Existence of the threshold is guaranteed by the continuity of $\Delta(v^*)$, and the fact that $\Delta(v^*)$ is positive and bounded.[4]

The esteem premium $\Delta(v^*)$ thus measures the reputation incentives in equilibrium, and plays an important role in the analysis. In the context of crime, $\Delta(v^*)$ would represent the stigma associated with criminal behavior. Since $\Delta(v^*)$ depends on the shape of the distribution $F(v)$. Some results on the shape are provided by Jewitt (2004) and Adriani and Sonderegger (2015), a unimodal shape of $f(v^*)$ will result in a single-dipped esteem premium, with the minimum corresponding to the mode of $f(v^*)$.

Here, we will make the reasonable assumption that types are distributed according to a normal distribution, i.e. we assume that v is distributed $v \sim \mathcal{N}(0, \sigma_v^2)$ with support on $[\underline{v}, \bar{v}]$, with $\underline{v} = -\bar{v}$. Technically, our normal distribution is truncated, but we will assume that $F(\underline{v})$ is small, and abstract from it in our analysis.[5] The assumption of a truncated normal distribution

[4] Pooling equilibria may exist alongside the threshold equilibrium, where the former are supported by low off-equilibrium beliefs for either compliance or non-compliance. Pooling equilibria on non-compliance are not likely to survive standard equilibrium refinements like D1, as it is mostly in the interests of high types to comply. Pooling equilibria on full compliance may exist, but as Adriani and Sonderegger (2015) point out, forward induction arguments will favor the semi-separating equilibrium we study here.

[5] To have a density function on $[\underline{v}, \bar{v}]$, we have to normalize the density by dividing by $1 - 2F(\underline{v})$. Abstracting from this normalizations does not affect our qualitative results. Moreover, to have $\Delta(v^*)$ defined and continuous at the boundaries we define $\Delta(\bar{v}) := \lim_{v \to \bar{v}} \Delta(v)$ and $\Delta(\underline{v}) := \lim_{v \to \underline{v}} \Delta(v)$. By setting $E_v = 0$, the aggregate value of reputation is 0. Thus, increasing visibility does not affect the total amount of esteem, and we avoid the

is not necessary for most of our qualitative findings, as most single peaked symmetric functions will yield similar results. However, using qualitatively different type distributions will lead to different qualitative results. Adriani and Sonderegger (2015) provide an extensive discussion of the relation between the shape of the type distribution and the esteem function.

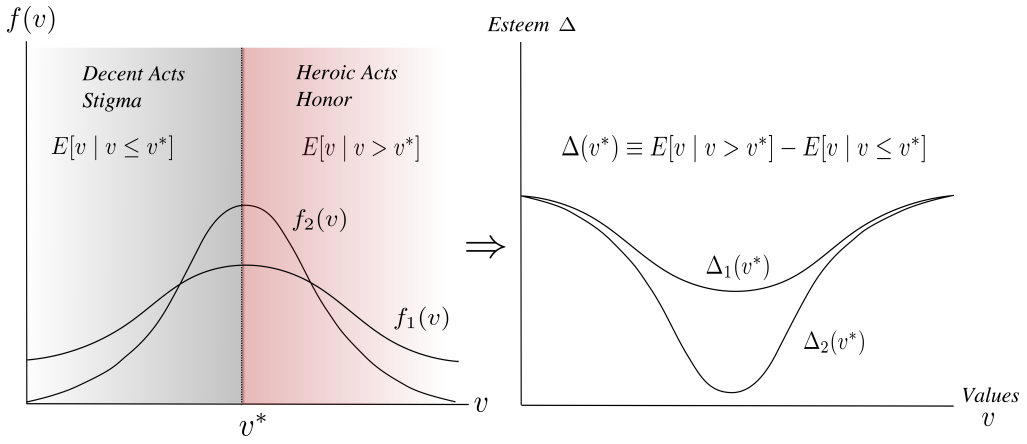


Figure 1: The esteem function $\Delta(v^*)$ of the truncated normal distribution. The function gets flatter if the variance σ_v^2 increases.

Under these assumptions, we can derive the following formula for the esteem function:

Lemma 1 *For the truncated normal distribution, the esteem function is given by*

$$\Delta(v^*) = \frac{\sigma_v^2 h(v^*)}{F(v^*)}, \tag{2}$$

where $h(v^*) := \frac{f(v^*)}{1-F(v^*)}$. This function is graphed in Figure 1. The higher the variance of the truncated normal, the flatter and higher the esteem function becomes, because for any intermediate levels of v^*, it is more likely that the type of the observed agent is somewhere in two tails. Note that in the extreme case, where the distribution is uniform, the esteem function is constant.

The shape of the esteem function reflects the change in esteem when the compliance level in society changes. When average prosocial behavior is low (v^* is high), prosocial acts are very informative as they signal an exceptionally high type. Below we refer to such as acts as "heroic". Similarly, antisocial actions generate strong negative esteem when v^* is low, as they signal a very low type. These are "decent" acts that most people would do. By contrast, when v^* is in an intermediate range, actions convey relatively little information and the esteem premium

question whether esteem is a good or bad thing in itself.

is low. Figure 1 also shows that these effects are muted when the tails of the distribution get thicker, as the esteem premium gets flatter if the variance σ_v^2 increases.

The equilibrium condition (EC) is depicted graphically in Figure 2. The straight, downward sloping line in Figure 2 is given by $\frac{c-v-y}{s\mu}$, that we will refer to as the EC-line. It represents all pairs $(v^*, \Delta(v^*))$ such that the threshold type is exactly indifferent between being pro-social or not, given the material costs c of contributing as well as the levels of y and s. Thus, equilibrium of the game is found on the intersection of this line with the $\Delta(v)$ curve. As we show below, policies y and s will determine compliance levels by shifting the EC line.

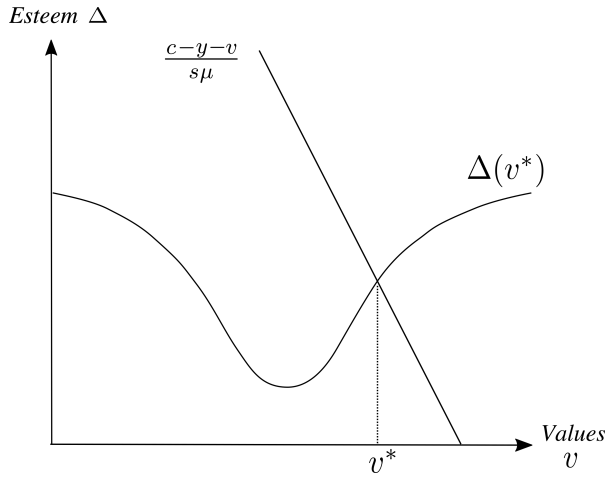


Figure 2: Equilibrium is found on the intersection of the (straight) EC-line and the esteem premium $\Delta(v^*)$.

5 Incentives and compliance in equilibrium

We will derive our results in two steps. In the first step we investigate the effect of incentives on the equilibrium level of prosocial behavior. This provides insight into how different incentives affect behavior, and whether they reinforce or dampen each other's effect. The second step is to derive the optimal government incentive, given the reaction of the agents. We focus on interior equilibria in which the first order conditions hold.

5.1 Monetary incentives and compliance

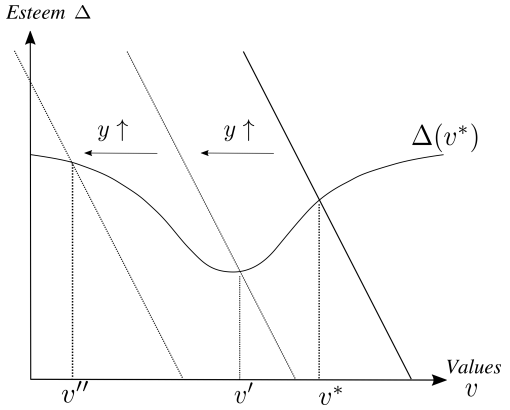
Proposition 1 replicates the result of Equation (6) in Bénabou and Tirole (2011: 7), and shows the effectiveness of monetary incentives on compliance.

Proposition 1 (Bénabou and Tirole, 2011) *In any interior equilibrium,*

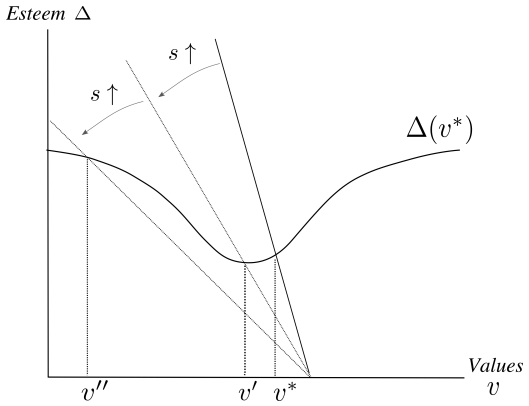
$$\frac{\partial a^*}{\partial y} = -f(v^*)\frac{\partial v^*}{\partial y} = \frac{f(v^*)}{1 + s^*\mu\Delta'(v^*)}. \tag{3}$$

If we look at the monetary incentive y, we see that the multiplier is given by $\frac{f(v^*)}{1+s\mu\Delta'(v^*)}$. This is equivalent to Equation (6) in Bénabou and Tirole (2011: 7). Thus, the effect of the monetary incentive depends on the slope of the esteem premium at the threshold, $\Delta'(v^*)$. If $\Delta'(v^*) > 0$, which is the case for low compliance rates (or high v^*), the effect of esteem counteracts the effect of the incentive. The reason is that for such "heroic" acts, the shift in behavior induced by monetary incentives "dilutes" the expected type of the small number of heroes more quickly than it dilutes the expected type the majority. Conversely, for "decent" pro-actions that most people do, esteem increases with the amount of pro-social behavior, as the expected type of the remaining deviants drops quickly with the threshold. The strength of these multiplier effects depend on the importance of stigma, which is determined by $s\mu$.

The left panel of Figure 3 demonstrates these results graphically. An increase in y shifts the EC line leftward. The shift in the equilibrium threshold depends on the slope of the esteem premium. A negative multiplier reduces the effects of a shift in the EC line that occurs on the downward-sloping part of the esteem function. The consequence is a modest increase in compliance from v^* to v'. An equivalent raise in y that occurs on the upward sloping part of the esteem function benefits from a positive multiplier, and hence produces a much larger shift in compliance from v' to v''.



(a) The monetary incentive shifts the EC curve to the left.



(b) The esteem incentive rotates the EC curve around its intersection with the x-axis.

Figure 3: Graphical illustration of the effect of the monetary incentive (left panel) and the esteem incentive (right panel) on the compliance level.

5.2 Esteem incentives and compliance

The next result shows the corresponding effect of esteem incentives on compliance.

Proposition 2 *In any interior equilibrium,*

$$\frac{\partial a^*}{\partial s} = -f(v^*)\frac{\partial v^*}{\partial s} = \frac{f(v^*)\mu\Delta(v^*)}{1 + s^*\mu\Delta'(v^*)}. \tag{4}$$

For the esteem incentive, the multiplier effect in the denominator is the same as for the monetary incentive. However, there is an additional (numerator) effect which depends on the *level* of esteem $\Delta(v^*)$. The higher the esteem premium, the higher the effect of raising the visibility of actions.

Both effects can be verified in the right panel of Figure 3, which graphically demonstrates the effect of the esteem incentive. An increase in s rotates the EC line inwards around the intersection with the v- axis. Increased visibility means that the esteem premium necessary to convince the agent to take the prosocial action is now lower for any level of intrinsic values v. Again, the shift in the equilibrium threshold depends on the slope of the esteem premium. A rotation in the EC line that occurs on the low and downward-sloping part of the esteem function produces only a modest increase in compliance from v^* to v'. An equivalent raise in y that occurs on the high and upward sloping part of the esteem function produces a much larger shift in compliance from v' to v''.

Thus, depending on the levels of μ and $\Delta(v^*)$, the esteem-based incentive can either be more or less effective than the monetary incentive. Generally, it will be most effective when esteem is high, i.e. either for heroic acts with high costs that few people do, or for decent acts with low costs that many people do.

5.3 Complementarities

Our previous results already showed that the two incentives are interdependent. We now investigate this interdependence in more detail, by looking at $\frac{\partial^2 v^*}{\partial y \partial s}$. If this expression is negative, then an increase in one of incentives makes the other incentive more effective in increasing compliance direction and we call the incentives "complements".

Proposition 3 *The two incentives are complements if and only if*

$$\Delta'(v^*) < -s\frac{\partial v^*}{\partial s}\Delta''(v^*). \tag{5}$$

The condition in Proposition 3 implies that complementarity depends on both the first and second derivative of the esteem premium. The reason the first derivative matters is that any

incentive that increases the esteem premium by decreasing v^* (i.e. $\Delta'(v^*) < 0$) will make the esteem incentive s more efficient. The reason the second derivative appears in (5) is that an incentive that decreases the *slope* of the esteem premium by decreasing v^* (i.e. $\Delta''(v^*) > 0$) will increase the social multiplier of both incentives, as shown in Proposition 3. Note that the size of the esteem incentive s will affect the size of the complementarity, but not the sign.

If we consider the esteem premium as shown in Figure 1, it is clear that it has one interior minimum, where $\Delta'(v^*)$ changes sign, and two inflection points where $\Delta''(v^*)$ changes sign. The area in which the two incentives reinforce each other is graphed as the shaded area in Figure 4. Either incentive increases the effect of the other incentive only when compliance is relatively high, i.e for actions where v^* is low, because in this case they make an antisocial act a more informative signal. By contrast, if compliance is low, the effect of increasing compliance is a decrease in the esteem-premium, which makes both incentives less effective.

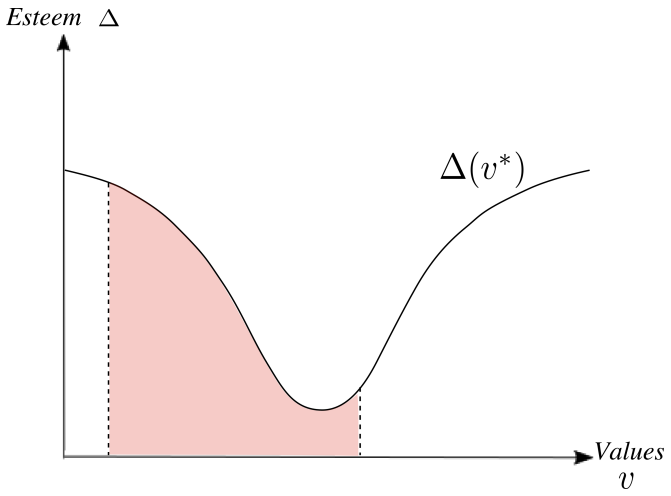


Figure 4: The two incentives reinforce each other only in the shaded region.

Summary 1 *The effect of both monetary and esteem incentives depends on the slope of the esteem premium, which depends on the amount of compliance. In addition, the effect of esteem incentives depends on the size of the esteem premium, which is higher for extreme actions. The two incentives are reinforcers if compliance is high, as the esteem premium increases in the level of compliance this domain.*

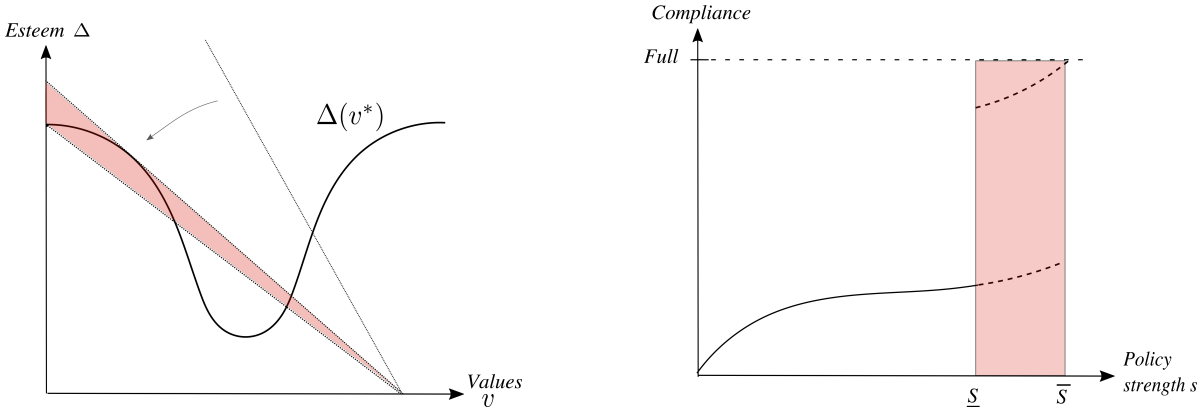
6 Multiple equilibria and loss of control

We now turn to the potential for a "loss of control" that various legal scholars have associated with esteem incentives. We operationalize the idea of "loss of control" by looking at

the existence of multiple interior equilibrium points v^*, which make the effect of incentives fundamentally unpredictable.

Before we develop our mathematical results, we illustrate the main ideas graphically in Figure 5. The left panel shows how an increase in s can lead to multiple equilibria. In the shaded zone, the EC line cuts the $\Delta(v^*)$ three times, twice from above and once from below. Of these three equilibrium points, only the most extreme ones are stable. The intermediate one is unstable, because a deviation by the threshold type will make a deviation optimal for other types, thus leading to an unraveling of the equilibrium.

Figure 5 elucidates the conditions for multiple equilibria to occur. First, the esteem premium must decline steeply for relatively high compliance levels, i.e. increase fast with compliance. Moreover, the EC line needs to be relatively flat, that is, the esteem incentive s needs to be high enough. If these conditions are satisfied, there is a possibility of co-existence of equilibria with either low levels of pro-sociality (high v^*) and a low esteem premium $\Delta(v^*)$, and equilibria with high levels of pro-sociality and a high esteem premium.



(a) Multiple equilibria when esteem incentives increase exposure.

(b) Compliance levels as a function of the level of the esteem incentive.

Figure 5: Graphical demonstration of the loss of control for high levels of esteem incentives.

Multiplicity of equilibria is an obvious problem for policy-making. Unless there are grounds to predict that agents can coordinate on a given equilibrium, welfare maximization becomes impossible. The right panel of Figure 5 illustrates this problem, by showing the compliance levels associated with the different equilibria depicted in the left panel as a function of s: for $s > \underline{s}$, the authority cannot predict the level of compliance and thus suffers a loss of control.

Figure 5 allows a few more observations. First, the occurrence of multiple equilibria necessitates relatively high levels of the esteem incentive, such that the EC line is relatively flat. When s is low and the EC line is near vertical, multiple equilibria cannot occur for any level

of the monetary incentive y. It is thus clear that a loss of control is indeed associated with high esteem incentives, in line with the intuition of legal scholars.

Second, the additional equilibria that appear when s increases have relatively high compliance levels and are associated with a high esteem premium. Thus, they can be considered a form of mobbing or "crowd justice", where increased visibility leads to high levels of shaming or stigma for a small minority, even for relatively minor offenses (for examples see Hess and Waller, 2014, and Ronson, 2015). While such strong applications of esteem generate high compliance, this is not necessarily efficient, as compliance may overshoot the optimal level (as we discuss in more detail below). Thus, these multiple equilibria reflect the concerns of critics like Whitman (1998, cited in the introduction) that outsourcing justice to the crowd may lead to unpredictable and heavy punishment.

Third, Figure 5 makes clear that when compliance levels are very low (heroic actions), there is not much risk of multiple equilibria. Thus, the use of rewards like medals or awards for very costly actions are unlikely to generate a loss of control.

Finally, note that an increase in σ_v^2 flattens the esteem premium, as Figure 1 shows. Thus, multiple equilibria are more likely to occur in populations with more homogeneous values. The intuition here is that when types are very concentrated, a switch by marginal types will have a relatively large impact on the expected types for each given action.

6.1 Uniqueness

We now investigate the conditions for multiple equilibria to occur more precisely. Lemma 2 characterizes a sufficient condition for uniqueness of an interior fixed point v^*, for any level of the policy variables s and y.

Lemma 2 *Let* $\hat{v} := \underset{v\in[\underline{v},\bar{v}]}{\arg\min}\, \Delta'(v)$. *If*

$$\Delta'(\hat{v}) > \frac{\Delta(\hat{v})}{\hat{v}-c}, \tag{6}$$

there is at most one internal equilibrium threshold $v^* \in (\underline{v}, \bar{v})$ *satisfying (EC), for all* $s, y \geq 0$.

According to this lemma, the equilibrium corresponding to any policy is unique if c is low enough and the esteem premium is "flat enough". Here, "flat enough" is measured by the derivative Δ' at the point $\hat{v}$, defined as the point where $\Delta'(\hat{v})$ is most negative on its domain and the esteem premium thus increases fastest with compliance. Putting a bound on $\Delta'(\hat{v})$ prevents the simultaneous existence of equilibria with low levels of pro-sociality (high v^*) and low esteem $\Delta(v^*)$, or with high levels of pro-sociality and a high esteem. If c is low, (6) is more

likely to hold, since the right hand side of (6) is now small or positive. The intuition is that a low cost increases compliance close to or beyond the point $\hat{v}$ where image changes the fastest, and around which multiple equilibria might occur. Thus, for high levels of s, any interior must be in a range where the esteem function is relatively flat.

6.2 Loss of control

When Lemma 2 is not satisfied, high levels of the esteem incentive may lead to multiple equilibria and a loss control for the authority, as illustrated graphically in Figure 5. The following formal result makes this intuition more precise, where again we use the definition $\hat{v} := \arg\min_{v \in [\underline{v}, \bar{v}]} \Delta'(v)$.

Proposition 4 *If*

$$\hat{v} - \frac{\Delta(\hat{v})}{\Delta'(\hat{v})} < c - y < \underline{v} - \frac{\Delta(\underline{v})}{\Delta'(\underline{v})}, \tag{7}$$

there exist $\underline{s} > 0$ and $\bar{s} > \underline{s}$, such that there exist multiple stable equilibrium thresholds if $s \in [\underline{s}, \bar{s}]$.

Both a graphical and mathematical proof are provided in the appendix. Proposition 4 shows that for multiple equilibria to occur, the cost of prosocial behavior $c - y$ can neither be too low nor too high. The first inequality of (7) is the converse of (6) if $y = 0$, showing that (6) is indeed necessary to guarantee uniqueness for all $y, s > 0$. It reflects that $c-y$ needs to be high enough, such that high levels of s imply compliance levels around $\hat{v}$ where multiple equilibria occur. As described above, a violation of this condition yields a single, unique equilibrium to the left of $\hat{v}$, or, if s and/or y are very high, a unique pooling equilibrium with compliance by all types.

The second inequality in Proposition 4 implies that multiple equilibria cannot occur if $c-y$ is extremely high. In this case, one needs very high levels of visibility s to generate compliance in the region around $\hat{v}$, where potential multiple equilibria occur. If the esteem curve decreases steeply enough close to $\underline{v}$, there is no stable interior equilibrium to the left of $\hat{v}$.[6]

Summary 2 *Esteem sanctions may lead to a loss of control for the authority for intermediate values of $c-y$. The additional equilibria are associated with high compliance and high levels of*

[6] In addition to the interior equilibrium to the right of $\hat{v}$, there may be an additional stable equilibrium featuring pooling by all types on compliance. While throughout this analysis we have focused on interior equilibria, one may of course characterize such a situation as one with multiple equilibria. In that sense, the second inequality in Proposition 4 can be considered a less stringent condition.

stigma. Thus, the model bears out the criticism associated with shaming sanctions, mentioned in the introduction.

7 Optimal incentives and welfare

In this section we consider the role of a government or an authority that maximizes welfare. While we will talk about the authority as a 'government', the model could also be applied to other (commercial) organizations, where the authority consists of (a team of) managers. Following BT with some small modifications, we assume that government maximizes the following social welfare function

$$W(y,s) = \bar{U} - \bar{a}\left(y + c_y(y) + c_s(s)\right) \tag{8}$$

where $\bar{U}$ is the average utility of the population, and the second term is a cost function. Costs are multiplied by $\bar{a}$, because incentives are only applied to those agents who behave pro-socially, i.e. choosing $a_v = 1$. Thus, the aggregate costs increase with the level of the incentive as well as the aggregate level of pro-social behavior. The linear cost of y reflects the transfer from the government to the agents. In addition the cost functions $c_y(y)$ and $c_s(s)$ reflect the cost of implementing policies. We do not make explicit assumptions on the cost functions $c_y(y)$ and $c_s(s)$, but we will argue below that an interior equilibrium is likely to exist only if both $c_y(y)$ and $c_s(s)$ are sufficiently convex. The costs of increasing the monetary incentive may be convex, as implementing higher incentive schemes may lead to increased efforts to conceal bad behavior and hence higher costs of implementing the sanction. Moreover, the distortionary losses from taxes necessary to finance these incentives are likely to increase with the level of incentives. For the case of esteem-based incentives, it is not clear whether there are returns to scale from increasing visibility. However, the political costs of reducing privacy are likely to be increasing in the amount of visibility.

We assume the government first sets incentive y and s, and then each agent chooses her action $a_v \in \{0,1\}$, after which payoffs are realized. When multiple equilibria occur, the maximization of the welfare function may not be possible without imposing further assumptions. In the following, we therefore assume that the uniqueness condition (6) holds. Thus, we only consider the case where the esteem premium Δ is relatively flat, and assume away any problems related to the loss of control discussed above.

To study the nature of this equilibrium, we first show the properties of the optimal incentives in the cases $c_y = c_s = 0$. In this case, it is straightforward to derive an explicit expression

for the optimal first-best incentive scheme (y^*, s^*) (see BT, Proposition 1):

$$y^* + s^* \mu \Delta(c - e) = e. \tag{9}$$

Equation (9) shows that when setting the optimal level of compliance, which is characterized by $v^* = c - e$, y^* and s^* can be traded off at rate $\mu\Delta$. Thus, if $\Delta(v^*)$ is high, i.e. only very few people comply or very few people do not comply, a small increase in s^* leads to a relatively large drop in y^*. By contrast, if $\Delta(v^*)$ is low, i.e. about half of the population complies, a small increase in s^* is compensated by only a small drop in y^*. The reason is that the social multiplier on s is much higher in the former case.

For the second-best case where implementing both incentives is costly, we are not able to derive explicit expressions for the individual level of the two incentives. The existence of an interior equilibrium requires that Δ is relatively flat, implying a rather stable marginal effect of s and y on compliance as shown by (3) and (4). Moreover, the cost functions $c_y(y)$ and $c_s(s)$ need to be sufficiently convex so that each policy is used in equilibrium.[7]

If these conditions are satisfied, the first-order conditions are sufficient and we can derive the following result.

Proposition 5 *In any interior equilibrium where the first order conditions are sufficient for a maximum, the optimal y and s of incentives satisfy*

$$\frac{c_s'(s^*)}{c_y'(y^*)} = \mu\Delta(v^*). \tag{10}$$

The expression in Proposition 5 equates the marginal benefits of each incentive to its marginal cost, where we know from Propositions (3) and (4) that the increase in pro-social behavior from a unit increase in s^* is $\mu\Delta(v^*)$ times the increase in pro-social behavior due to a unit increase in y^*.

Under the assumption that $c_y(y)$ and $c_s(s)$ are convex, Proposition 5 has several important implications. First, since $\Delta(v^*)$ is positive and bounded, the optimal incentive mix features positive levels of both incentives. Second, the use of the esteem incentive s is associated with 'extreme' levels of v^*, i.e. behaviors that either very few or very many people do.[8] The reason is that in this case the esteem premium for pro-social behavior is very high, and a change in

[7] One can show that some optimal policy exists by using versions of the extreme value theorem. However, the exact conditions for the existence of an interior equilibrium are implicit as they depend on the shape of Δ and the cost functions $c_y(y)$ and $c_s(s)$.

[8] This result depends crucially on the shape of the esteem function, and hence the distribution of types. The same qualitative results from any single-peaked function with declining density on either side of the mode, see Jewitt (2004) and Adriani and Sonderegger (2015).

visibility therefore has a large incentive effect. This rationalizes the use of award for heroic behaviors that only very few people do, and the use of shaming sanctions for very deviant acts.

Third, an exogenous increase in the importance of esteem μ implies a higher relative use of esteem-based incentives. While not surprising as a result, this implies that shaming sanctions are most effective in close-knit communities where news spreads fast and interactions are repeated. Furthermore, it gives support for an increased use of shaming sanctions for white-collar criminals (Skeel, 2001). White-collar workers such as businessmen have much more to lose from a ruined reputation, which is essential to secure business contacts.

Finally, an interesting corollary of Proposition 5 is that when values v become more heterogeneous the relative use of esteem-based sanctions increases. To make this more precise, we formally define heterogeneity.

Definition 1 *Society 1 is more heterogeneous than Society 2 if $F_1(v)$ second-order stochastically dominates $F_2(v)$.*

For the case of the normal distribution, this definition implies that an increase in σ^2 (while keeping the mean fixed), results in an increase in heterogeneity of a society. We can now make use of a result derived in Adriani and Sonderegger (2015, Lemma 3): if distribution $F_1(v)$ second-order stochastically dominates distribution $F_2(v)$ and the two distributions have an identical mean, then $\Delta_1(v) < \Delta_2(v)$ for all $v \in [\underline{v}, \bar{v}]$.

Corollary 1 *If Society 1 is more heterogeneous than Society 2, $s_1^* > s_2^*$ in any interior equilibrium.*

The intuition behind the result is simple: a more heterogeneous society will have higher levels of esteem $\Delta(v^*)$, as for any given v^*, the conditional expectations of types of compliant and non-compliant agents are further apart. Thus, an increase in the polarization of values, perhaps due to more ethnic diversity of cultural disagreements, implies an increased use of reputation and shame relative to other kinds of punishments.[9]

Summary 3 *If the first order conditions are sufficient and both policies are costly to implement, both esteem and monetary incentives are used in equilibrium. Esteem incentives are associated with more extreme behaviors and are used more when the distribution of values is more heterogeneous.*

[9]Note that we are assuming that while intrinsic values or motivations for a given action are further spread out when polarization increases, agents still agree on which actions are worthy of esteem.

8 Discussion and conclusion

In this paper we have considered a framework to analyze the deterrent effect of both monetary and esteem incentives. This framework rationalizes some existing intuitions about esteem incentives and provides a number of new insights. First, we show that the effect of esteem incentives and monetary incentives both depend on the compliance level. They reinforce each other when levels of prosocial behavior are relatively high, as in this case they make antisocial act a more informative signal.

Second, esteem incentives can indeed lead to a loss of control for the authority, as critics of such incentives have pointed out. Ramping up levels of visibility can lead to coexistence of equilibria with high levels of compliance and high levels of stigma for deviant actions, and equilibria with lower stigma and lower compliance. This captures incidents of "mob justice" or the "digital pillory", i.e. the unpredictable effect of online shaming.

Third, the model provides some support for the conjecture by Kosfeld and Neckermann (2011: 97), who write that "it is likely that social status and monetary aspects reinforce each other and that optimal incentives are based on the combination of social as well as monetary elements." While we show that the two instruments are complements only for relatively high levels of compliance, we show that under some regularity assumptions, both incentives are indeed used by the authority. More generally, our analysis shows the interdependence between the two kinds of incentives, which has been documented in recent empirical literature (Ashraf and Bandiera, 2018).

Fourth, we find that esteem-based incentives are associated in equilibrium with rare actions, for which esteem concerns are high. In this case, because extreme behaviors are committed by extreme types, reputation concerns are important and esteem-based incentives are particularly more effective. This is in accordance with observed real world practices. Medals of honor and awards are only given out for exceptionally virtuous behavior, not for wearing a seat-belt. Names and addresses of offenders are published only for extremely undesirable behavior like sex offenses against children, as several American States do under the so-called 'Megan's law', and the U.K. government (with some limitations) does under 'Sarah's law'.

Finally, an interesting implication of the model is that the heterogeneity of moral values in society matters for the optimal policy mix. Specifically, relatively homogeneous values imply a lower esteem premium, reducing the level of esteem incentives. Moreover, homogeneous values also make the esteem premium steeper on some parts of the domain and are more likely to lead to a loss of control.

The interplay between esteem and financial incentives offers much scope for further research. In this paper we have analyzed the deterrent effects of both incentive schemes. However,

as Kahan (1996) stresses, the expressive value of both types of sanctions may be equally important. To this end, one can analyze how the optimal policy mix is affected by information asymmetries between the authority and the agents about the distribution of values in society, as considered for example in Sliwka (2007), Bénabou and Tirole (2011), and Van der Weele (2012b). Moreover, the interplay between the two kinds of sanctions may be more intricate than we have assumed, as the size of the financial sanction may itself influence the visibility of an action and the amount of esteem associated with it. Another issue that would be interesting to incorporate in the model is recidivism. While reputational punishments may deter crime or anti-social behavior, it may also encourage recidivism by lowering the outside options of ex-offenders (Funk, 2004).

References

Adriani, Fabrizio and Silvia Sonderegger. 2015. "A theory of esteem-based peer pressure", mimeo, University of Leicester.

Ali, Nageeb S. and Roland Bénabou. 2016. Image Versus Information: Changing Societal Norms and Optimal Privacy (No. w22203). National Bureau of Economic Research.

Andreoni, James and Douglas B. Bernheim. 2009. "Social esteem 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: 1605-623.

Ashraf, Nava and Oriana Bandiera. 2018. "Social Incentives in Organizations?" Annual Review of Economics, In press.

Bénabou, Roland and Jean Tirole. 2006. "Incentives and Prosocial Behavior," *American Economic Review*, 96:5, 1652-78.

Bénabou, Roland and Jean Tirole. 2011. "Law and Norms", NBER working paper 17579.

Bowles, Samuel and Sandra Polania-Reyes, 2012. "Economic incentives and social preferences: substitutes or complements?", *Journal of Economic Literature*, 50:2, 368-425.

Bradler, Christiane, Robert Dur, Susanne Neckermann and Arjan Non. 2016. "Employee Recognition and Performance: A Field Experiment". *Management Science*, 62:11, 3085 - 3099.

Brennan, Geoffrey and Philip Pettit. 2004 *The economy of esteem: An Essay on Civil and Political Society*, Oxford University Press.

Cooter, Robert and Ariel Porat. 2001. "Should Courts Deduct Nonlegal Sanctions from Damages?" *Journal of Legal Studies*, 30: 401-22.

Daughety, Andrew F. and Jennifer F. Reinganum. 2010. "Public Goods, Social Pressure, and the Choice Between Privacy and Publicity", *American Economic Journal: Microeconomics*, 2, 191-221.

Jewitt, Ian. 2004. "Notes on the Shape of Distributions", unpublished manuscript.

Karlan, Dean and Margaret A. McConnell. 2014 "Hey look at me: The effect of giving circles on giving," *Journal of Economic Behavior and Organization*, 2014, 106, 402?412.

Dur, Robert and Joël J. van der Weele. 2013. "Status-Seeking in Criminal Subcultures and the Double Dividend of Zero-Tolerance," *Journal of Public Economic Theory*, 15:1, 77-93.

Ellickson, R. 1998. "Law and Economics Discovers Social Norms," *Journal of Legal Studies*, 27:2, 537-52.

Funk, Patricia. 2004. "On the Effective Use of Stigma as a Crime Deterrent," *European Economic Review*, 48(4): 715-728.

Harbaugh, William T. 1998. "What Do Donations Buy? A Model of Philanthropy Based on Prestige and Warm Glow," *Journal of Public Economics*, 67, 269?284.

Harel, Alon and Alon Klement. 2007. "The Economics of Stigma: Why More Detection of Crime May Result in Less Stigmatization", *Journal of Legal Studies*, 36:2, 355-77.

Hess, Kristy, and Lisa Waller. 2014. "The digital pillory: media shaming of ?ordinary' people for minor crimes," *Continuum: Journal of Media & Cultural Studies*, 28:1, 101-110.

Holton, Viki, Fiona Dent, and Jan Rabbetts. 2009. *Motivation and employee engagement in the 21st century: A survey of management views.* Ashridge.

Iacobucci, Edward M. 2014. "On the Interaction between Legal and Reputational Sanctions," *The Journal of Legal Studies*, 43:1, 189-207.

Jann, Ole and Christoph Schottmüller. 2016. "An informational theory of privacy", Mimeo, University of Copenhagen.

Kahan, Dan. 1996. "What Do Alternative Sanctions Mean?", *Chicago Law Review*, 63, 591-653.

Kahan, Dan. 1997. "Social Influence, Social Meaning and Deterrence", *Virginia Law Review*, 83:2, 349-95.

Kahan, Dan. 2006. "What's Really Wrong with Shaming Sanctions," *Texas Law Review*, 84: 2075.

Kahan, Dan M., and Eric A. Posner. 1999. "Shaming White-Collar Criminals: A Proposal for Reform of the Federal Sentencing Guidelines", *Journal of Law and Economics*, 42:365-91.

Kaplow, Louis and Steven Shavell. 2007. "Moral Rules, the Moral Sentiments and Behavior: Toward a Theory of an Optimal Moral System," *Journal of Political Economy*, 115:3, 494-514.

Kosfeld, Michael and Susanne S. Neckermann, 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance", *American Economic Journal: Microeconomics*, 3: 86-99.

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

Markham, Steven E., K. Dow Scott, and Gail H. McKee. 2002. "Recognizing Good Attendance: A Longitudinal, Quasi-Experimental Field Study", *Personnel Psychology*, 55:3, 639-60.

Neckermann, Susanne and Bruno Frey. 2013. "And the winner is ...? The motivating power of employee awards", *The Journal of Socio-Economics*, 46: 66-77.

Neckermann, Susanne, Reto Cueni and Bruno Frey. 2014. "Awards at Work." *Labour Economics* 31: 205-217.

Nussbaum, Martha C. 2009. *Hiding from humanity: Disgust, shame, and the law.* Princeton University Press.

Posner, Eric A. 2000. *Law and Social Norms.* Harvard University Press: Cambridge.

Rasmusen, Eric. 1996. "Stigma and Self-fulfilling Expectations of Criminality", *Journal of Law and Economics*, 39:2, 519-44.

Rege, Mari and Kjetil Telle, 2004. "The impact of social approval and framing on cooperation in public good situations," *Journal of Public Economics*, 88, 1625-1644.

Ronson, J., 2015. *So you've been publicly shamed*, Londen: Picador.

Skeel, David A. 2001. "Shaming in Corporate Law", *University of Pensylvania Law Review*, 149: 1811-1869.

Sliwka, Dirk, 2007. "Trust as a signal of a social norm and the hidden costs of incentive schemes", *American Economic Review*, 97:3, 999-1012.

Van der Weele, Joël J. 2012a. "Beyond the state of nature: introducing social interactions in the economic model of crime," *Review of Law and Economics*, 8:1, 401-32.

Van der Weele, Joël J. 2012b. "The Signaling Power of Sanctions in Social Dilemmas", *Journal of Law, Economics and Organization*, 28:1, 103-25.

Whitman, James Q. 1998. "What Is Wrong with Inflicting Shame Sanctions", *Yale Law Journal*, 107: 4, 1055-92.

Appendix with proofs

Proof of Lemma 1.
Suppose v has a truncated normal distribution $\mathcal{N}(0,\sigma_v^2)$ in $[\underline{v},\overline{v}]$ such that $\underline{v} = -\overline{v}$. The distribution is given by $f(v) = \frac{\alpha}{\sigma\sqrt{2\pi}} e^{-\frac{v^2}{2\sigma_v^2}}$ in which α is a multiplier that ensures the density to sum up to one.
It's first derivative is given by $f'(v) = -\frac{v}{\sigma_v^2} f(v)$, so we can write $\int_{\underline{v}}^{v^*} v f(v) dv = \int_{\underline{v}}^{v^*} -\sigma_v^2 f'(v) dv = \sigma_v(f(\underline{v}) - f(v^*))$.
This implies that $E(v|v < v^*) = \frac{\int_{\underline{v}}^{v^*} v f(v) dv}{F(v^*)} = \frac{\sigma_v^2(f(\underline{v}) - f(v^*))}{F(v^*)}$. Similarly, we can write $E(v|v > v^*) = \frac{\sigma_v^2(f(v^*) - f(\bar{v}))}{1 - F(v^*)}$. Using symmetry ($f(\underline{v}) = f(\bar{v})$) and the definition of $\Delta(v^*) := E(v|v > v^*) - E(v|v < v^*)$, we find:

$$\Delta(v^*) = \frac{\sigma_v^2(f(v^*) - f(\underline{v}))}{(1 - F(v^*))\, F(v^*)}.$$

We will approximate $f(\underline{v})$ to be zero, which will be true if $\underline{v}$ is small enough. Then, the first derivate can be written as

$$\Delta'(v^*) = -\frac{\Delta(v^*)}{\sigma_v^2} v^* + \frac{(\Delta(v^*))^2}{\sigma_v^2}(2F(v^*) - 1).$$

■

Proof of Proposition 3.
The two incentives are reinforcers if and only if $\frac{\partial^2 v^*(y,s)}{\partial y \partial s} < 0$.
Since $\frac{\partial^2 v^*(y,s)}{\partial y \partial s} = \mu \frac{\Delta'(v^*) + s\mu(\Delta'(v^*))^2 - s\mu\Delta(v^*)\Delta''(v^*)}{(s\mu\Delta'(v^*)+1)^2}$ the two incentives are reinforcers if and only if

$$\begin{aligned}
\Delta'(v^*) + s\mu\left(\Delta'(v^*)\right)^2 - s\mu\Delta(v^*)\Delta''(v^*) &< 0 \\
\Delta'(v^*)\left(-\frac{1}{\frac{\partial v^*}{\partial s}}\right) - s\Delta''(v^*) &< 0 \\
\Delta'(v^*) &< -s\frac{\partial v^*}{\partial s}\Delta''(v^*).
\end{aligned}$$

■

Proof of Lemma 2. We start with a few remarks. First, we define v^T as the threshold type associated with the internal equilibrium with the highest compliance level. Note that this equilibrium is always stable since $\Delta(\bar{v}) > 0$ and the EC line given by $\frac{-v-y+c}{s\mu}$ slopes downward. Thus, the EC line crosses $\Delta(v)$ from above, which is a sufficient and necessary condition for a stable equilibrium. Second, it is easy to verify that $\Delta'(\hat{v})$ is decreasing on $[\underline{v},\hat{v}]$ and $[-\hat{v},\bar{v}]$ and increasing on $[\hat{v},-\hat{v}]$. Third, our proof assumes $y = 0$. This is in fact a sufficient condition, as it follows from the EC condition that we can replace c by $c - y$ whenever $y > 0$, which will not cause a violation of (6).
We now show that (6) is sufficient for uniqueness. To rule out additional equilibria with compliance level below v^T, it is sufficient that the slope of $\Delta(v)$ is higher (less negative) on

$[\underline{v}, v^T]$ than the slope of the EC, where the latter is given by $\frac{\Delta(v^T)}{v^T - c}$. We now confirm that this is the case, where we restrict our analysis to $v^T < \hat{v} - \frac{\Delta(\hat{v})}{\Delta'(\hat{v})}$, since $v^T > \hat{v} - \frac{\Delta(\hat{v})}{\Delta'(\hat{v})}$ implies $c > \hat{v} - \frac{\Delta(\hat{v})}{\Delta'(\hat{v})}$, which violates (6). We distinguish two cases:

1. Suppose $v^T \in [\underline{v}, \hat{v}]$. Since $\Delta'(\hat{v})$ is decreasing on this interval, $\Delta'(\hat{v})$ is lower than the slope of the EC on $[\bar{v}, v^T]$. This rules out the existence of a second equilibrium to the left of v^T, and hence v^T is unique.

2. Suppose $v^T \in \left[\hat{v}, \hat{v} - \frac{\Delta(\hat{v})}{\Delta(\hat{v})}\right]$. In this case, a sufficient condition for uniqueness is

$$\Delta'(\hat{v}) > \frac{\Delta(v^T)}{v^T - c},$$
$$c < v^T - \frac{\Delta(v^T)}{\Delta'(\hat{v})}. \tag{A.11}$$

   For this case, (6) implies $c < \hat{v} - \frac{\Delta(\hat{v})}{\Delta'(\hat{v})}$. Thus (6) implies (A.11) if

$$\hat{v} - \frac{\Delta(\hat{v})}{\Delta'(\hat{v})} < v^T - \frac{\Delta(v^T)}{\Delta'(\hat{v})},$$
$$\Delta'(\hat{v}) < \frac{\Delta(\hat{v}) - \Delta(\ v^T)}{\hat{v} - v^T}. \tag{A.12}$$

   By the definition of $\hat{v}$, $\Delta'(\hat{v})$ is lower than $\Delta'(v)$ for all v, so (A.12) always holds. This insures that the slope of the EC is smaller than $\Delta'(\hat{v})$, and rules out the existence of a second equilibrium to the left of v^T. Hence, v^T is unique.

■

Proof of Proposition 4.

A graphical illustration of this proof is provided below. To establish multiple equilibria, it is sufficient to find an $\underline{s}$ and $v_1 \in (\underline{v}, \hat{v}]$ such that the EC line is exactly tangent to $\Delta(v)$ at $(v_1, \Delta(v_1))$. Since $\Delta(v)$ is concave on $[\underline{v}, \hat{v}]$ as we noted in Lemma 2, there exists an $\epsilon > 0$ that pivots the EC line downward such that $\underline{s} + \epsilon$ will result in intersections between the EC and $\Delta(v)$ on either side of v_1. The intersection with the threshold $v < v_1$ is associated with a stable equilibrium, while the intersection with the threshold $v > v_1$ is associated with an unstable equilibrium as the EC crosses $\Delta(v^*)$ from below. However, since $\Delta(\bar{v}) > 0$ and the EC slopes downward, there exists another stable equilibrium where the EC crosses $\Delta(v^*)$ from above, associated with a higher equilibrium threshold than that of the unstable equilibrium. This establishes multiplicity.

We now investigate the conditions for the existence of $\underline{s}$. First, consider an equilibrium $v^* = \underline{v}$. The EC line crosses $\Delta(v)$ from above at $(\underline{v}, \Delta(\underline{v}))$ if $\Delta'(\underline{v}) > -\frac{\Delta(\underline{v})}{c - y - \underline{v}}$ or

$$c - y < \underline{v} - \frac{\Delta(\underline{v})}{\Delta'(\underline{v})}. \tag{A.13}$$

Second, consider an equilibrium $v^* = \hat{v}$. The EC crosses $\Delta(\hat{v})$ from below at $(\hat{v}, \Delta(\hat{v}))$ if

$$c - y > \hat{v} - \frac{\Delta(\hat{v})}{\Delta'(\hat{v})}, \tag{A.14}$$

(see also the proof of Lemma 2).
The concavity of $\Delta(v)$ on $[\underline{v}, \hat{v}]$ implies that

$$\begin{aligned}
\frac{\Delta(\hat{v}) - \Delta(\underline{v})}{\hat{v} - \underline{v}} &< \Delta'(\underline{v}) \\
\frac{\Delta(\hat{v}) - \Delta(\underline{v})}{\Delta'(\underline{v})} &> \hat{v} - \underline{v} \\
\frac{\Delta(\hat{v})}{\Delta'(\hat{v})} - \frac{\Delta(\underline{v})}{\Delta'(\underline{v})} &> \hat{v} - \underline{v} \\
\underline{v} - \frac{\Delta(\underline{v})}{\Delta'(\underline{v})} &> \hat{v} - \frac{\Delta(\hat{v})}{\Delta'(\hat{v})}.
\end{aligned} \tag{A.15}$$

Thus, there exists an nonempty interval V_0, such that if $c - y \in V_0$, both (A.13) and (A.14) are satisfied.
Consider $c - y \in V_0$. Since the slope of the EC equals $-\frac{1}{\mu s}$ and is continuous on $(-\infty, 0)$, and since $\Delta'(\hat{v}) < \Delta'(\underline{v})$, there exists an $\underline{s}$ such that the EC is exactly tangent to $\Delta(v)$.
Finally, the strict concavity of $\Delta(v)$ implies that the width of V_0 is positive and bounded away from zero. Therefore we can find an $\bar{s} > \underline{s}$ and bounded away from $\underline{s}$ such that multiple equilibria exist for any $s \in [\underline{s}, \bar{s}]$.
The proof is illustrated in Figure A.1. If $c - y$ lies in the shaded area, multiple equilibria will occur for some values of s.

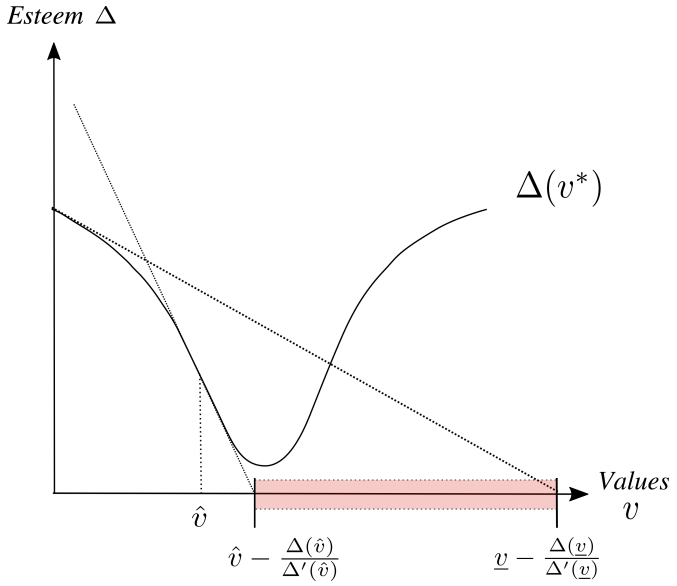


Figure A.1: When $c - y$ lies in the shaded area, multiple equilibria exists for some values of s.

■

Proof of Proposition 5. We denote the equilibrium by $v^* = v^*(y, s)$. Welfare maybe rewritten by

$$W(y, s, v^*) = \int_{v^*}^{\bar{v}} (v - c + e - c_y(y) - c_s(s)) f(v) dv. \tag{A.16}$$

We derive first order conditions of the welfare function w.r.t. s and y:

$$\frac{\partial W}{\partial y} = -\frac{\partial v^*}{\partial y} f(v^*) \left[e + v^* - c - c_y(y) - c_s(s)\right] - c_y'(y) \left[1 - F(v^*)\right] = 0 \tag{A.17}$$

$$\frac{\partial W}{\partial s} = -\frac{\partial v^*}{\partial s} f(v^*) \left[e + v^* - c - c_y(y) - s_y(s)\right] - c_s'(s) \left[1 - F(v^*)\right] = 0 \tag{A.18}$$

Combining (A.17) and (A.18), we obtain

$$\frac{\frac{\partial v^*}{\partial s}}{\frac{\partial v^*}{\partial y}} = \frac{c_y'(y^*)}{c_s'(s^*)}$$

Substituting the equations from Propositions 1 and 2, we obtain the result.

■