2020-14

Laws and Norms with (Un)Observable Actions

Claude Fluet
Murat C. Mungan

Octobre / October 2020

Centre de recherche sur les risques
les enjeux économiques et les politiques publiques

www.crrep.ca

ABSTRACT

We analyze the interactions between social norms, the prevalence of regulated acts, and policies. These interactions are impacted by people's inability to directly observe actors' behavior. Norms are ineffctive incentivizers when acts are committed either very frequently or very infrequently, because noisy signals of behavior are then too weak to alter people's beliefs about others' behavior. This cuts against the dynamics of the 'honor-stigma' model (Bénabou and Tirole 2006, 2011) and reverses its implications with even moderately noisy signals. With unobservable acts, the review process through which incentives are provided becomes an additional policy variable whose optima we characterize.

Keywords: Norms, social concerns, reputation, esteem, stigma, signaling, regulation.

Claude Fluet : Université Laval, CRREP, CRED.
Murat C. Mungan : George Mason University, Antonin Scalia Law School.

For valuable comments and suggestions, we thank Jesse Bull, Ezra Friedman, Andreea Cosnita-Langlais, Carlos Oyarzun, Jennifer Reinganum, Sarath Sanga, Abraham Wickelgren, and the participants of the 7th Annual Law and Economic Theory Conference, the 2017 Economics Seminar at UniversitÈ Paris Nanterre, the 2020 Soshnick Colloquium on Law and Economics at Northwestern Pritzker School of Law, and the 2020 University of Queensland Economic Theory Seminar.

Laws and Norms with (Un)Observable Actions[*]

Claude Fluet[†]
Université Laval
CRREP, CRED

Murat C. Mungan[‡]
George Mason University,
Antonin Scalia Law School

August 2020

Abstract

We analyze the interactions between social norms, the prevalence of regulated acts, and policies. These interactions are impacted by people's inability to directly observe actors' behavior. Norms are ineffective incentivizers when acts are committed either very frequently or very infrequently, because noisy signals of behavior are then too weak to alter people's beliefs about others' behavior. This cuts against the dynamics of the 'honor-stigma' model (Bénabou and Tirole 2006, 2011) and reverses its implications with even moderately noisy signals. With unobservable acts, the review process through which incentives are provided becomes an additional policy variable whose optima we characterize.

Keywords: Norms, social concerns, reputation, esteem, stigma, signaling, regulation.

1 Introduction

While making decisions, people are often motivated by how their behavior may impact their social standing. The decision to buy an expensive watch, to refrain from committing crimes, to donate money, to work hard, and many other decisions, are motivated by the effect they may have on one's reputation in addition to their immediate direct effect on material well-being. It is not surprising, therefore, that economists have formalized concepts like norms, reputation, esteem, and status through signaling models.[1]

[*] For valuable comments and suggestions, we thank Jesse Bull, Ezra Friedman, Andreea Cosnita-Langlais, Carlos Oyarzun, Jennifer Reinganum, Sarath Sanga, Abraham Wickelgren, and the participants of the 7th Annual Law and Economic Theory Conference, the 2017 Economics Seminar at Université Paris Nanterre, the 2020 Soshnick Colloquium on Law and Economics at Northwestern Pritzker School of Law, and the 2020 University of Queensland Economic Theory Seminar.

[†] Email: claude.fluet@fsa.ulaval.ca.

[‡] Email: mmungan@gmu.edu

[1] See Bernheim (1994), Ireland (1994), Glazer and Konrad (1996), Bénabou and Tirole (2006), Ellingsen and Johanneson (2008), among others.

An observation that emerged in this literature is that direct material incentives (whether provided through the law or private arrangements) and norm based reputational concerns will interact with each other. First, the size of direct rewards and punishments can influence the social prevalence of different types of acts. This affects social norms in the sense that it impacts the social-status gains or losses associated with the action being incentivized or discouraged; see Bénabou and Tirole (2006, 2010, 2011), Rasmusen (1996), Iacobucci (2014), Mazyaki and van der Weele (2019), Ali and Bénabou (2020). Secondly, as pointed out in earlier work, the presence of reputational incentives requires an adjustment of the optimal legal sanctions or rewards, e.g. a subsidy, tax or fine; see Cooter and Porat (2001). However, determining the optimal adjustments is not straightforward. This is because increasing direct material incentives may either increase or decrease reputational incentives, depending on whether behavior in the reputation game displays strategic complementarity or substitutability. Here, and in subsequent descriptions, we use the phrase "reputational incentives" loosely to refer to the expected change in one's social status or esteem that comes about from being perceived as having committed a 'bad' rather than a 'good' act, and which is a function of prevailing social norms. An issue at the heart of prior scholarship is whether formal incentives reinforce or mitigate reputational incentives, since the answer to this question is, a priori, ambiguous.

However, it has been noted that there is a predictable relationship between formal incentives and the equilibrium reputational incentives, when the propensities of individuals to engage in pro-social behavior are well-behaved, i.e. have a single-peaked distribution. Both the size of the reputational incentives and the way it responds to the prevalence of the act then depend in an intuitive way on how frequently the act is committed. Three types of equilibrium acts can be distinguished in such cases with reference to what is often referred to as the 'honor-stigma' model of Bénabou and Tirole (2006, 2011); see also Adriani and Sonderegger (2019).[2]

Frequently committed pro-social acts (corresponding to rarely committed bad acts) are considered 'normal' (e.g. not stealing) and their omission is associated with large reputational sanctions. The size of these sanctions increases further when bad acts become even more rare. The reputational sanctions are then mainly driven by the stigma associated with the bad act, rather than the honor associated with the good act, in the sense that committing the bad act signals a much greater negative deviation from the average person than the good act signals a positive deviation from the same. Similarly, large reputational rewards emerge for pro-social acts committed rarely, but this is driven by the great honor conveyed by the commission of the pro-social act, whereas the omission of the act does not signal a large negative deviation from the mean. Finally, there are 'modal acts' where the marginal actor's propensity is close to the modal actor's propensity. This last category of acts generates the lowest

[2] The following description of the three types of acts reproduces the discussion in Bénabou and Tirole (2011, p. 7). See also section 2.1 below.

reputational incentives.

These dynamics are quite appealing because they resonate with our intuition: we attach large stigma to rarely committed bad acts and we honor good deeds when only few people have the courage or willingness to engage in such behavior. However, these observations, and, more generally, the literature on the interaction between formal and reputational incentives, make predictions under the assumption that people's acts are perfectly observable. This contrasts with the assumptions employed in the moral-hazard context as well as the related literature on enforcement errors. In these contexts, the primary source of the principal-agent problem is the unobservability of the agent's actions.

In this article, we consider reputational incentives in a setting where an actor's behavior is unobservable by third parties, who need to rely on noisy signals of behavior to form beliefs about actors. We use this model to re-visit the findings of both the literature on the interaction between formal incentives and reputation, as well as the literature on enforcement errors. Our analysis reveals that the honor-stigma model's classification of acts in equilibrium does not hold when acts are unobservable. Moreover, except when the signals that third parties must rely on are extremely informative, the size of reputational incentives vis-à-vis the commonality of the act has the exact opposite properties as those that emerge in the honor-stigma model. Specifically, reputational effects are larger for acts which are committed with moderate frequency, and are smaller for acts committed frequently or rarely.

These results are driven by the dynamics of statistical inference. When acts are not directly observable, third parties have to rely on noisy proxies for a person's actions. In addition to this information, they also know the proportion of individuals who commit the act (as in the honor-stigma model) to make inferences regarding people's types. The inference process of third parties is a Bayesian one where they update their 'prior', i.e. the equilibrium proportion of people who commit the act, by using noisy information regarding actors' behavior. When the commission rate is extreme, priors are strong or very 'informed', so that the marginal impact of noisy information is limited. Conversely, when the commission rate is close to fifty percent, priors are 'uninformed' and the noisy signal becomes relatively important. Therefore, the statistical inference process generates dynamics which cut in the opposite direction as those that emerge due to the shape of the honor and stigma attached to being known as an actor versus non-actor, respectively. The two effects compete with each other. The statistical inference dynamics always dominate for extreme acts, i.e. those which are committed very frequently or very infrequently. This is because, for these acts, third parties' priors are so strong that they cannot be reversed by an imperfectly informative signal. As a result, even moderately noisy signals cause the statistical dynamics to overturn the honor-stigma dynamics previously described.

Figure 1 provides an illustration. We plot the magnitude of reputational incentives as a function of the commission rate of the pro-social act. The plot uses a symmetric unimodal beta-distribution for people's intrinsic propensities to behave pro-socially (see the modeling section for the details). The thick curve at the top of the figure represents reputational incentives with observable acts, as in the standard honor-stigma model. The thin curve depicts the reputational incentives with unobservable actions when third parties receive information through a binary signal characterized by symmetric type-1 and type-2 errors with a 0.05 probability.[3]

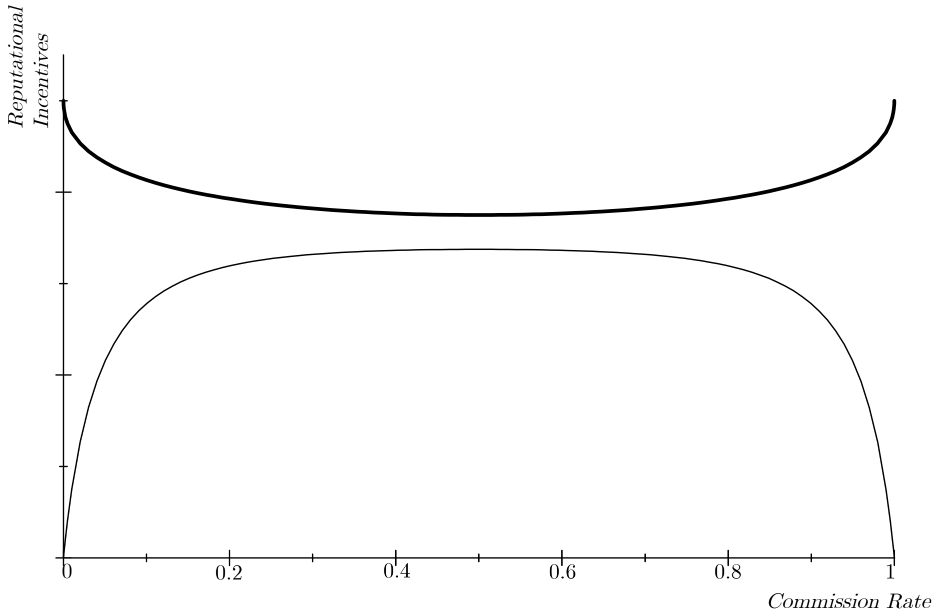

Fig. 1: Reputational incentives as a function of the prevalance of the pro-social act. Thick: Observable acts. Thin: Signals with 5% type-1 and type-2 errors.

A better appreciation of the two competing effects can be gained by walking through how reputational incentives are impacted by the participation rate with and without observable acts. First, consider the case where actions are observable and suppose material incentives are such that the equilibrium commission rate is 50%. When formal incentives are increased, the commission rate increases and reputational incentives increase because of strategic complementarity in the reputation game, i.e. reputational effects reinforce the effect of greater formal incentives. Conversely, when formal incentives are reduced, the commission rate decreases. Now, reputational incentives increase because of strategic substitutability in the reputation game, i.e. reputational effects will compensate to some extent for the smaller formal incentives. However, when actions are 'observable' with a small risk of mistake, such as the 0.05 probability of our example, the relationship between the frequency of the act in equilibrium and the magnitude of reputational incentives is completely reversed. Starting

[3] For instance, suppose the signal is *Pass* or *Fail.* The probability of generating *Pass* (resp. *Fail*) given the anti-social (resp. pro-social) action is then equal to 0.05.

from the 50% commission rate benchmark, greater formal incentives now crowd out reputational incentives; conversely, the effect of smaller formal incentives is reinforced by smaller reputational incentives.

Our observations thus far relate to the behavioral dynamics that emerge in a model with unobservable acts. We use these findings to asses normative implications. First, we ask how the optimal subsidy or tax must be altered when information is noisy. Trivially, the optimal Pigovian subsidy is larger than in the case where acts are observable. However, it also responds to changes in exogenous considerations (e.g. the resource cost of engaging in the pro-social act or greater intrinsic motivations to do so) in the opposite manner to that which would be predicted with observable acts. For instance, when acts are observable, the optimal subsidy is quasiconcave in the resource cost of engaging in the pro-social act and reaches a maximum at intermediate values of the resource cost. When acts are unobservable, the optimal subsidy is quasiconvex in the resource cost and reaches a minimum at intermediate values.

Subsequently, we analyze how the social planner can influence the magnitude of reputational sanctions, when he can trade off the type-1 and type-2 errors characterizing the binary signals received by third parties. Specifically, the social planner receives a continuous signal about the individuals' behavior and chooses a review standard that governs whether or not a person is rewarded. Third parties only observe the social planner's decision, i.e. whether or not a person is rewarded. This analysis is closely related to the economics literature on optimal standards of proof, which has been utilized to study law enforcement with imperfect monitoring.[4] We extend this literature by considering the impact of review standards on both formal and reputational incentives.

A preliminary result that follows from this analysis relates to the use of *symbolic rewards* wherein the good deeds of a person are acknowledged through the conferral of a reward whose monetary value is far smaller than the reputational benefits that it confers. Such rewards are used frequently in reality, with examples ranging from the Légion d'honneur to employee of the month awards. Our analysis provides a rationale for these rewards as well as an explanation of the circumstances under which their use is desirable. Specifically, we note that a degree of pro-social behavior can be incentivized exclusively through reputational concerns. Incentivizing further participation through the use of monetary incentives is costly. How these costs compare to the marginal benefits from increased participation naturally depend on the marginal cost of financing monetary incentives. When this cost is large, it becomes optimal to exclusively rely on reputational incentives, even when such incentives are insufficient to generate first best participation in pro-social acts. In these cases, the review standard used in allocating symbolic rewards is chosen solely for the purpose of maximizing reputational incentives.

In other cases where incentivizing pro-social behavior through the use of monetary incentives is cost justified, we find that the role of informal incentives

[4] See Rubinfeld and Sappington (1987), Demougin and Fluet (2006, 2008), Kaplow (2011), and Mungan (2020) among others.

in the determination of optimal policies depends crucially on the properties of the signal on which the review process is based and on whether the desired participation rate is extreme or modal. Moreover, when harnessing reputational concerns is an important consideration, the optimal policies will often be characterized by relatively weak review standards, implying large type-1 errors in erroneously rewarding agents.

Overall, our findings suggest that the unobservability of acts changes the dynamics in the honor-stigma model as well as its normative implications, and reveals additional insights when the review standard can be chosen optimally. In the next section, we present a simple model wherein agents commit acts, which generate noisy signals received by third parties. A special case of this model, where the signals are perfectly informative of agents' acts, reproduces Bénabou and Tirole's honor-stigma model. In section 3, we contrast the comparative statics with observable and unobservable acts and draw implications about the optimal Pigovian and Ramsey subsidies. In section 4, we endogenize the type-1 and type-2 errors and consider a model where third parties use a principal's decision on whether to reward an agent to update their beliefs about the agent's type. In section 5, we use this model to identify optimal policies consisting of a combination of a review process and a monetary incentive. We provide concluding remarks in section 6. All proofs are in the Appendix.

2 Model

To facilitate comparison, we follow the notation in Bénabou and Tirole (2011), henceforth BT, as closely as possible. Agents choose a discrete action $a \in \{0, 1\}$ where 1 is pro-social and 0 is anti-social in the sense that act 1 improves social welfare relative to act 0. Engaging in the pro-social act involves a resource cost of c for the agent and generates a benefit of e to be shared equally by all agents. In addition, the agent has intrinsic motivations, such that he receives an additional pay-off of v from engaging in the pro-social act. Agents differ from each other only with respect to their intrinsic motivation, so v is referred to as the agent's type. Types are distributed according to the cumulative function $G(v)$ with finite support $[v_{\min}, v_{\max}]$ and a continuously differentiable density $g(v) > 0$. The mean is denoted $\bar{v} \equiv E[v]$.

The principal incentivizes the agents by paying a bonus of y for engaging in the pro-social act, based on a decision rule which causes him to incorrectly offer a reward with probability α (type-1 error, since we consider these 'false positives') and to incorrectly refrain from rewarding an individual with probability $1 - \beta$ (type-2 error or 'false negative'), with $\beta > \alpha$. Thus, a person who engages in act $a = 1$ receives the bonus with probability β; if he does not, he receives the bonus with probability α. The bonus and the error rate pair can be endogenously determined through the choices of the principal. In this section and section 3, we take the error rates as given. We consider the case where they are endogenously determined in sections 4 and 5.

The final component which affects an agent's incentives reflects his reputational concerns. Third parties can make inferences regarding the agent's intrinsic motivation by observing whether or not he received a bonus. To incorporate these inferences we denote with $b \in \{0, 1\}$ whether the agent received a bonus, where $b = 1$ indicates receipt. Thus, $E[v|b]$ denotes third parties' expectations of the agent's type, conditional on whether or not he received a bonus. From the agent's perspective, engaging in act a causes this estimate to have an expectation of $E[E[v|b]|a]$. When $a = 1$, this expectation is

$$\beta E[v|b = 1] + (1 - \beta)E[v|b = 0]$$

When $a = 0$, the expectation is

$$\alpha E[v|b = 1] + (1 - \alpha)E[v|b = 0]$$

Below, we explain how these expected values are related to the principal's and other agents' actions.

Given this set-up, the agent's preferences are represented as follows:

$$U = (v + \beta y - c)a + \alpha y(1 - a) + e\bar{a} + \mu E[E[v|b]|a] \tag{1}$$

where $\bar{a}$ is the proportion of agents engaging in the pro-social action and μ is a positive parameter denoting the importance of reputation relative to other considerations (i.e. v, y and c). The agent's utility function in BT emerges as a specific case of (1) when $\beta = 1$ and $\alpha = 0$.

2.1 Reputational Incentives

As a first step, we derive explicit expressions for reputational gains or losses. From (1), the utility from engaging in the pro-social act is increasing in v. Thus, in equilibrium, individuals with intrinsic valuations above a threshold commit the act, and those with lower valuations do not. In deriving specific expressions, we therefore restrict our attention to behavior profiles where agents choose $a = 1$ if, and only if, their intrinsic value exceeds some threshold v^*.

The measure of individuals who receive a bonus is

$$\psi(v^*, \alpha, \beta) \equiv \alpha G(v^*) + \beta(1 - G(v^*)) \tag{2}$$

and $1 - \psi$ agents do not receive bonuses. Denote the conditional expectations below and above the cut-off by

$$\mathcal{M}^+(v^*) \equiv E[v|v > v^*] \text{ and } \mathcal{M}^-(v^*) \equiv E[v|v < v^*] \tag{3}$$

Using Bayes' rule, we then obtain $E[v|b = 1]$, which we denote as H, referring to the fact that the person is *honored* through the receipt of a bonus. This yields:

$$H(v^*, \alpha, \beta) \equiv \frac{\alpha G(v^*)\mathcal{M}^-(v^*) + \beta(1 - G(v^*))\mathcal{M}^+(v^*)}{\psi(v^*, \alpha, \beta)} \tag{4}$$

Similarly, we can calculate $E[v|b=0]$, which we denote as S, referring to the fact that the person has been *stigmatized* by not receiving a bonus:

$$S(v^*, \alpha, \beta) = \frac{(1-\alpha)G(v^*)\mathcal{M}^-(v^*) + (1-\beta)(1-G(v^*))\mathcal{M}^+(v^*)}{1-\psi(v^*, \alpha, \beta)} \tag{5}$$

Finally, because they play key roles in the analysis, we define both the difference between H and S and the difference between $\mathcal{M}^+(v^*)$ and $\mathcal{M}^-(v^*)$ as follows:

$$\Delta(v^*) \equiv \mathcal{M}^+(v^*) - \mathcal{M}^-(v^*) \tag{6}$$

$$\begin{aligned} \Lambda(v^*, \alpha, \beta) &\equiv H(v^*, \alpha, \beta) - S(v^*, \alpha, \beta) \\ &= \frac{(\beta-\alpha)(1-G(v^*))G(v^*)}{\psi(v^*, \alpha, \beta)(1-\psi(v^*, \alpha, \beta))}\Delta(v^*) \end{aligned} \tag{7}$$

We note that Λ multiplied by the importance of reputation, μ, is the reputational gain associated with receiving a bonus versus not receiving one. The honor, stigma, and expected reputational benefits considered in BT emerge in the special case where actions are perfectly observable, since

$$H(v^*, 0, 1) = \mathcal{M}^+(v^*);\ S(v^*, 0, 1) = \mathcal{M}^-(v^*);\ \text{and}\ \Lambda(v^*, 0, 1) = \Delta(v^*) \tag{8}$$

As shown in the literature (Jewitt 2004, Bénabou and Tirole 2006, 2011), when the distribution of types is strictly unimodal with an interior maximum, Δ is quasiconvex with a unique interior minimum.[5] Based on these observations Bénabou and Tirole draw some important implications regarding the relationship between how frequently acts are committed and reputational incentives. They introduce the following categorization, which we reproduce verbatim:

> “For concreteness, we shall refer to the “desired” behavior $a = 1$ as being (in equilibrium):
>
> –“Respectable” or “normal”, if v^* is in the lower tail, for instance because the cost c is low. These are things that “everyone but the worst people do”, such as not abusing one’s spouse and children, and which are consequently normative, in the usual sense that the pressure to conform rises with their prevalence.
>
> –“Admirable”or “heroic”, if v^* is in the upper tail, for instance because the cost c is very high. These are actions that “only the best do”, such as donating a kidney to a stranger or risking one’s life to rescue others.
>
> –“Modal” if v^* in the middle range around the minimum of Δ: Both $a = 1$ and $a = 0$ are then common behaviors, leading to weak inferences about agent’s types.” (Bénabou and Tirole 2011, p. 7)

[5] An example is the thick curve of Fig. 1 in the introduction.

However, when third parties cannot directly observe actions, they must rely on whether the agent received a bonus to make inferences. Since the conferral of bonuses is subject to error, they do not perfectly indicate whether the recipient has intrinsic motivations above or below the equilibrium cut-off. When a large fraction of agents participate in the pro-social action, conferral of a bonus will tend to be a frequent event. However, as suggested in the introduction, it no longer follows that non-conferral of a bonus then imposes a large stigma. Similarly, when few agents participate in the pro-social action, bonuses will be infrequent but it no longer follows that receipt of a bonus confers much honor.

To disentangle the different effects, we rewrite (7) as

$$\Lambda(v^*, \alpha, \beta) = \delta(v^*, \alpha, \beta)\Delta(v^*)$$

where

$$\delta(v^*, \alpha, \beta) \equiv \frac{(\beta - \alpha)(1 - G(v^*))G(v^*)}{\psi(v^*, \alpha, \beta)(1 - \psi(v^*, \alpha, \beta))} \tag{9}$$

The preceding expression is a measure of the predictive value of bonuses regarding the agents' behavior. This depends on their discriminatory value, as captured by the type-1 and 2 errors, and on the prevalence of the pro-social action. The predictive value is between zero and unity, with $\delta(v^*, 0, 1) = 1$ when conferral and non-conferral of a bonus are perfectly informative about the agent's action. Equation (9) can also be expressed as

$$\delta(v^*, \alpha, \beta) = P(a = 1|b = 1) - P(a = 1|b = 0) \tag{10}$$

In this form, δ is a well-known metric of the degree of associative relationship between pairs of events, and has the following properties.[6]

Lemma 1 *Let $\alpha > 0$ and $\beta < 1$. Then $\delta(v^*, \alpha, \beta) < 1$ and is quasiconcave in v^* with a strict interior maximum, with $\delta(v^*, \alpha, \beta) = 0$ when v^* equals $v_{\min}$ or $v_{\max}$.*

It follows that, when the distribution of types is unimodal with an interior mode, the reputational gain as defined in (7) is the product of two functions with opposite properties. The following proposition summarizes the implications.

Proposition 1 *Suppose the distribution of types is strictly unimodal with an interior mode.*
(i) When acts are observable (i.e. $\alpha = 1 - \beta = 0$), the reputational benefit equals the difference in the average intrinsic value of actors and non-actors, i.e. $\Lambda = \Delta$, and is quasiconvex in v^* with a unique interior minimum.
(ii) When acts are unobservable and bonuses serve as noisy signals of acts (i.e. $\alpha > 0$ and $\beta < 1$), the reputational benefit $\Lambda < \Delta$ and never has an interior minimum. When $\beta - \alpha$ is not too large, Λ is quasiconcave in v^* with a unique interior maximum.

[6] In experimental psychology, it is known as Delta P owing to the difference in probability in (10). See Powers (2011).

The preceding results reveal the contrast between the way reputational benefits interact with the prevalence of the act when acts are observable and unobservable. The qualitative relationship with reputational benefits described in the prior literature emerges as a knife-edge case, when $\alpha = 0$ and $\beta = 1$. Small errors cause the relationship to vanish. Moreover, as noted in part (ii) of the proposition, the opposite of the interactions described in the literature is obtained when bonuses are sufficiently noisy signals of the agents' behavior, expressed in terms of the probability differential $\beta - \alpha$. Note that this is a standard measure of accuracy in dichotomous discrimination tests.[7]

The proposition naturally raises the question of what happens for a reasonable range of non-extreme v^* and for reasonably small type-1 and type-2 errors. Answering this question generally is difficult. We provide numerical examples illustrating how the introduction of small errors is sufficient to modify the qualitative nature of the interactions between reputational costs and the prevalence of the pro-social act. We plot the Λ and Δ functions when types follow the Beta distribution $\mathcal{B}(w, z)$ with parameters $w, z \geq 1$. We consider only three illustrative cases: the symmetric case ($w = z = 2$), the skewed case ($w = 2 < z = 4$), and the uniform distribution ($w = z = 1$). The probability density functions are represented in Figure 2. The uniform distribution is useful to illustrate the statistical inference dynamics in the absence of the competing honor-stigma dynamics, since Δ is then a constant unaffected by the prevalence of the pro-social act.

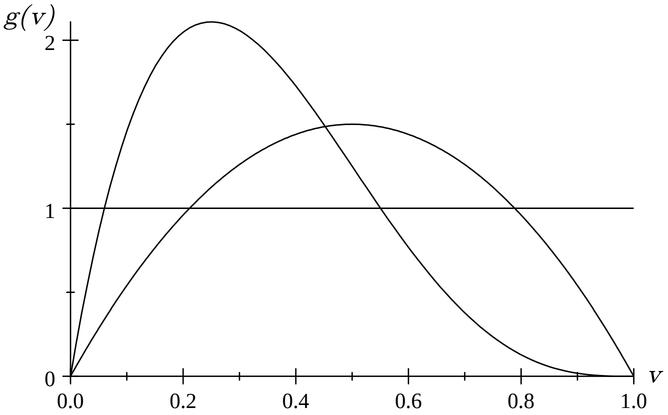

Fig. 2: Three beta-distriburions (symmetric, skewed, and uniform).

The Figures 3a to 3c plot $\Delta(v^*)$ and $\Lambda(v^*)$ where the latter is computed with symmetric type-1 and 2 errors, i.e. $\alpha = 1-\beta = \varepsilon$, where $\varepsilon \in \{0, 0.01, 0.05, 0.1\}$.[8] These values are chosen because they either correspond to standard significance

[7] It is often referred to as Youden's index, in reference to Youden (1950). It can also be written as $1 - (\alpha + 1 - \beta)$, i.e. as one minus the sum of type-1 and 2 errors.

[8] Note that we now follow the literature by drawing the reputational benefit with respect to v^*, by contrast with Figure 1 in the introduction where it is drawn with respect to the participation rate $1 - G(v^*)$.

levels or would be considered acceptable in a dichotomous diagnostic test. The exercise reveals two insights. First, even small error rates cause reputational benefits to exhibit the same behavior near the boundary of the type interval. Second, an error rate of 5% may be enough to cause the reputational benefit to be quasiconcave.

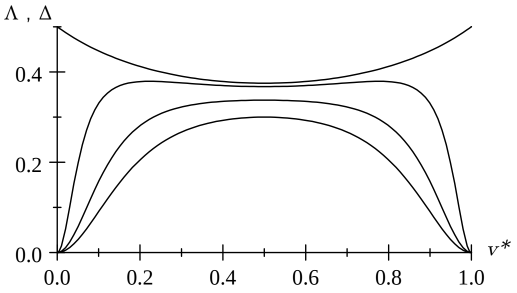

Fig. 3a: Symmetric uni-modal $w = z = 2$.

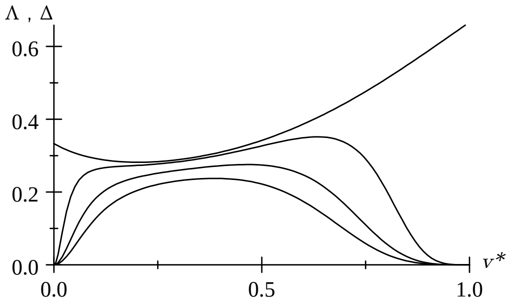

Fig. 3b: Skewed uni-modal $w = 2$ and $z = 4$.

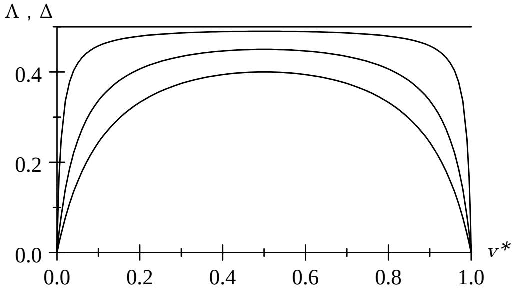

Fig. 3c: Uniform $w = z = 1$.

2.2 Equilibrium and the Social Multiplier

To conclude the section, we use the agents' expected utility (1) to characterize the equilibrium threshold and investigate how incentives are shaped by the social prevalence of the anti- and pro-social acts. We start with the best response of an agent to the other agents' behavior profile as summarized by v^*. Thus, we express the agent's utility as $U = U(a, v^*, v)$ and note that he weakly prefers the pro-social act if, and only if, $U(1, v^*, v) \geq U(0, v^*, v)$, which corresponds to the condition:

$$v + (\beta - \alpha)(y + \mu\Lambda(v^*, \alpha, \beta)) - c \geq 0 \tag{11}$$

Engaging in the pro-social act instead of the anti-social act increases the probability of receiving a bonus from α to β. In (11), the probability differential multiplies the bonus as well as the reputational gain $\mu\Lambda$.

Given (11), a perfect Bayesian equilibrium is characterized by a cut-off intrinsic value v^* solving

$$\varphi(v, y, \alpha, \beta) \equiv v + (\beta - \alpha)(y + \mu\Lambda(v, \alpha, \beta)) - c = 0 \tag{12}$$

The existence of an interior solution $v^* \in (v_{\min}, v_{\max})$ is ensured by the condition

$$\varphi(v_{\min}, y, \alpha, \beta) < 0 < \varphi(v_{\max}, y, \alpha, \beta) \tag{13}$$

When $\alpha > 0$ and $\beta < 1$, the reputational gain vanishes when v^* approaches the boundaries of the support of types, so that (13) reduces to

$$v_{\min} < c - (\beta - \alpha)y < v_{\max} \tag{14}$$

We assume this condition holds.[9]

When acts are perfectly observable, a unique equilibrium obtains if the reputation parameter μ is not too large, which ensures that φ is everywhere increasing in v. With positive type-1 and 2 errors, the same upper bound on μ may not be sufficient to guarantee that (12) has a unique solution. We select the equilibrium as follows:

$$v^* = \min\{v : \varphi(v, y, \alpha, \beta) \geq 0 \text{ and } \varphi_v(v, y, \alpha, \beta) > 0\} \tag{15}$$

where φ_v denotes a partial derivative. When (12) has more than one solution, (15) selects the 'stable' equilibrium with the highest participation rate in the pro-social activity. Given (13), a solution satisfying (15) always exists.

An important question is whether the incentives to engage in pro-social acts increase or decrease as more people commit it. Totally differentiating (12), one obtains

$$-\frac{\partial v^*}{\partial y} = \frac{\beta - \alpha}{1 + (\beta - \alpha)\mu\Lambda_v(v^*, \alpha, \beta)} \tag{16}$$

where Λ_v denotes a partial derivative. A unit increase in the bonus y increases by $\beta - \alpha$ the expected material reward from engaging in the pro-social act. Taking others' behavior as given, an individual's best-response threshold would therefore decrease by the same amount. However, taking social interactions into account, the total effect on everyone's behavior is (minus) the change in expected reward multiplied by

$$M(v^*, \alpha, \beta) = \frac{1}{1 + (\beta - \alpha)\mu\Lambda_v(v^*, \alpha, \beta)} \tag{17}$$

[9] When acts are perfectly observable (see BT, footnote 12), condition (13) reduces to

$$v_{\min} + \mu(\overline{v} - v_{\min}) < c - y < v_{\max} + \mu(v_{\max} - \overline{v})$$

In our subsequent analysis, the bonus (and the type-1 and 2 errors in section 5) will be optimally chosen, so that (13) will hold at the optimal solution obviating the need for these assumptions.

which henceforth will be referred to as the social multiplier.[10]

The denominator equals $\varphi_v(v^*, y, \alpha, \beta)$, so the preceding expressions are always positive reflecting the fact that at a stable equilibrium an increase in monetary incentives always increases participation. When agents have no reputational concerns (i.e. $\mu = 0$), the multiplier equals one. In the presence of reputational concerns, the multiplier is larger (resp. smaller) than one if $\Lambda_v < 0$ (resp. $\Lambda_v > 0$). Thus, for any given participation rate, whether reputational concerns and formal incentives reinforce each other or mitigate each other's effectiveness depends entirely on how reputational concerns change with the participation rate, i.e. the shape of Λ_v.

Proposition 1, in turn, notes how these interactions depend on whether or not acts are observable. Figures 3a-c depict, for instance, a wide range of high participation rates (i.e. small values of v^*) for which $\Lambda_v > 0$ when acts are unobservable, but where $\Lambda_v < 0$ when they are observable. In these examples, for large participation rates the social multiplier is smaller than what would emerge absent reputational concerns when acts are observable, and the opposite is true when acts are unobservable. Similar observations can be made regarding small participation rates. This naturally causes assumptions regarding act observability to play a pivotal role in how formal incentives ought to be adjusted in the presence of reputational concerns. We analyze this issue, next.

3 Optimal Rewards and Comparative Statics

How formal incentives affect behavior depends crucially on the observability of agents' actions as revealed by our discussion of the social multiplier. This has important policy implications, which we illustrate here by comparing optimal policies as well as how they respond to changes in the environment when acts are observable versus unobservable. This section also introduces concepts which are useful in section 5 where the principal optimally chooses the review standard (i.e. the α, β pair).

3.1 Social Welfare and Optimal Rewards

To identify the optimal bonus, we consider a welfare function defined as the sum of agents' equilibrium expected utilities net of the cost of financing the bonuses. This is expressed as

$$W = \bar{U} - (1 + \lambda)y\psi \tag{18}$$

where $\bar{U}$ is the sum of the utilities defined in (1) and $1 + \lambda$ is the marginal cost of a dollar to be used as incentives, with $\lambda \geq 0$ denoting the shadow cost of public funds. Substituting (1) into (18),

$$W = \int_{v^*}^{v_{\max}} (v + e - c)g(v)dv - \lambda y\psi(v^*) + \mu\bar{v} \tag{19}$$

[10] To quote Scheinkman (2006, p. 12554): "The *social multiplier* measures the ratio of the effect on the average action caused by a change in a parameter to the effect on the average action that would occur if individual agents ignored the change in actions of their peers."

The first term is the net direct benefit from the pro-social act. The second term is the deadweight loss of financing bonuses. The last term is the average reputational benefit, which is independent of v^* because reputational gains and losses cancel out when summed over all individuals.[11] The equilibrium threshold satisfies (12), i.e. $v^* = v^*(y)$.

In the special case where rewarding agents involves no social cost, $\lambda = 0$ and welfare is maximized if agents participate in the pro-social act whenever $v + e \geq c$ and otherwise abstain, i.e. the sum of the private and social benefit must exceed the cost. The first-best threshold is therefore

$$v_{FB} = c - e \tag{20}$$

We make the following assumptions, as in BT, to ensure a well behaved maximization problem.

Assumption 1 $v_{\min} < v_{FB} < v_{\max}$ and $e > \max_v \mu\Delta(v)$.

In the first-best, some agents participate in the pro-social act but not all of them. Moreover, when acts are perfectly observable, positive bonuses are required to implement the first best, i.e. relying solely on reputational incentives is not sufficient. Because $\Lambda(v) \leq \Delta(v)$, the same is also true when acts are unobservable. We maintain assumption 1 in the remainder of our analysis.

Assuming λ is not too large, an interior solution $y > 0$ obtains satisfying the first-order condition:

$$\frac{1}{g(v^*(y))}\frac{\partial W}{\partial y} = [e + v^*(y) - c - \lambda(\beta - \alpha)y]\left(\frac{-\partial v^*(y)}{\partial y}\right) - \lambda\rho(v^*(y)) = 0 \tag{21}$$

where

$$\rho(v) \equiv \frac{\psi(v)}{g(v)} = \frac{\alpha G(v) + \beta[1 - G(v)]}{g(v)} \tag{22}$$

is a weighted sum of the reciprocals of hazard rates[12] and tracks the size of the population receiving bonuses relative to the measure of the individuals who are on the margin. The first-order condition is normalized by the density $g(v^*)$ to ease the interpretation of various components, including ρ.

The condition in (21) is similar to that obtained in many contexts with Ramsey taxation: incentives must be raised to the point where the social benefits from additional participation equal the marginal social cost of financing incentives. The first term in the middle expression contains all benefits and costs associated with changes in the participation rate, following a marginal increase in the bonus. The second term is the extra deadweight loss of increasing the bonuses paid to inframarginal agents (those with $v > v^*(y)$), as well as the bonuses erroneously paid to non-participating agents (those with $v < v^*(y)$).

[11] This follows from the assumption that reputational utility is linear in the posterior beliefs of third parties. See Deffains and Fluet (2020) for a framework where reputational gains and losses do not cancel out.

[12] The hazard rate is $g/(1 - G)$ and the so-called reverse hazard rate is g/G.

When $\lambda = 0$, the second term vanishes and the optimal bonus, or Pigovian subsidy, solves $v^*(y) = c - e = v_{FB}$. Absent reputational considerations, it would be optimal to provide an expected bonus $(\beta - \alpha)y$ equal to the externality e, i.e. the bonus would equal $e/(\beta - \alpha)$. However, when agents have reputational concerns, substituting (20) into (12) yields

$$y = \frac{e}{\beta - \alpha} - \mu\Lambda(v_{FB}, \alpha, \beta) \tag{23}$$

where $\mu\Lambda(v_{FB}, \alpha, \beta)$ is the reputational benefits one obtains upon receiving a bonus. Thus, the optimal subsidy depends on the reputational benefits associated with the first-best participation rate, which in turn depends on the observability of acts. Since $v_{FB} = c - e$, a comparison between both the size of Pigovian subsidies and how they change with v_{FB} can be illustrated by plotting the subsidies, as in figure 4, as a function of the resource cost c when acts are observable (*thick curve*) and when they are not (*thin curve*).[13]

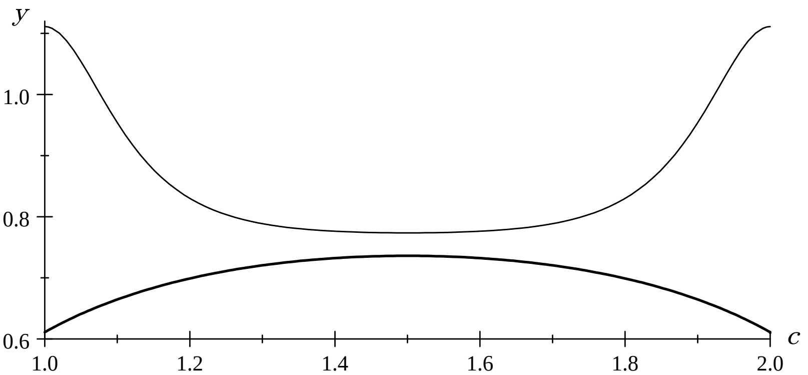

Fig 4: Pigovian Subsidies with obervable acts and with unobservable acts with $\beta = 0.95$, $\alpha = 0.05$.

Obviously, the optimal subsidy is greater when acts are unobservable. Moreover, it moves in opposite directions in response to a change in c when acts are observable and unobservable, respectively. These observations illustrate which policy implications discussed in the literature ought to be adjusted when acts are unobservable versus observable. For instance, the conclusion that tax reductions due to charitable givings ought to be lower than the standard Pigovian subsidy, and that the gap between the optimal tax reduction and the standard Pigovian subsidy ought to be decreasing in the importance of reputational considerations (i.e. μ) remain intact. However, the relationship between optimal subsidies for using harm-reducing technologies whose cost of adoption is declining over time must be revisited. With observable actions, the optimal subsidy

[13] Figure 4 depicts the case where $\mu = e = 1$ and $G = \mathcal{B}(2, 2)$.

for the adoption of these technologies ought to first increase (assuming a low initial participation rate) and subsequently fall.[14] If the adoption of the technology is inferred subject to sufficient error, then the optimal subsidy should first fall, and then rise over time, as depicted in figure 4.

When financing bonuses is socially costly, the analysis is not as straightforward. This is because the presence of the second term in (21) causes the optimal participation rate to be lower than the first-best rate and to be responsive to other considerations. We discuss these in the context of exogenous changes in the agents' intrinsic motivations.

3.2 Shifts in Norms

Here we study how the equilibrium behavior of agents and optimal rewards respond to changes in people's attitudes towards the pro-social act and explain how these responses depend on whether their acts are observed with or without errors. To enable these comparative statics, we consider uniform shifts in the distribution of types. A rightward shift means that individuals are on average more intrinsically motivated. Because dispersion between types does not change, such shifts have a simple interpretation in terms of exogenous changes in norms.[15] First, we consider the impacts of such shifts on the equilibrium behavior of agents, for a given bonus. Subsequently, we consider the case where the bonus is chosen optimally.

In both cases, let $G(v-\theta)$ be the original distribution of types shifted to the right when θ is positive, so that the support is then $[v_{\min}+\theta, v_{\max}+\theta]$. It is easily seen that the reputational benefits are given by the same functions as in the preceding section, but taking the shift into account. Specifically, the benefits are $\Delta(v^*-\theta)$ or $\Lambda(v^*-\theta)$, where in the latter we omit reference to the type-1 and 2 errors.

An implication of this observation is that a rightward shift θ in intrinsic motivations has the same effect on the equilibrium reputational benefits as an identical increase in the expected bonus, i.e. a bonus increase equal to $\frac{\theta}{\beta-\alpha}$. The shift in the distribution of motivations causes people to infer the same honor-stigma differential from an equilibrium threshold of $v^*+\theta$ as they used to infer from the equilibrium threshold of v^* prior to the change. In other words, the shift changes the *supply of reputational benefits,* and this is reflected by the fact that reputational benefits are given by $\Lambda(v^*-\theta)$. The same impact can be generated by an increase in bonuses, which causes more people to participate in the pro-social act, inducing third parties to adjust their inferences regarding the expected type of actors. We demonstrate this insight graphically via figures 5a and 5b for the cases where acts are unobservable and observable, respectively.

[14] Both of these policies are discussed in BT, p. 10.

[15] See BT and in particular Adriani and Sonderegger (2019) who also consider other changes in the distribution of types.

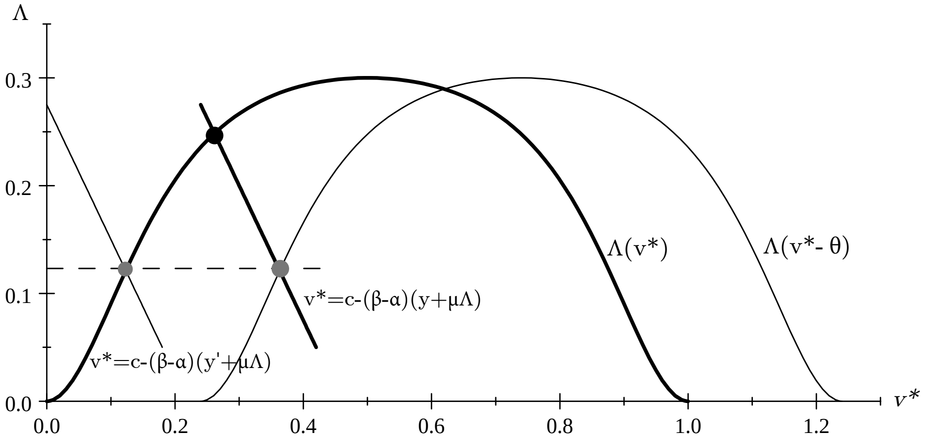

Fig. 5.a: Unobservable acts. Changes in eq. threshold and reputation caused by shifts in norms and a monetary bonus that generates equivalent changes

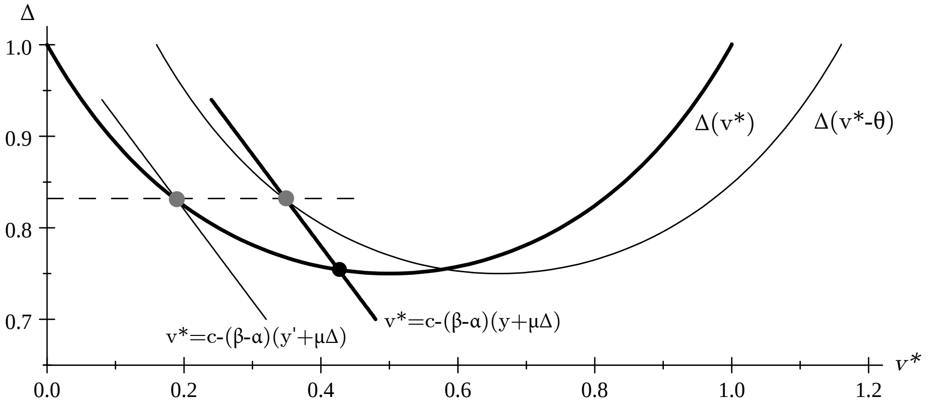

Fig. 5.b: Observable acts. Changes in eq. threshold and reputation caused by shifts in norms and a monetary bonus that generates equivalent changes

In both graphs, the thicker curves are associated with an initial distribution of types and with some initially given bonus. The thinner curves correspond to the shift in the distribution of types and to an 'equivalent' change in the bonus, respectively.[16] To illustrate, in the (v^*, Λ) plane of fig. 5a , the initial thick reputational curve is $\Lambda = \Lambda(v^*)$. The thick negatively sloped straight line is the best response threshold $v^* = c - (\beta - \alpha)(y + \mu\Lambda)$, given the initial bonus y. The initial equilibrium is the intersection of both curves. The thin reputational curve is $\Lambda = \Lambda(v^* - \theta)$. As depicted, the shift in intrinsic motivations would increase v^* and yield a smaller reputational benefit at equilibrium. Now, starting from the initial situation, an increase in the bonus to $y' = y + \frac{\theta}{\beta - \alpha}$ corresponds to a leftward θ shift in the best response line. Such a policy change would reduce

[16] The graphs use $G = \mathcal{B}(2,2)$ for the intitial distribution of types. In 5a, $\alpha = 0.1$, $\beta = 0.9$, $c - y = 0.4$, $\theta = 0.24$, $\mu = 1$. In 5b, $c - y = 1.18$, $\theta = 0.16$, $\mu = 2$.

v^* and result in the same drop in the equilibrium reputational benefit as the θ shift in the distribution of types. In either case, the *proportion* of agents participating in the pro-social act is the same, which explains why reputational benefits are the same.

To formalize these observations, we modify the equilibrium condition in (12) to incorporate the possibility of a shift in the distribution of types (and we omit references to type-1 and type-2 errors as arguments):

$$\varphi(v^*, y, \theta) \equiv v^* + (\beta - \alpha)[y + \mu\Lambda(v^* - \theta)] - c = 0 \tag{24}$$

such that the equilibrium threshold can now be expressed as $v^*(y, \theta)$, which solves (24). Using this equilibrium condition, we formalize our previous observations as follows.

Lemma 2 $v^*(y, \theta) = \theta + v^*\left(y + \frac{\theta}{\beta - \alpha}, 0\right)$.

Next, we examine equilibrium effects of shifts in norms by taking the monetary reward y as given. Applying lemma 2,

$$\left.\frac{\partial v^*(y,\theta)}{\partial \theta}\right|_{\theta=0} = 1 + \frac{\partial v^*\left(y,0\right)}{\partial y}\frac{1}{\beta - \alpha} = 1 - M(v^*\left(y,0\right)) \tag{25}$$

where the second equality follows from (17). As noted in section 2, whether the social multiplier, i.e. $M(v^*)$, exceeds or is smaller than unity is determined by whether acts are observable and whether the equilibrium threshold is small or large. Thus, an immediate implication of (25) is that the same factors also determine the direction of the impact of changes in social norms on the equilibrium threshold.

Proposition 2 *Assume the distribution of types is strictly unimodal and let* v_0^* *be the initial equilibrium. Following a small shift* $G(v - \theta)$*:*
(i) v^* *decreases (resp. increases) if* $v_0^* < \arg\min \Delta(v^*)$ *(resp.* $>$*), when acts are observable; and*
(ii) v^* *increases (resp. decreases) if* $v_0^* < \arg\max \Lambda(v^*)$ *(resp.* $>$*), when acts are unobservable and* $\beta - \alpha$ *is not too large.*

Proposition 2 highlights the contrast between comparative statics when acts are observable versus when they are unobservable: the equilibrium participation rate moves in opposite directions for small as well as large participation rates, provided $\beta - \alpha$ is not too large. These results are graphically illustrated in figures 5a and 5b for the case where the initial participation rate is large.[17]

Equilibrium responses to changes in social norms naturally play an important role in the determination of how optimal policies respond to such changes.

[17] The figures illustrate the impacts of discrete changes in social norms whereas proposition 2 is formulated in terms of marginal changes. Results pertaining to discrete changes in θ would be very similar. See, fo instance, Proposition 1 in Adriani and Sonderegger (2019) for the case of observable actions with both quasiconvex and quasiconcave reputation functions. In their analysis, quasiconcavity follows from a U-shaped distribution of types.

Consider first the case where $\lambda = 0$. Rewriting (23), and omitting reference to the type-1 and 2, the Pigovian subsidy is

$$y = \frac{e}{\beta - \alpha} - \mu\Lambda(v_{FB} - \theta) \tag{26}$$

Thus,

$$\left.\frac{dy}{\partial\theta}\right|_{\theta=0} = -\mu\Lambda_v(v_{FB}) \tag{27}$$

The change in the optimal bonus may be interpreted as a pure 'substitution effect' between formal and reputational incentives, in implementing the same target threshold v_{FB}.

For the case of a positive λ, we rewrite the first-order condition (21) as $F(v^*(y,\theta), y, \theta) = 0$, where

$$\begin{aligned} F(v^*(y,\theta), y, \theta) \quad &\equiv \quad [e + v^*(y,\theta) - c](\beta - \alpha) \\ &\quad - \lambda\left[(\beta-\alpha)^2 y + \frac{\rho(v^*(y,\theta))}{M(v^*(y,\theta)-\theta)}\right] \end{aligned} \tag{28}$$

For λ not too large, the above function is decreasing in y, so that $dy/d\theta$ has the same sign as

$$\frac{dF}{d\theta} = \frac{(\beta-\alpha)^2\mu\Lambda_v}{1 + (\beta-\alpha)\mu\Lambda_v} - \lambda\frac{d\left(\frac{\rho}{M}\right)}{d\theta} \tag{29}$$

We note that when the marginal cost of financing bonuses is small the optimal equilibrium threshold, v^*, is close to the first best solution v_{FB}. Moreover, the term multiplying λ in (29) is bounded in an interior solution. Thus, for appropriately small λ, the sign of F_θ tracks the sign of Λ_{v^*}, which implies the following result.

Proposition 3 *The optimal bonus and the equilibrium threshold move in the same direction as the partial impact of social norms on the equilibrium threshold (i.e.* $\frac{\partial v^*(y,\theta)}{\partial\theta}$*) described in proposition 2, as long as the marginal cost of financing bonuses is sufficiently small.*

The rationale behind this result can be uncovered by interpreting the first term in (29) as *first order reputation effects,* which arise from the direct impact of a shift in social norms on reputational benefits. The shift also affects the hazard rate, ρ, as well as the social multiplier M, a second order reputation effects so to speak.[18] Hazard rate effects and second order reputation effects are jointly captured by the second term in (29). When λ is small, first order reputation effects dominate, and thus determine the sign of (29).

Proposition 3 reveals how our insights regarding the impact of changes in social norms extends to changes in optimal policies. When the marginal cost of financing rewards is small, the sign of comparative statics are determined by the relationship between the equilibrium threshold and the extremum of reputational benefits, as described in proposition 2. Thus, the observability of acts continue to play a crucial role.

[18] The effect depends on the sign of Λ_{vv}.

4 Endogenous Errors

Up to this point, we analyzed how reputational concerns interact with the frequency of the pro-social act, given an exogenously determined pair of errors in the mapping between behavior and the receipt of bonuses. Here, we allow the two errors to emerge endogenously from the principal's choice.

4.1 Decision rule

We consider a situation where the principal receives a noisy signal from each agent that is imperfectly informative of his action. The signal has realizations $x \in [\underline{x}, \bar{x}]$ with probability density function f_a when the agent engages in act a, and associated cumulative distribution function F_a. The pair of probability distributions satisfies the monotone likelihood ratio property (MLRP) with $f_1(x)/f_0(x)$ decreasing in x. Thus, smaller realizations of the signal are more indicative of the agent having chosen the pro-social act. The principal uses a threshold rule $x_r \in [\underline{x}, \bar{x}]$, such that an agent receives the reward whenever the signal he generates satisfies $x < x_r$. This allows the principal to achieve any targeted probability of incorrectly rewarding an anti-social actor while maximizing the probability with which a pro-social actor is rewarded.

Given this type of decision rule, the probability of incorrectly rewarding an anti-social actor is $\alpha = F_0(x_r)$. Conversely, for any targeted level of type-1 error, the signal threshold that the principal chooses satisfies:

$$x_r(\alpha) \equiv F_0^{-1}(\alpha) \tag{30}$$

where F_0^{-1} denotes the inverse of F_0. It follows that the probability of correctly rewarding a pro-social agent satisfies:

$$\beta(\alpha) = F_1(F_0^{-1}(\alpha)) \tag{31}$$

Therefore, even though the principal chooses a signal threshold x_r for purposes of determining whom to reward, we may use (31) to express β as a function of α. This allows us to more conveniently focus on the two probabilities of receiving rewards, which are of key interest as was made apparent in the previous sections.

We may further investigate the relationship between these two probabilities, by noting that

$$\beta'(\alpha) = \frac{f_1(x_r(\alpha))}{f_0(x_r(\alpha))} > 0 \tag{32}$$

Because the likelihood ratio is decreasing in the signal threshold, and the latter is increasing in the targeted type-1 error, it also follows that $\beta(\alpha)$ is a strictly concave function:

$$\beta''(\alpha) = \frac{d\left(\frac{f_1(x_r(\alpha)))}{f_0(x_r(\alpha))}\right)}{dx_r(\alpha)} \frac{dx_r(\alpha)}{d\alpha} < 0 \tag{33}$$

Finally, (30) implies that $\beta(0) = 0$ and $\beta(1) = 1$. The strict concavity of the function therefore implies that $\beta(\alpha) > \alpha$ for all $\alpha \in (0, 1)$.

Thus, the set of review policies that the principal may choose from is summarized by the error pairs along an increasing and concave $\beta(\alpha)$ curve such as the ones depicted in Figure 6. Each curve plots the 'true positive rate' against the 'false positive rate', for a given underlying signal. In the figure, the middle and top curves are such that symmetric type-1 and 2 errors obtain when $\alpha = 0.1$ and $\alpha = 0.05$ respectively[19], which connects with the numerical examples in Figure 3; for the bottom curve, corresponding to the least informative signal, symmetric errors obtain at $\alpha = 0.3$. In the practice of discrimination tests (e.g. diagnostic tests, credit scoring, and the like), the $\beta(\alpha)$ functions in Figure 6 are known as Receiver Operating Characteristic (ROC) curves. As a useful typology, a test would be considered as 'acceptable' if the area under the curve (AUC) is between 0.7 and 0.8, as 'good' if the AUC is between 0.8 and 0.9, and as excellent or outstanding if it is over 0.9.[20] In this view, our examples range from good for the bottom curve to excellent for the two top ones.

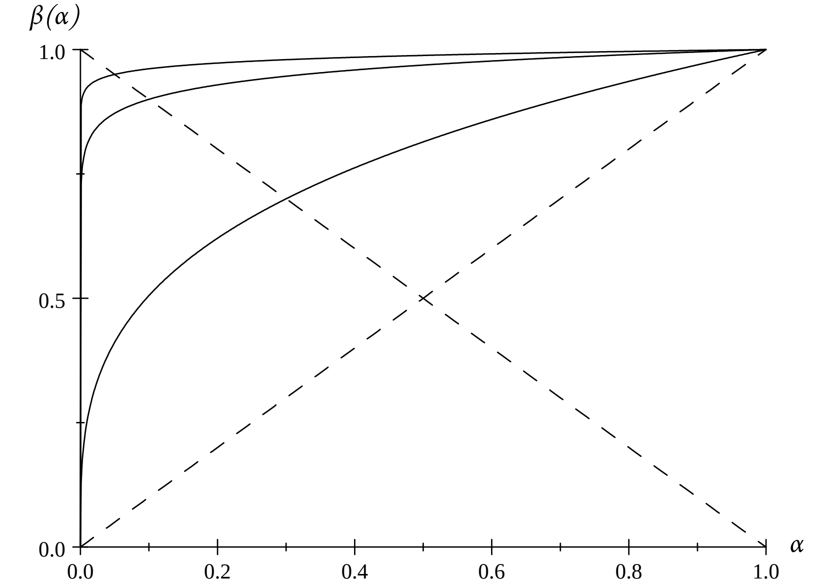

Fig. 6: β as a function of α.

4.2 Review Process and Incentives

The error pair chosen by the principal affects the agents' incentives. First, it does so by altering the expected material reward from engaging in the pro-social act, i.e. $(\beta(\alpha)-\alpha)y$. Secondly, it also impacts the expected reputational benefit $(\beta(\alpha)-\alpha)\mu\Lambda(v^*,\alpha,\beta(\alpha))$. In the sequel, given the function $\beta(\alpha)$, we shorten our notation to $\Lambda(v^*,\alpha)$. Similarly, $\psi(v^*,\alpha)$ is the proportion of agents receiving a bonus.

[19] The symmetric error level is where the $\beta(\alpha)$ curve intersects the straight line $\beta = 1 - \alpha$ drawn from $(0,1)$ to $(1,0)$.

[20] For instance, Hosmer and Lemeshow (2000).

For a given level of bonus, the expected material reward is maximized by the type-1 error $\overline{\alpha}$ that maximizes the probability differential $\beta(\alpha) - \alpha$. This solves

$$\beta'(\overline{\alpha}) = 1 \tag{34}$$

The review process associated with $\overline{\alpha}$ is an important benchmark. From (32), this benchmark defines a signal threshold x_r satisfying

$$\frac{f_1(x_r(\overline{\alpha}))}{f_0(x_r(\overline{\alpha}))} = 1$$

Thus, when $x_r(\overline{\alpha})$ is used as the threshold rule, an agent gets a bonus if, and only if, the signal received is more likely to have been emitted when the agent acted pro-socially rather than anti-socially.[21] This decision rule can also be characterized as minimizing the sum of the type-1 and type-2 errors, which equals $\alpha + 1 - \beta(\alpha)$.

The expected reputational benefit depends both on the probability differential, as for the expected material reward, and on the effect of the review process on the reputational benefit itself. It is useful to rewrite the latter as

$$\mu\Lambda(v^*, \alpha) = \frac{\beta(\alpha) - \alpha}{\psi(v^*, \alpha)(1 - \psi(v^*, \alpha))}\mu\tau(v^*) \tag{35}$$

where

$$\tau(v^*) \equiv (1 - G(v^*))G(v^*)\Delta(v^*) \tag{36}$$

is the part of the reputational benefit that does not depend on α. Observe that the numerator of (35) is concave in the type-1 error and maximized by $\overline{\alpha}$, while the denominator is concave and maximized by α such that $\psi(v^*, \alpha) = 1/2$. These properties imply the following result.

Lemma 3 *For any given v^*, the review process that maximizes the expected reputational benefit either (i) leads to infrequent bonuses (i.e. $\psi \leq 1/2$) and a type-1 error not larger than $\overline{\alpha}$, or (ii) leads to frequent bonuses (i.e. $\psi > 1/2$) and a type-1 error greater than $\overline{\alpha}$.*

Which of the two possibilities in the lemma holds will essentially depend on the functional form of β. To illustrate, we consider the following functions

$$\beta = \alpha^{1-\gamma} \text{ and } \beta = 1 - (1 - \alpha)^{\frac{1}{1-\gamma}},\ \gamma \in (0, 1)$$

In both cases, γ captures the level of informativeness of the underlying signal.[22] In Figures 7a and 7b, the parameter value is $\gamma = 0.5$ for both functions. We plot the expected reputational benefit (assuming $\mu = 1$) for the cases where v^* equal 0.1 (*thick*), 0.3 (*medium*), 0.5 (*thin*), and 0.7 (*dashed*). The distribution of types used to compute $\psi(v^*, \alpha)$ is the unimodal symmetric Beta distribution with parameters $w = z = 2$. The vertical straight line identifies $\overline{\alpha}$.

[21] See Demougin and Fluet (2005, 2006) on the incentive properties of such a decision rule.

[22] For any level of the type-1 error, β is larger the larger the value of γ.

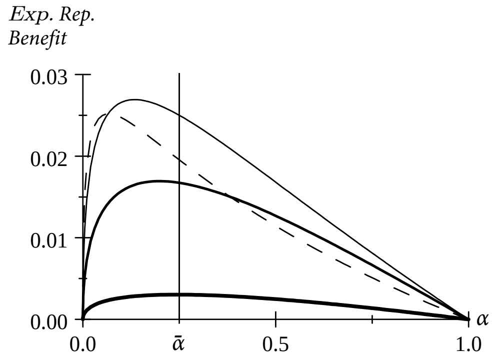

Fig. 7a: Expected reputational benefit and $\overline{\alpha}$ when $\beta = \alpha^{0.5}$

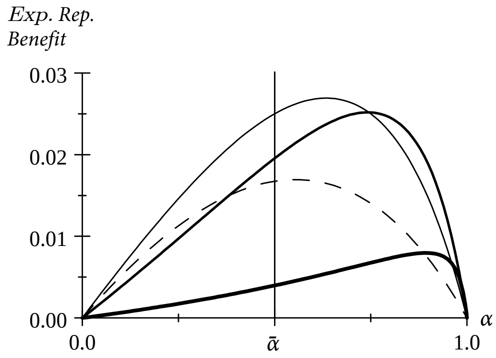

Fig. 7b: Expected reputational benefit and $\overline{\alpha}$ when $\beta = 1 - (1-\alpha)^2$

Observe that in the left-hand figures expected reputational benefits tend to be maximized by a type-1 error smaller than $\overline{\alpha}$, irrespective of the participation rate in the pro-social activity. The converse holds in the right-hand side figures. These examples illustrate how reputational benefits respond to changes in the review process, taking the participation rate as given. The shape of these responses play an important role in the determination of optimal review processes, but now also taking into account the effect of the review process on the participation rate.

5 Endogenous Rewards and Review Standards

Having explained the mechanics of the review standard, we now consider how the principal may maximize welfare by choosing both the review standard and the size of the formal incentives. Both affect the equilibrium participation threshold, written as $v^*(y, \alpha)$. As will become apparent, it is useful to decompose welfare, previously expressed by (19), into two components: (i) the direct net-benefit from the act:

$$D(v^*) \equiv \int_{v^*}^{v_{\max}} (e + v - c) g(v) dv \tag{37}$$

and (ii) the expected deadweight loss of financing bonuses:

$$C(v^*, y, \alpha) \equiv \lambda y \psi(v^*, \alpha) \tag{38}$$

Quite importantly, the net-benefit depends only on the achieved threshold, v^*, while the cost depends on the threshold as well as the mix of instruments, y and α.

5.1 Minimizing Cost

First, we identify cost minimizing policies that a principal would choose to implement some target threshold, v^*. This analysis reveals insights that are also useful in discussing policies chosen by the principal who targets an optimal threshold.

From (12), the required bonus for implementing the targeted threshold v^* is

$$y = \frac{c - v^*}{\beta(\alpha) - \alpha} - \mu\Lambda(v^*, \alpha) \tag{39}$$

where the reputation term is as defined in (35). Assuming the required bonus is positive and substituting in (38), the deadweight loss is

$$C(v^*, \alpha) = \frac{\lambda(c - v^*)\psi(v^*, \alpha)}{\beta(\alpha) - \alpha} - \lambda\mu\Lambda(v^*, \alpha)\psi(v^*, \alpha) \tag{40}$$

We note that the first term in (40) is monotonically increasing as the review standard is made weaker, since[23]

$$\frac{\partial\left(\frac{\lambda(c-v^*)\psi(v^*,\alpha)}{\beta(\alpha)-\alpha}\right)}{\partial\alpha} = \frac{\lambda(c - v^*)(\beta - \alpha\beta')}{(\beta - \alpha)^2} > 0 \tag{41}$$

In the extreme case where individuals have no reputational concerns, $\mu = 0$, and thus the relationship between the review standard and deadweight loss is governed entirely by (41). Next, we analyze this case in more detail, since it has important implications for cases where $\mu > 0$.

The case $\mu = 0$

When individuals have no reputational concerns, it immediately follows from (40) and (41) that minimized costs equal

$$C_N(v^*) \equiv \lambda(c - v^*)k(v^*)(1 - G(v^*)) \tag{42}$$

where

$$k(v^*) \equiv \lim_{\alpha \to 0^+} \frac{\psi(v^*, \alpha)}{(\beta(\alpha) - \alpha)(1 - G(v^*))} = \frac{G(v^*) + \beta'(0)(1 - G(v^*))}{(\beta'(0) - 1)(1 - G(v^*))} \tag{43}$$

It is easily seen that $k(v^*) \geq 1$, with strict inequality if $\beta'(0)$ is finite.

Thus, when $\mu = 0$, the cost minimizing bonus is unbounded, a familiar result (Becker 1968). For practical purposes, the interpretation may be that there is a minimum 'granular' level of type-1 error, some small value, below which an informative review standard cannot go. Alternatively, there is a large upper bound on allowable bonuses. In other words, when there are no reputational concerns, the optimal policy has a very demanding review standard, equivalently a very small type-1 error, and a very large bonus. We will refer to it, loosely speaking, as the 'no type-1 error' policy.

The possibility that $\beta'(0)$ is unbounded cannot be excluded, as it corresponds to situations where the likelihood ratio $f_1(x)/f_0(x)$ is arbitrarily large for some

[23] The strict concavity of $\beta(\alpha)$, together with the boundary condition $\beta(0) = 0$, implies $\beta(\alpha) > \alpha\beta'(\alpha)$ for all $\alpha \in (0, 1)$.

realizations of the signal on which the review process is based. In such cases, $k(v^*) = 1$ and the cost of the 'no type-1 error' policy reduces to

$$C_N(v^*) = \lambda(c - v^*)(1 - G(v^*))$$

This is the the same as if acts were perfectly observable, even though the signal is not perfectly informative. When acts are observable, all agents who behave pro-socially receive a bonus equal to $c - v^*$. Here, with the 'no type-1 error' policy, the same agents (and only they) face a very small probability of receiving a very large bonus. In expectation, they get $c - v^*$ from engaging in the pro-social act.

The case $\mu > 0$

When people have reputational concerns, the monetary benefits that must be financed to achieve a behavior profile is naturally decreasing in the size of the equilibrium reputational concerns. This is reflected by the second term in (40). Under a 'no type-1 error' policy, this term vanishes because the policy does not provide any information about agents' types. Thus, the 'no type-1 error' policy implements the target threshold at the same cost as when agents have no reputational concerns. However, this policy need not be cost minimizing because the reduction in monetary incentives required to implement the targeted level of participation, captured by the second term in (40), is strictly increasing in α. We note sufficient conditions under which the 'no type-1 error' is not cost minimizing, as follows.

Lemma 4 *Either the cost of implementing v^* is minimized with the 'no type-1 error' policy in which case it equals $C_N(v^*)$; or it is minimized by some $\alpha > 0$ in which case it is lower than $C_N(v^*)$. Two different sufficient conditions for the latter are*

$$c - v^* \leq \max_{\alpha}(\beta(\alpha) - \alpha)\mu\Lambda(v^*, \alpha) \tag{44}$$

and

$$\frac{\mu G(v^*)(1 - G(v^*))\Delta(v^*)}{c - v^*} > \frac{-\beta''(0)}{2(\beta'(0) - 1)^3} \tag{45}$$

When (44) holds, reputational incentives are sufficient by themselves to implement v^*. The cost minimizing bonus is then nil and so is the cost. Abstracting from this case, the issue is whether the cost minimizing policy relies solely on formal incentives or whether reputational incentives also play a role. Condition (45) ensures that $C(v^*, \alpha)$ is strictly decreasing in α for small values. When the right-hand side of (45) is nil, the inequality is satisfied for all $v^* \in (v_{\min}, v_{\max})$; when the right-hand side is positive and not too large, the condition can be satisfied only if the target participation rate is not too extreme, that is, $G(v^*)$ is not too close to zero or one.[24]

[24] When $\beta(\alpha) = 1 - (1 - \alpha)^{\frac{1}{1-\gamma}}$, $\gamma \in (0, 1)$, the right-hand side of (45) is positive and finite. When $\beta(\alpha) = \alpha^{1-\gamma}$, the right-hand side equals zero if $\gamma < 1/2$ and is infinite if $\gamma > 1/2$.

Lemma 4 shows that reputational benefits may cause the optimal policies to deviate from Beckerian results discussed in detail in the law enforcement context (see, e.g. Polinsky and Shavell (2007)). As noted, when (44) holds, the cost minimizing policy is a 'no bonus' policy (i.e. $y = 0$). Otherwise, bonuses are necessary since reputational incentives are insufficient on their own to achieve the targeted participation rate. Then, when (45) holds (and note this is only a sufficient condition) the cost minimization problem has an interior solution with $\alpha > 0$ and $y > 0$. In such cases, the cost minimizing mix of instruments satisfies

$$\frac{\partial v^*(y,\alpha)/\partial\alpha}{\partial v^*(y,\alpha)/\partial y} = \frac{y\psi_\alpha(v^*,\alpha)}{\psi(v^*,\alpha)} \tag{46}$$

where $\partial v^*/\partial y$ is as defined in (16) and

$$-\frac{\partial v^*}{\partial\alpha} = \frac{(\beta'-1)(y+\mu\Lambda)+(\beta-\alpha)\mu\Lambda_\alpha}{1+(\beta-\alpha)\mu\Lambda_{v^*}} \tag{47}$$

is a similar expression for the effect of a marginal change in the type-1 error. Specifically, following a marginal increase in α, the increase in the expected total benefit (both material and reputational) is equal to the numerator in (47). The effect on the equilibrium threshold is (minus) the increase in expected benefits times the social multiplier.

Equation (46) is a standard 'production theory' condition: the rate of substitution between α and y, keeping participation constant, must equal the relative effects on the cost. The right-hand side of (46) is positive, $\partial v^*/\partial y$ is negative, therefore $\partial v^*/\partial\alpha$ must also be negative. Thus, in the cost minimizing policy, marginally relaxing the review standard (a larger α) increases the total expected benefit from participation in the pro-social activity. This decomposes into two effects. The numerator in (47) can be rewritten as

$$(\beta'(\alpha)-1)y+\mu\frac{\partial[(\beta(\alpha)-\alpha)\Lambda(v^*,\alpha)]}{\partial\alpha}$$

The first term relates to the formal incentives and is the change in the expected bonus for participating in the pro-social act, when the review standard is relaxed. This is positive for $\alpha < \overline{\alpha}$, but negative for less demanding review standards. The second term is the change in the expected reputational benefit for participating in the act due to the increase in α, which upon further examination can be shown to always be positive in a cost minimizing policy. To see this, note that (46) can be re-written as

$$\frac{(\beta'-1)(y+\mu\Lambda)+(\beta-\alpha)\mu\Lambda_\alpha}{\beta-\alpha} = \frac{y\psi_\alpha(v^*,\alpha)}{\psi(v^*,\alpha)} \tag{48}$$

which is easily shown to yield

$$\frac{\partial[(\beta-\alpha)\mu\Lambda]}{\partial\alpha} = \frac{(\beta-\alpha\beta')y}{\psi} > 0$$

In Figure 7a, for various thresholds v^*, the expected reputational benefit is maximized at some $\alpha < \overline{\alpha}$. Therefore, minimizing costs then requires a review standard below $\overline{\alpha}$. In Figure 7b, by contrast, the expected reputational benefit is maximized at $\alpha > \overline{\alpha}$. One cannot then exclude the possibility that the cost minimizing policy involves a review standard laxer than $\overline{\alpha}$. We summarize the implications of these observations, and related results, as follows.

Proposition 4 *In a cost minimizing policy with $y > 0$, either the type-1 error α is less than $\bar{\alpha}$ or rewards are frequent, i.e. $\psi(v^*, \alpha) > 1/2$.*

Both conditions in the proposition may hold simultaneously. However, a policy with $\alpha > \bar{\alpha}$ can be cost minimizing policy only when receipt of a bonus is the norm, in the sense that a majority of individuals are rewarded.[25]

It may be remarked that the relationship between the optimal review standard and $\bar{\alpha}$ (which minimizes the sum of errors) has received considerable attention in the law and economics literature on standards of proof (Posner 2007, Rizzolli and Saraceno 2013). Specifically, many articles have sought to identify economic rationales for the use of strong standards, such as beyond a reasonable doubt, in criminal trials. An implication of proposition 4 is that reputational concerns can supply a rationale for such standards.

We note, however, that the frequency with which formal incentives are provided in the cost minimizing policy depend crucially on the properties of the signal generating process. We demonstrate this via numerical examples summarized through tables in Appendix B. These tables describe the cost minimizing policy for a range of target thresholds v^* and values of the private cost c, and for different signals underlying the review process. The results illustrate, in particular, that the optimal α may be non-monotonic in the target participation rate.

5.2 Optimal Policy

An optimal policy implements the participation threshold that maximizes $D(v^*) - \min_{v^*} C(v^*, y(v^*), \alpha(v^*))$ where y and α refer to the cost minimizing policies. As described in the previous section, any one of three broad types of policies ('no type-1 error' policy, 'no bonus' policy, or $\alpha, y > 0$) may be used depending on the underlying signal and on the targeted participation rate. When the participation rate itself is chosen optimally, which of these three broad categories of policies the targeted rate will be associated with naturally depends on the comparison between the marginal deadweight loss from increasing the participation rate (given that the policy instruments are chosen to minimize costs) and the marginal benefits from additional participation.

When, contrary to assumption 1, reputational concerns are large enough to allow the first best participation rate to be achieved without the aid of any

[25] This will arise, for instance, when $1 - G(v^*) > 1/2$, i.e. the pro-social act is majoritarian, and $\alpha \geq 1 - \beta(\alpha)$. In Figure 6, for the top curve, the preceding inequality is satisfied for all $\alpha \geq 0.05$; for the middle curve, it is satisfied for all $\alpha \geq 0.1$.

bonuses, i.e. when (44) holds at $v^* = v_{FB}$, implementing the first best leads to no deadweight loss. The optimal solution is then trivial. No bonuses are used and first best incentives are provided exclusively through the optimal choice of the review standard.[26]

Thus, a meaningful trade-off between the marginal costs and benefits from targeting a particular participation rate arises only when assumption 1 holds. This trade-off depends crucially on the shadow cost of public funds, λ. We provide an exhaustive description of possibilities and show how λ affects the optimal policy.

Proposition 5 *The optimal policy consists of either: (i) a 'no type-1 error' policy, (ii) an interior solution (i.e. $y > 0$ and $\alpha > 0$), or (iii) a 'no bonus' policy with $y = 0$ and where α maximizes the expected reputational benefit. (iv) There exists a critical value $\bar{\lambda} > 0$ such that the optimal solution is either of the form (i) or (ii) when $\lambda < \bar{\lambda}$ and of the form (iii) when $\lambda > \bar{\lambda}$.*

In addition to highlighting how the size of λ may affect optimal policies, proposition 5 imposes a restriction on the type of 'no bonus' policies that can be optimal. The review standard must then be chosen so as to maximize reputational incentives. This is because the maximum expected reputational benefit that is achievable falls short of incentivizing the first best participation rate. Starting with a policy which only harnesses reputational incentives, and makes no use of monetary bonuses, the introduction of a small bonus leads to marginal benefits due to increased participation. However, the marginal cost of financing such bonuses may be larger than their benefits in terms of incentives. In this case the best option may be to use no bonuses at all and remain at a participation rate below the first best, harnessing reputational incentives as much as possible to mitigate the sub–optimal participation. These no bonus solutions mimic symbolic conferral of rewards which carry no monetary value (or monetary values which are negligible in comparison to the reputational rewards conferred) for which we have listed some examples in the introduction.

As noted in part (iv) of the proposition, purely symbolic rewards become sub-optimal when the marginal cost of financing material rewards is sufficiently small. How small λ needs to be is naturally context dependent. For instance, the magnitude of the resource cost c and of the externality e, as well as the relative importance of reputational benefits μ and the informativeness of the signal all affect the critical λ. When the cost associated with financing bonuses is small given the context, welfare can be improved upon compared to the 'no bonus' policy by providing some monetary incentives to increase participation. The optimal policy is then either an interior solution or a 'no type-1 error' policy, and cannot be pinned down any further absent additional restrictions. Loosely speaking, the optimal policy will be an interior solution when the signals that are indicative of the pro-social act are sufficiently informative or when the second-best participation rate targets are not too extreme. For instance, condition (45)

[26] The α that yields the appropriate reputational incentives is then generally not unique. See Appendix B.

in lemma 4 is satisfied for any signal defined by $\beta = \alpha^{1-\gamma}$ with $\gamma > 1/2$. In such cases, it follows that when the cost of financing rewards is sufficiently small, the optimal solution is interior. Next, we identify the impacts of shifts in norms on optimal policies.

5.3 Shifts in Norms

Our previous discussion shows that the optimal policy can be one of three types. Therefore, we study the impact of a shift in norms (as described in section 3) on the optimal participation rate and policy instruments given that one starts with one of the three types of optimal policies as an initial solution. We start by summarizing comparative statics in each case and relegate the technical derivation and explanation of these results to the appendix. Subsequently, we describe some of the complexities that prohibit generally signing the direction of effects in some cases.

Proposition 6 *(i) When the optimal solution is a 'no bonus' policy, a small positive shift $G(v-\theta)$ leads to an increase in α and a decrease in v^*.*

(ii) When the optimal solution is interior and λ is sufficiently small, the impact of a small positive shift $G(v-\theta)$ is a decrease in either α or y if $\Lambda_v(v^*, \alpha) < 0$ and an increase in either α or y if $\Lambda_v(v^*, \alpha) > 0$.

(iii) When the optimal solution is a 'no type-1 error policy', a small positive shift $G(v-\theta)$ decreases v^* if the reciprocal hazard rate $h(v^*) \equiv \frac{1-G(v^*)}{g(v^*)}$ is either increasing or decreasing slowly in v^*, i.e. if $h'(v^*) > -\frac{k(v^*)-1}{k(v^*)}$; otherwise, v^* increases.

When the initial optimal solution involves a purely symbolic reward, shifts in norms have predictable effects because they impact a single interior optimal policy tool, namely the review standard. An increase in the population's sentiments towards the pro-social act then causes the optimal review policy to be 'laxer'. This is because, as noted in proposition 5, optimal 'no bonus' policies maximize expected reputational benefits, and a shift in social norms increases the marginal impact of relaxing the review standard. Thus, the population's intrinsic motivations and symbolic rewards act as, in a sense, complements, and an increase in the former causes the latter to be utilized more generously.[27]

The second part of the proposition reveals how the possibility of adjusting α in addition to y in response to shifts in norms causes complexities in the optimal responses. Unlike in the exogenous α case, one cannot ascertain the direction towards which monetary bonuses must be adjusted and the same is true for α. Shifts in norms alter the welfare maximizing monetary bonus, holding α constant, as explained in section 3. However, they also impact the optimal review process, and thereby indirectly further affect the optimal bonus, which is responsive to changes in the review process as well.

[27] More formally, a positive shift in norms increases the 'productivity' of a marginal increase in α, i.e. $\frac{\partial(\beta-\alpha)\Lambda}{\partial\alpha}$.

Nevertheless, a more precise characterization can be obtained by focusing on the pure substitution effect associated with shifts in norms.[28] Denote the optimal policies as $y(\theta) = \hat{y}(\theta, v^*(y(\theta), \alpha(\theta), \theta))$ and $\alpha(\theta) = \hat{\alpha}(\theta, v^*(y(\theta), \alpha(\theta), \theta))$, where $\hat{y}$ and $\hat{\alpha}$ refer to the cost-minimizing policies described in section 5.2. Thus, the changes in optimal policies are given by

$$\begin{aligned} &\frac{\partial \hat{y}}{\partial \theta} + \frac{\partial \hat{y}}{\partial v^*}\frac{dv^*}{d\theta}; \text{ and} \\ &\frac{\partial \hat{\alpha}}{\partial \theta} + \frac{\partial \hat{\alpha}}{\partial v^*}\frac{dv^*}{d\theta} \end{aligned} \tag{49}$$

The first term in each expression can be interpreted as the direct effect of a change in θ, a substitution effect. The second terms can be interpreted as a threshold effect analogous to an income effect. We note that the direction of the substitution effect can be fully characterized, as follows.

Corollary 1 *When the optimal policy is interior, there exists $\varepsilon_2 > \varepsilon_1 > 0$ such that the substitution effect of a positive shift $G(v^* - \theta)$ satisfies: $\partial\hat{y}/\partial\theta < 0 < \partial\hat{\alpha}/\partial\theta$ if $\Lambda_v(v^*, \alpha) < \varepsilon_1$; $\partial\hat{y}/\partial\theta > 0$ and $\partial\hat{\alpha}/\partial\theta > 0$ if $\Lambda_v(v^*, \alpha) \in (\varepsilon_1, \varepsilon_2)$; and $\partial\hat{\alpha}/\partial\theta < 0 < \partial\hat{y}/\partial\theta$ if $\Lambda_v(v^*, \alpha) > \varepsilon_2$.*

The rationale behind corollary 1 can be noted by observing that substitution effects correspond to changes in cost-minimizing policies. From the cost minimization problem in 5.1, it can be shown that positive shifts in norms reduce the relative marginal cost of increasing α versus y (captured by the right hand side of (46)) while also increasing the marginal rate of substitution (captured by the left hand side of (46)) unless Λ_v is positive and sufficiently large. Positive shifts in norms then cause the cost minimizing review standard to be laxer and the bonus to be smaller. In other words, one relies more on reputational incentives and less on formal incentives. These cases correspond to situations where the initial v^* is not too small. The opposite response will emerge when the initial participation rate is large, i.e. Λ_v is positive and large. The preceding observations will apply to the optimal policy with endogenous v^* as long as the optimal threshold is relatively unresponsive to changes in social norms. However, when 'threshold effects' are sizeable, it is not possible to make statements that are more general than those summarized via proposition 6.

Finally, when the initial optimal policy is the 'no type-1 error policy', meaningful statements cannot be made about the impact of a shift in social norms on the individual instruments. Nevertheless, one can still make observations regarding the optimal threshold, as noted in part (iii) of the proposition. As explained in section 2.3, the reciprocal hazard rate[29] captures the relative importance of marginal financing costs compared to marginal benefits associated with increasing participation. A shift in θ causes a change in the participation rate, and hence h. When h is an increasing function, the relative importance of

[28] Recall our discussion of the effect of such shifts on the Pigovian subsidy.

[29] In section 2.3, the relevant concept was captured by ρ, which in the current context reduces to h due to the 'no type-1 error' policy.

participation is increased, and the optimal threshold goes down when $k(v^*) = 1$. These dynamics are only slightly more complicated when $\beta'(0)$ is finite, since then $k(v^*) > 1$ as noted in the discussion of this term following (43).

6 Concluding Remarks

The nature of the interactions between reputational incentives and the frequency with which good or bad acts are committed is quite sensitive to whether acts are perfectly observable by third parties. When they are perfectly observable as shown in Bénabou and Tirole (2011), norms emerge which attach large reputational consequences to acts that are committed by a very large or a very small fraction of actors. We have shown that the opposite conclusion holds when acts are unobservable by third parties who must rely on noisy signals in forming opinions about others. Perhaps more importantly, we have shown that, even with small errors in the conveyance of information, reputational sanctions are inverse-U shaped rather than U-shaped in the social prevalence of the act.

These results can be viewed as disturbing, because they undermine an intuitive economic explanation as to why acts with extreme participation rates, in fact, have large reputational consequences attached to them. Our results, which rely on identical assumptions as the previous literature, with the exception of incorporating unobservable actions, do not suggest that there is no economic rationale for this stylized fact, but, instead, suggests that the explanation behind this relationship perhaps lies elsewhere. Specifically, both the honor-stigma model, and our model, take the magnitude of reputational incentives relative to bonuses, namely μ, as given. This value, which is exogenously given, need not be constant *across* acts. In fact, holding all else constant (e.g. the social prevalence of the act, the informativeness of signals about commission of the act, etc.) a signal that is equally informative of the commission of one act versus another may implicate much greater reputational consequences than another. It would be senseless to assume that murder and practicing interior design without a license would generate the same degree of reputational harm for a person, if they were committed equally frequently. This is because the commission of the two acts reveal different kinds of information about the character of the person committing it. Why people respond by imposing greater reputational sanctions on murderers versus illegal-interior-designers is a question not about *how* third parties make inferences, but *about what* they make inferences.

Thus, our model, as well as the honor-stigma model preceding it, should presumably be used not to make categorizations across acts, but to analyze how policies ought to be designed given the presence of reputational incentives. Our analysis suggests that, when formal rewards or sanctions are noisy signals of behavior, it will often be optimal to rely substantially or perhaps exclusively on formal incentives if the target is a high rate of pro-social behavior. Such a policy will typically involve large rewards and a stringent standard for rewarding agents. By contrast, when the target is a more 'modal' rate of pro-social behavior, it will often be optimal to rely on a mix of formal and informal or reputational incentives. The policy will then involve moderate rewards and less stringent review standards. Less demanding standards increase the visibility of rewards, which helps in sustaining reputational incentives, thereby reducing the need for socially costly formal rewards. Exogenous changes in the intrinsic motivations to behave pro-socially will also affect the optimal mix of formal and informal incentives. When intrinsic motivations improve and pro-social behavior becomes more modal, the optimal policy may tilt towards greater reliance on informal incentives, e.g. the review standard is relaxed and rewards are reduced. By contrast, when the improvement in intrinsic motivations implies that pro-social behavior becomes even less modal, it may be optimal to tilt the policy towards a more stringent standard and a larger formal reward.

Appendix A

Proof of Lemma 1: From (2), $\psi(v^*_{\min}, \alpha, \beta) = \beta$ and $\psi(v^*_{\max}, \alpha, \beta) = \alpha$. Therefore, given $\alpha > 0$ and $\beta < 1$, the denominator in (9) is always greater than zero, so that $\delta(v^*_{\min}, \alpha, \beta) = \delta(v^*_{\max}, \alpha, \beta) = 0$. That $\delta < 1$ follows from the fact that δ can be rewritten as

$$\delta = \frac{\beta(1-G)}{\alpha G + \beta(1-G)} - \frac{(1-\beta)(1-G)}{(1-\alpha)G + (1-\beta)(1-G)}$$

where the first term is less than one.

From (9), $\partial\delta/\partial v^* > 0$ is equivalent to

$$g(1-2G)\psi(1-\psi) - \frac{\partial\psi}{\partial v^*}(1-2\psi)(1-G)G > 0 \tag{50}$$

Substituting $\partial\psi/\partial v^* = -g(\beta-\alpha)$, the preceding inequality becomes

$$(1-2G)\psi(1-\psi) + (\beta-\alpha)(1-2\psi)(1-G)G > 0$$

or equivalently

$$(1-G)(1-\psi)[\psi + (\beta-\alpha)G] > G\psi[(\beta-\alpha)(1-G) + (1-\psi)]$$

Substituting for $\psi = \beta(1-G)+\alpha G$ in the squared brackets and cancelling terms then yields

$$A \equiv G\psi\frac{[1-\beta-\alpha]}{\beta} + (G+\psi) < 1$$

Thus, (50) is equivalent to $A < 1$. It is easily checked that $A = \beta < 1$ when $v^* = v_{\min}$ and $A = 1+\alpha(1-\alpha)/\beta$ when $v^* = v_{\max}$. Therefore, if A is everywhere strictly increasing in v^*, then δ must be first strictly increasing in v^*, then strictly decreasing. To show that A is indeed strictly increasing, let $N = \frac{1-\beta-\alpha}{\beta}$ and note that

$$\begin{aligned}\frac{\partial A}{\partial v^*} &= g(N\psi+1) + \frac{\partial\psi}{\partial v^*}(1+NG) \\ &= g(N\psi+1) - g(\beta-\alpha)(1+NG)\end{aligned}$$

Thus, $\partial A/\partial v^* > 0$ is equivalent to

$$1-\beta+\alpha > N[(\beta-\alpha)G-\psi]$$

Substituting for $\psi \equiv \beta(1-G)+\alpha G$ again, the preceding inequality is equivalent to

$$1-\beta+\alpha > N[2(\beta-\alpha)G-\beta] \tag{51}$$

If $N \geq 0$, the right hand side is non decreasing in G. A sufficient condition for (51) to hold everywhere is then that it holds at $G=1$, i.e.

$$1-\beta+\alpha > N(\beta-2\alpha) = 1-\beta-\alpha-2N\alpha$$

which is true for $N \geq 0$. If $N < 0$, it suffices that (51) holds at $G=0$, i.e.

$$1-\beta+\alpha > -N\beta = -(1-\beta-\alpha)$$

which reduces to $2 > 2\beta$.■

Proof of Proposition 1: For part (i) we refer the reader to BT or Adriani and Sonderegger (2019). The first two claims in part (ii) follow trivially from Lemma 1. To prove strict quasiconcavity, let

$$\varsigma(v^*,\alpha,\beta) \equiv \frac{(1-G(v^*))G(v^*)}{\psi(v^*,\alpha,\beta)(1-\psi(v^*,\alpha,\beta))}\Delta(v^*)$$

Note that $\psi(v^*,q,q)=q$. Thus, for any $q \in (0,1)$,

$$\begin{aligned}\varsigma(v^*,q,q) &= \frac{(1-G(v^*))G(v^*)}{q(1-q)}\left(\frac{\int_{v^*}^{v_{\max}} vg(v)dv}{1-G(v^*)} - \frac{\int_{v\min}^{v^*} vg(v)dv}{G(v^*)}\right) \quad (52)\\ &= \frac{G(v^*)\int_{v_{\min}}^{v_{\max}} vg(v)dv - \int_{v\min}^{v^*} vg(v)dv}{q(1-q)}\end{aligned}$$

which reveals that

$$\frac{\partial\varsigma}{\partial v^*} = g(v^*)\frac{\bar{v}-v^*}{q(1-q)} \tag{53}$$

Because $\varsigma(v^*,q,q)$ is strictly increasing for $v^* < \bar{v}$ and strictly decreasing for $v^* > \bar{v}$, it is strictly quasiconcave in v^*. Therefore, for any v_0^* and v_1^* and $t \in (0,1)$,

$$\varsigma(tv_0^*+(1-t)v_1^*,q,q+\varepsilon) - \min\{\varsigma(v_0^*,q,q+\varepsilon),\varsigma(v_1^*,q,q+\varepsilon)\} > 0 \tag{54}$$

when $\varepsilon = 0$. By continuity, the inequality also holds for $\varepsilon > 0$ and sufficiently small. Letting $\alpha = q$ and $\beta = q+\varepsilon$, it follows that $\Lambda = (\beta-\alpha)\varsigma(v^*,\alpha,\beta)$ is strictly quasiconcave in v^* for sufficiently small $\beta-\alpha > 0$.■

Proof of Lemma 2: Let $v^*(y, \theta)$ solve

$$v^*(y, \theta) + (\beta - \alpha)\,(y + \mu\Lambda(v^*(y, \theta) - \theta) = c \tag{55}$$

Then $\widetilde{v} \equiv v^*(y + \theta/(\beta - \alpha), 0)$ solves

$$\widetilde{v} + (\beta - \alpha)\left(y + \frac{\theta}{\beta - \alpha} + \mu\Lambda(\widetilde{v})\right) = c$$

which reduces to

$$\widetilde{v} + \theta + (\beta - \alpha)(y + \mu\Lambda(\widetilde{v})) = c$$

Therefore, $v^*(y, \theta) = \theta + \widetilde{v}$ solves (55). In particular, $\Lambda(v^*(y, \theta) - \theta) = \Lambda(\widetilde{v})$.■

Proof of Proposition 2: From (25), at $\theta = 0$,

$$\frac{\partial v^*}{\partial \theta} = 1 - M(v^*) = \frac{(\beta - \alpha)\mu\Lambda_v(v^*)}{1 + (\beta - \alpha)\mu\Lambda_v(v^*)}$$

The rest of the argument then follows directly from proposition 1.■

Proof of Proposition 3: See the argument in the text.

Proof of Lemma 3: Maximizing $(\beta(\alpha) - \alpha)\mu\Lambda(v^*, \alpha)$ with respect to α yields an interior solution $\alpha \in (0, 1)$ satisfying the first-order condition:

$$(\beta'(\alpha) - 1)\mu\Lambda(v^*, \alpha) + (\beta(\alpha) - \alpha)\mu\Lambda_\alpha(v^*, \alpha) = 0$$

Substituting from (35) and (36), the above condition is equivalent to

$$2(\beta' - 1)\psi(1 - \psi) - (\beta - \alpha)(1 - 2\psi)\psi_\alpha = 0$$

which implies the statements in the lemma, given that $\psi_\alpha = G + \beta'(1 - G) > 0$.■

Proof of Lemma 4: When (44) holds, $y = 0$ and $C(v^*) = 0$ with α solving

$$(\beta(\alpha) - \alpha)\mu\Lambda(v^*, \alpha) = c - v^*$$

When this equation has no solution, $y > 0$ and can be expressed in terms of α, yielding $C(v^*, \alpha)$ as defined in (40). The minimum is characterized by $\alpha > 0$ if $C_\alpha(v^*, 0^+) \equiv \lim_{\alpha \to 0^+} C_\alpha(v^*, \alpha) < 0$. Let $K(v^*, \alpha) \equiv C(v^*, \alpha)/\lambda$. Substituting from (35) in the cost function (40) and defining $K(v^*, \alpha) \equiv C(v^*, \alpha)/\lambda$,

$$\begin{aligned} K_\alpha(v^*, \alpha) &= \frac{(c - v^*)[(\beta - \alpha)\psi_\alpha - (\beta' - 1)\psi]}{(\beta - \alpha)^2} \\ &\quad - \frac{[(\beta' - 1)(1 - \psi) + (\beta - \alpha)\psi_\alpha]\mu\tau}{(1 - \psi)^2} \\ &= \frac{(c - v^*)(\beta - \alpha\beta')}{(\beta - \alpha)^2} - \frac{(\beta - \alpha\beta' + \beta' - 1)\mu\tau}{(1 - \psi)^2} \end{aligned} \tag{56}$$

where the strict concavity of $\beta(\alpha)$ implies $\beta - \alpha\beta' > 0$ and $\beta - \alpha\beta' + \beta' - 1 > 0$ for all $\alpha \in (0, 1)$. Therefore, $C_\alpha(v^*, \alpha) < 0$ or equivalently if $K_\alpha(v^*, \alpha) < 0$ if

$$\frac{\mu\tau}{c - v^*} > \frac{(\beta - \alpha\beta')(1 - \psi)^2}{(\beta - \alpha)^2(\beta - \alpha\beta' + \beta' - 1)}$$

where the right-hand side is positive. Taking the limit, $C_\alpha(v^*, 0^+) < 0$ if

$$\frac{\mu\tau}{c - v^*} > \lim_{\alpha\to 0^+} \frac{(\beta - \alpha\beta')(1 - \psi)^2}{(\beta - \alpha)^2(\beta - \alpha\beta' + \beta' - 1)} = \frac{-\beta''(0)}{2(\beta'(0) - 1)^3}$$

which is equivalent to (45).■

Proof of Proposition 4: When the solution is interior, from (47), the inequality $-\partial v^*/\partial\alpha > 0$ requires

$$(\beta' - 1)(y + \mu\Lambda) + (\beta - \alpha)\mu\Lambda_\alpha > 0 \tag{57}$$

where

$$\Lambda_\alpha = \frac{(\beta' - 1)\psi(1 - \psi) - (\beta - \alpha)(1 - 2\psi)\psi_\alpha}{[\psi(1 - \psi)]^2} \tag{58}$$

and where $\psi_\alpha = G + (1 - G)\beta' > 0$. Substituting from (58) in (57) implies that either $\beta' > 1$ (equivalently $\alpha < \overline{\alpha}$) or $\psi > 1/2$.■

Proof of Proposition 5: The first three claims follow from lemma 4 and the discussion in the text. To prove claim (iv), write welfare as

$$W(y, \alpha, \lambda) \equiv \int_{v^*}^{v_{\max}} (e + v - c)g(v)\, dv - \lambda y\psi(v^*, \alpha)$$

where $v^* = v^*(y, \alpha)$. Welfare is maximized subject to $y \geq 0$. The no-bonus policy is obviously optimal for λ sufficiently large. So suppose it is optimal for some λ_0 and let α_0 denote the optimal standard, i.e. $W(0, \alpha_0, \lambda_0) \geq W(y, \alpha, \lambda_0)$ for all y and α. First, observe that the no-bonus policy remains optimal for any $\lambda > \lambda_0$ because

$$W(0, \alpha_0, \lambda) = W(0, \alpha_0, \lambda_0) \geq W(y, \alpha, \lambda_0) > W(y, \alpha, \lambda), \text{ for all } y > 0 \text{ and } \alpha$$

The equality on the left follows from the fact that W does not vary with λ in a no-bonus policy; the strict inequality on the right, from the fact that W is decreasing in λ for any positive y. Now, if $y = 0$ and α_0 is optimal for λ_0, then assumption 1 implies that α_0 solves $-v^*_\alpha(0, \alpha_0) = 0$. Moreover, with respect to y, the following first-order condition must be satisfied

$$W_y(0, \alpha_0, \lambda_0) = (e + v^*(0, \alpha_0) - c)g(v^*(0, \alpha_0))\left(-v^*_y(0, \alpha_0)\right) - \lambda_0\psi(v^*(0, \alpha_0), \alpha) \leq 0 \tag{59}$$

As is obvious from (59), given that the first term in the middle expression is positive, there exists $\overline{\lambda} \leq \lambda_0$ satisfying $W_y(0, \alpha_0, \overline{\lambda}) = 0$ and such that

$W_y(0, \alpha_0, \lambda) > 0$ for all $\lambda < \overline{\lambda}$. Altogether, therefore, there exists $\overline{\lambda}$ such that the optimal policy is no-bonus if $\lambda \geq \overline{\lambda}$ and otherwise involves a positive bonus.■

Proof of Proposition 6:

(i) Let $R(v^*, \alpha) \equiv (\beta(\alpha) - \alpha)\Lambda(v^*, \alpha)$ denote the expected reputational benefit. In the no bonus policy, α solves $-v^*_\alpha(0, \alpha) = 0$. This is equivalent to α maximizing $R(v^*, \alpha)$ at $v^* = v^*(0, \alpha)$. Therefore,

$$R_\alpha(v^*, \alpha) \equiv (\beta'(\alpha) - 1)\mu\Lambda(v^*, \alpha) + (\beta - \alpha)\mu\Lambda_\alpha(v^*, \alpha) = 0 \qquad (60)$$

We assume the second-order condition holds strictly, i.e. $R_{\alpha\alpha}(v^*, \alpha) < 0$.

Introducing a small shift θ, we have

$$R_\alpha(v^*_\theta - \theta, \alpha) = (\beta'(\alpha) - 1)\mu\Lambda(v^*_\theta - \theta, \alpha) + (\beta(\alpha) - \alpha)\mu\Lambda_\alpha(v^*_\theta - \theta, \alpha) \qquad (61)$$

where v^*_θ is shorthand for $v^*(0, \alpha, \theta)$ and where, using lemma 2,

$$v^*_\theta = \theta + v^*\left(\frac{\theta}{\beta(\alpha) - \alpha}, \alpha, 0\right)$$

Thus, evaluated at $\theta = 0$,

$$\frac{d\alpha}{d\theta} = -\frac{R_{\alpha\theta}}{R_{\alpha\alpha}}$$

where, using (25),

$$R_{\alpha\theta} = \left[(\beta' - 1)\mu\Lambda_{v^*}(v^*, \alpha) + (\beta - \alpha)\mu\Lambda_{\alpha v^*}(v^*, \alpha)\right](-M(v^*, \alpha)) \qquad (62)$$

The expression inside the square brackets in (62) is equal to $\partial^2 R/\partial v^* \partial\alpha$. Recall that

$$\Lambda = \frac{(\beta - \alpha)\tau}{\psi(1 - \psi)}$$

so that

$$R = \frac{(\beta - \alpha)^2\tau}{\psi(1 - \psi)}$$

This yields

$$\begin{aligned} \frac{\partial R}{\partial \alpha} &= \frac{(\beta - \alpha)\tau}{\psi(1 - \psi)} \cdot \left[2(\beta' - 1) - \frac{(\beta - \alpha)(1 - 2\psi)\psi_\alpha}{\psi(1 - \psi)}\right] \\ &= \Lambda \cdot Q \end{aligned}$$

where Q denotes the expression in the square brackets. Thus,

$$\frac{\partial^2 R}{\partial v^* \partial \alpha} = \Lambda_v Q + \Lambda Q_v = \Lambda Q_v$$

where the last equality follows from the fact that (60) implies $Q = 0$.

Now

$$sign(Q_v) = -sign\left[\frac{\partial}{\partial v^*}\left(\frac{(1 - 2\psi)\psi_\alpha}{\psi(1 - \psi)}\right)\right]$$

One can show that

$$\frac{\partial}{\partial v^*}\left(\frac{(1-2\psi)\psi_\alpha}{\psi(1-\psi)}\right)=\frac{\{[\beta-\alpha\beta'](1-\psi)^2+[\beta'(1-\alpha)-(1-\beta)]\psi^2\}g}{\psi^2(1-\psi)^2}$$

where the expressions in both square brackets are positive owing to the strict concavity of β. Therefore, $Q_v < 0$ so that $R_{\alpha\theta} > 0$, which in turn implies that $d\alpha/d\theta$ is positive.

(ii) For the interior solution, let (y, α), both positive, maximize

$$W(y,\alpha) \equiv \int_{v^*(y,\alpha)}^{v_{\max}} (e+v-c)g(v)\,dv - \lambda y\psi(v^*(y,\alpha),\alpha)$$

Then, at (y, α),

$$W_y = (e+v^*-c-\lambda(\beta-\alpha)y)g(v^*)\left(-\frac{\partial v^*}{\partial y}\right)-\lambda\psi(v^*,\alpha)=0 \quad (63)$$

$$W_\alpha = (e+v^*-c-\lambda(\beta-\alpha)y)g(v^*)\left(-\frac{\partial v^*}{\partial \alpha}\right)-\lambda y\psi_\alpha(v^*,\alpha)=0 \quad (64)$$

We assume the second-order conditions hold strictly, i.e. $W_{yy} < 0$, $W_{yy}W_{\alpha\alpha} - W_{\alpha y}^2 > 0$.

Let $B(y, \alpha, v^*) \equiv (\beta(\alpha)-\alpha)(y+\mu\Lambda(v^*, \alpha))$ denote the total expected benefit from participation. Then the first-order conditions can be rewritten as

$$F_y \equiv \frac{W_y}{gM} = (e+v^*-c)B_y - \lambda[(\beta-\alpha)yB_y+\rho/M]=0 \quad (65)$$

$$F_\alpha \equiv \frac{W_\alpha}{gM} = (e+v^*-c)B_\alpha - \lambda y[(\beta-\alpha)B_\alpha+\hat{\rho}/M]=0 \quad (66)$$

where $\rho = \psi/g$ and $\hat{\rho} = \psi_\alpha/g$ are mixtures of reciprocal of hazard rates. At the solution, $F_{yy} < 0$, $F_{yy}F_{\alpha\alpha} - F_{\alpha y}^2 > 0$.

Introducing a small shift θ, the functions defined (65) and (66) are $F_y(y, \alpha, v_\theta^* - \theta)$ and $F_\alpha(y, \alpha, v_\theta^* - \theta)$, where v_θ^* is shorthand for $v^*(y, \alpha, \theta)$. At $\theta = 0$, and using lemma 2,

$$F_{y\theta} = B_y^2\mu\Lambda_v M - \lambda\frac{d(\rho/M)}{d\theta} \quad (67)$$

$$F_{\alpha\theta} = B_yB_\alpha\mu\Lambda_v M + (e+v^*-c)\frac{dB_\alpha}{d\theta} - \lambda y\left[(\beta-\alpha)\frac{dB_\alpha}{d\theta}+\frac{d(\hat{\rho}/M)}{d\theta}\right] \quad (68)$$

Substituting from (65) in (68) yields

$$F_{\alpha\theta} = B_yB_\alpha\mu\Lambda_v M + \lambda\left[\frac{(\rho/M)}{B_y}\frac{dB_\alpha}{d\theta} - y\frac{d(\hat{\rho}/M)}{d\theta}\right] \quad (69)$$

In an interior solution (implying that α is bounded away from zero), the terms multiplied by λ are bounded. Hence, for λ sufficiently small, the sign of both $F_{y\theta}$ and $F_{\alpha\theta}$ is given by the sign of Λ_v.

Differentiating the system of first-order conditions,

$$\frac{dy}{d\theta} = \frac{F_{\alpha y}F_{\alpha\theta} - F_{\alpha\alpha}F_{y\theta}}{F_{yy}F_{\alpha\alpha} - F_{\alpha y}^2} \tag{70}$$

$$\frac{d\alpha}{d\theta} = \frac{F_{\alpha y}F_{y\theta} - F_{yy}F_{\alpha\theta}}{F_{yy}F_{\alpha\alpha} - F_{\alpha y}^2} \tag{71}$$

We consider in turn the case where $\Lambda_v > 0$ and then the case $\Lambda_v < 0$.

a) The case $\Lambda_v > 0$

Then $F_{y\theta} > 0$ and $F_{\alpha\theta} > 0$. If $F_{\alpha y} \geq 0$, it immediately follows from (70) and (71) that both $dy/d\theta > 0$ and $d\alpha/d\theta > 0$. So suppose $F_{\alpha y} < 0$. We show that $dy/d\theta \leq 0$ and $d\alpha/d\theta \leq 0$ cannot simultaneously hold. From (71), $d\alpha/d\theta \leq 0$ implies

$$F_{\alpha y}F_{y\theta} - F_{yy}F_{\alpha\theta} \leq 0$$

Multiplying by $F_{\alpha y}/F_{yy}$, a positive quantity, yields

$$\frac{F_{\alpha y}^2 F_{y\theta}}{F_{yy}} \leq F_{\alpha y}F_{\alpha\theta} \tag{72}$$

From (71), $dy/d\theta \leq 0$ implies

$$F_{\alpha y}F_{\alpha\theta} \leq F_{\alpha\alpha}F_{y\theta} \tag{73}$$

Combining (72) and (73) yields

$$\frac{F_{\alpha y}^2 F_{y\theta}}{F_{yy}} \leq F_{\alpha\alpha}F_{y\theta}$$

or equivalently $F_{yy}F_{\alpha\alpha} \leq F_{\alpha y}^2$, which contradicts the strict second-order condition.

b) The case $\Lambda_v < 0$

Then $F_{y\theta} < 0$ and $F_{\alpha\theta} < 0$. If $F_{\alpha y} \geq 0$, then (70) and (71) imply $dy/d\theta < 0$ and $d\alpha/d\theta < 0$. For the case $F_{\alpha y} < 0$, we show that $dy/d\theta \geq 0$ and $d\alpha/d\theta \geq 0$ cannot simultaneously hold. Together, the inequalities imply

$$\frac{F_{\alpha y}F_{y\theta}}{F_{yy}} \leq F_{\alpha\theta} \leq \frac{F_{\alpha\alpha}F_{y\theta}}{F_{\alpha y}}$$

yielding $F_{yy}F_{\alpha\alpha} \leq F_{\alpha y}^2$, which again contradicts the second-order condition.

(iii) In the 'no type-1 error' policy, v^* maximizes $D(v^*) - C_N(v^*)$. Note that

$$\frac{d[k(v^*)(1 - G(v^*)]}{dv^*} = -g(v^*) \tag{74}$$

so that

$$C_N'(v^*) = -\lambda[k(v^*)(1 - G(v^*) + (c - v^*)g(v^*)]$$

Assuming an interior optimum (i.e. λ is not too large), the optimal v^* satisfies the first-order condition

$$-(e+v^*-c)g(v^*)+\lambda[k(v^*)(1-G(v^*)+(c-v^*)g(v^*)]=0$$

Equivalently, given $h \equiv (1-G)/g$,

$$Z(v^*) \equiv -[e+v^*-c-\lambda(c-v^*)]+\lambda k(v^*)h(v^*)=0 \tag{75}$$

The second-order condition is taken to hold strictly,

$$Z'(v^*)=-(1+\lambda)+\lambda\frac{d[k(v^*)h(v^*)]}{dv^*}<0 \tag{76}$$

Allowing for a shift θ,

$$Z(v^*,\theta)=-[e+v^*-c-\lambda(c-v^*)]+\lambda k(v^*-\theta)h(v^*-\theta)$$

Totally differentiating with respect to θ,

$$\frac{dv^*}{d\theta}=-\frac{Z_\theta(v^*,0)}{Z_v(v^*,0)}$$

where

$$Z_\theta(v^*,0)=-\lambda\frac{d[k(v^*)h(v^*)]}{dv^*}$$

The equality (74) can be rewritten as

$$k'(1-G)-kg=-g$$

equivalently

$$k'=\frac{(k-1)g}{1-G}=\frac{k-1}{h}$$

Therefore

$$\frac{d[k(v^*)h(v^*)]}{dv^*}=k'(v^*)h(v^*)+k(v^*)h'(v^*)=k(v^*)-1+k(v^*)h'(v^*) \tag{77}$$

It follows that $dv^*/d\theta<0$ if, and only if, the right-hand side of (77) is positive, which is the condition in part (iii) of the proposition.■

Proof of Corollary 1: Let the pair y and α refer to the optimal policy in the initial situation, with associated threshold $v^*=v^*(y,\alpha)$. The cost minimizing instruments at the same threshold, following a shift θ are denoted $\hat{\alpha}(v^*,\theta)$ and $\hat{y}(v^*,\theta)$. Then $\hat{\alpha}(v^*,0)=\alpha$ and $\hat{y}(v^*,0)=y$.

Define $K(v^*,\alpha)\equiv C(v^*,\alpha)/\lambda$ as in the proof of lemma 4. In an interior solution, the cost minimizing review standard satisfies

$$K_\alpha(v^*,\alpha)=\frac{(c-v^*)(\beta-\alpha\beta')}{(\beta-\alpha)^2}-\mu\frac{\partial[(\Lambda(v^*,\alpha)\psi(v^*,\alpha)]}{\partial\alpha}=0$$

We assume the second-order condition $K_{\alpha\alpha} > 0$ is strictly satisfied. Following a small shift θ, the change in the cost minimizing review standard is

$$\frac{\partial\hat{\alpha}}{\partial\theta} = -\frac{\partial K_\alpha(v^* - \theta, \alpha)/\partial\theta|_{\theta=0}}{K_{\alpha\alpha}(v^*, \alpha)} = \frac{K_{\alpha v}(v^*, \alpha)}{K_{\alpha\alpha}(v^*, \alpha)}$$

Define

$$\xi(v^*, \alpha) \equiv \frac{\partial[\Lambda(v^*, \alpha)\psi(v^*, \alpha)]}{\partial\alpha} \tag{78}$$

so that $K_{\alpha v}(v^*, \alpha) = -\mu\xi_v(v^*, \alpha)$. Writing Λ as in (35),

$$\begin{aligned} \xi &= \frac{[(\beta' - 1)(1 - \psi) + (\beta - \alpha)]\psi_\alpha]\tau}{(1 - \psi)^2} \\ &= \frac{(\beta - \alpha)\tau}{\psi(1 - \psi)} \cdot \left[\frac{\beta' - 1}{\beta - \alpha} + \frac{\psi_\alpha}{1 - \psi}\right]\psi \\ &= \Lambda S \end{aligned}$$

where

$$\begin{aligned} S &\equiv \left[\frac{\beta' - 1}{\beta - \alpha} + \frac{\psi_\alpha}{1 - \psi}\right]\psi \\ &= \left[\frac{\beta - \alpha\beta' + \beta' - 1}{\beta - \alpha}\right]\frac{\psi}{1 - \psi} \end{aligned}$$

where the last equality borrows (56) in lemma 4. The strict concavity of β implies that the expression inside the square brackets is positive for all $\alpha \in (0, 1)$. Therefore

$$S_v = \left[\frac{\beta - \alpha\beta' + \beta' - 1}{\beta - \alpha}\right]\frac{\psi_v}{(1 - \psi)^2} < 0$$

because $\psi_v = -(\beta - \alpha)g < 0$. Thus,

$$\xi_v(v^*, \alpha) = \Lambda_v(v^*, \alpha)S(v^*, \alpha) + \Lambda(v^*, \alpha)S_v(v^*, \alpha) \tag{79}$$

where the second term on the right-hand side is always negative while the sign of the first term depends on Λ_v. Define

$$\varepsilon_2 \equiv -\frac{\Lambda(v^*, \alpha)S_v(v^*, \alpha)}{S(v^*, \alpha)} > 0$$

It follows that $\xi_v(v^*, \alpha) < 0$, and therefore $\partial\hat{\alpha}/\partial\theta > 0$ if, and only if, $\Lambda_v(v^*, \alpha) < \varepsilon_2$.

Rewriting (39), the cost minimizing bonus $\hat{y}(v^*, \theta)$ satisfies

$$[\beta(\hat{\alpha}(v^*, \theta)) - \hat{\alpha}(v^*, \theta))[\hat{y}(v^*, \theta) + \mu\Lambda(v^* - \theta, \hat{\alpha}(v^*, \theta))] + v^* - c = 0$$

Therefore

$$\frac{\partial\hat{y}}{\partial\theta} = \mu\Lambda_v(v^*, \alpha) - \left[\frac{(\beta' - 1)(y + \mu\Lambda) + (\beta - \alpha)\mu\Lambda_\alpha}{\beta - \alpha}\right]\frac{\partial\hat{\alpha}}{\partial\theta}$$

Using the cost-minimization condition (48), this reduces to

$$\frac{\partial \hat{y}}{\partial \theta} = \mu \Lambda_v(v^*, \alpha) - \frac{y\psi_\alpha(v^*, \alpha)}{\psi(v^*, \alpha)} \frac{\partial \hat{\alpha}}{\partial \theta} \tag{80}$$

From the definition of ε_2 and the results concerning the sign of $\partial\hat{\alpha}/\partial\theta$, observe that $\partial\hat{y}/\partial\theta < 0$ if $\Lambda_v(v^*, \alpha) \leq 0$ and $\partial\hat{y}/\partial\theta > 0$ if $\Lambda_v(v^*, \alpha) \geq \varepsilon_2$. Now define

$$\varepsilon_1 \equiv \frac{y\psi_\alpha(v^*, \alpha)}{\mu\psi(v^*, \alpha)} \frac{\partial \hat{\alpha}}{\partial \theta}$$

so that $\partial\hat{y}/\partial\theta > 0$ if, and only if, $\Lambda_v(v^*, \alpha) > \varepsilon_1$. Combining this with the previous observations, it follows that $\varepsilon_1 \in (0, \varepsilon_2)$. Putting everything together, $\partial\hat{\alpha}/\partial\theta > 0$ and $\partial\hat{y}/\partial\theta < 0$ if $\Lambda_v(v^*, \alpha) < \varepsilon_1$, both $\partial\hat{\alpha}/\partial\theta > 0$ and $\partial\hat{y}/\partial\theta > 0$ if $\Lambda_v(v^*, \alpha) \in (\varepsilon_1, \varepsilon_2)$, while $\partial\hat{\alpha}/\partial\theta < 0$ and $\partial\hat{y}/\partial\theta > 0$ if $\Lambda_v(v^*, \alpha) > \varepsilon_2$.■

Appendix B

Table 1a: Cost minimizing policies when $\beta(\alpha) = 1 - (1-\alpha)^{\frac{1}{1-\gamma}}$

	v^*	α	$\beta(\alpha)$	$\psi(v^*, \alpha)$	$C_N(v^*)$	$C(v^*)$
$c = 12$						
$\gamma = .80$	8	.053	.240	.090	1.80	1.78
$\overline{\alpha} = .331$	6	.078	.330	.180	3.90	3.85
	4	0^+	0^+	0^+	6.80	6.80
	2	0^+	0^+	0^+	10.5	10.5
$\gamma = .90$	8	.112	.698	.230	1.24	0.96
$\overline{\alpha} = .226$	6	.146	.793	.405	3.07	2.45
	4	.170	.845	.575	5.69	4.91
	2	.197	.888	.750	9.11	8.63
$\gamma = .93$	8	.103	.806	.243	1.08	.645
$\overline{\alpha} = .176$	6	.129	.874	.427	2.83	1.88
	4	.151	.914	.609	5.37	4.04
	2	.182	.951	.797	8.71	7.34
$c = 8$						
$\gamma = .80$	6	.303	.835	.516	1.30	0.61
$\overline{\alpha} = .331$	4	.260	.778	.571	3.40	2.96
	2	0^+	0^+	0^+	6.30	6.30
$\gamma = .9$ $\overline{\alpha} = .226$	6	.134* .357*	.763 .988	.387 .611	1.02	0 0
	4	.238	.934	.655	2.84	1.32
	2	.263	.953	.815	5.27	4.11
$\gamma = .93$ $\overline{\alpha} = .176$	6	.067* .374*	.647 .999	.299 .624	0.94	0 0
	4	.194	.960	.654	2.69	0.53
	2	.217	.975	.823	5.23	3.09

Notes: $G(v) = v$ for $v \in [0, 10]$; $\mu = 1$; 0^+ denotes the no type-1 error policy.
* Two solutions, both relying solely on reputational incentives.

Table 1b: Cost minimizing policies when $\beta(\alpha) = \alpha^{1-\gamma}$

	v^*	α	$\beta(\alpha)$	$\psi(v^*, \alpha)$	$C_N(v^*)$	$C(v^*)$
$c = 12$						
$\gamma = .7$	8	.002	.158	.033	0.80	0.72
$\overline{\alpha} = .179$	6	.003	.166	.072	2.40	2.28
	4	.001	.140	.076	4.80	4.70
	2	$.000^\dagger$	.000	.000	8.00	7.96
$\gamma = .9$	8	.011	.635	.136	0.80	0.29
$\overline{\alpha} = .077$	6	.015	.661	.272	2.40	1.48
	4	.018	.661	.409	4.80	3.70
	2	.015	.656	.529	8.00	7.14
$\gamma = .95$	8	.009	.789	.165	0.80	0.10
$\overline{\alpha} = .043$	6	.013	.807	.330	2.40	1.08
	4	.018	.818	.489	4.80	3.07
	2	.021	.825	.664	8.00	6.35
$c = 8$						
$\gamma = .7$	6	.028	.342	.154	0.80	0.53
$\overline{\alpha} = .179$	4	.008	.239	.144	2.40	2.33
	2	.001	.134	.101	4.80	4.74
$\gamma = .9$ $\overline{\alpha} = .007$	6	$.002^*$ $.212^*$	.531 .856	.212 .470	0.80	0 0
	4	.042	.729	.454	2.40	1.14
	2	.031	.376	.571	4.80	3.81
$\gamma = .95$ $\overline{\alpha} = .043$	6	$.000^{*\dagger}$ $.300^*$	.526 .942	.211 .556	0.80	0 0
	4	.038	.849	.524	2.40	0.54
	2	.039	.850	.688	4.80	3.01

Notes: $G(v) = v$ for $v \in [0, 10]$; $\mu = 1$.

* Two solutions, both relying solely on reputational incentives;.

† An interior solution with $\alpha < 0.0005$.

References

[1] Adriani, F. and S. Sonderegger (2019) "A Theory of Esteem Based Peer Pressure" 115 Games and Economic Behavior 314-335.

[2] Ali, S. N. and R. Bénabou (2020) "Image versus Information: Changing Societal Norms and Optimal Privacy" 12 American Economic Journal: Microeconomics: 1–49.

[3] Becker, G. (1968) "Crime and Punishment: An Economic Approach" 76 Journal of Political Economy 169-217.

[4] Bénabou, R. and J. Tirole (2006) "Incentives and Prosocial Behavior" 96 American Economic Review 1652-1678.

[5] Bénabou, R. and J. Tirole (2010) "Individual and corporate social responsibility." 77 Economica 1-19.

[6] Bénabou, R. and J. Tirole (2011) "Laws and Norms" . National Bureau of Economic Research Working Paper 17579.

[7] Bernheim, B. D. (1994) "A Theory of Conformity" 102 Journal of Political Economy 841-877.

[8] Cooter, R. and A. Porat (2001) "Should Courts Deduct Nonlegal Sanctions from Damages?" 30 The Journal of Legal Studies 401-422.

[9] Demougin, D. and C. Fluet (2006) "Preponderance of Evidence" 50 European Economic Review 963-976.

[10] Demougin, D. and C. Fluet (2008) "Rules of Proof, Courts, and Incentives" 39 RAND Journal of Economics 20-40.

[11] Deffains, B. and C. Fluet (2020) "Social Norms and Legal Design" 36 Journal of Law, Economics, and Organization 139-169.

[12] Ellingsen, T., and M. Johannesson (2008) "Anticipated Verbal Feedback Induces Altruistic Behavior" 29 Evolution and Human Behavior 100-105.

[13] Glazer, A. and K. Konrad. (1996) "A Signaling Explanation for Charity" 86 The American Economic Review 1019-1028.

[14] Hosmer D.W. and S. Lemeshow S. (2000) Applied Logistic Regression, 2nd Ed., New York: John Wiley and Sons.

[15] Iacobucci, E. (2014) "On the Interactions between Legal and Reputational Sanctions" 43 Journal of Legal Studies 189-207.

[16] Ireland N.J. (1994) "On Limiting the Market for Status Signals" 53 Journal of Public Economics 91-110.

[17] Jewitt, I. (2004) "Notes on the 'Shapes' of Distributions" Unpublished.

[18] Kaplow, L. (2011) "On the Optimal Burden of Proof" 119 Journal of Political Economy, 1104-40.

[19] Mazyaki, A. and J. van der Weele (2019) "On Esteem-Based Incentives" 60 International Review of Law and Economics 105848.

[20] Mungan, M. (2020) "Justifications, Excuses, and Affirmative Defenses" 36 The Journal of Law, Economics, and Organization 343–377.

[21] Polinsky, A. M., and S. Shavell. (2007) "The theory of public enforcement of law" Handbook of Law and Economics, North Holland: Elsevier, 403-454 (A. M. Polinsky & S. Shavell, eds. 2007).

[22] Posner, R. A. (2007) "Economic Analysis of Law" New York: Aspen Publishers.

[23] Powers, D. (2011) "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation" 1 Journal of Machine Learning Technologies 37-63.

[24] Rasmusen, E. (1996) "Stigma and Self-Fulfilling Expectations of Criminality" 39 Journal of Law and Economics 519-543.

[25] Rizzolli, M., and M. Saraceno (2013) "Better That Ten Guilty Persons Escape: Punishment Costs Explain the Standard of Evidence." 155 Public Choice 395-411.

[26] Rubinfeld, D.L. and D.E. Sappington (1987) "Efficient Awards and Standards of Proofs in Judicial Proceedings" 18 RAND Journal of Economics 308-315.

[27] Scheinkman J. (2018) "Social Interactions (Theory)" in Macmillan Publishers Ltd (eds), *The New Palgrave Dictionary of Economics.* Palgrave Macmillan, London.

[28] Youden, W.J. (1950) "Index for Rating Diagnostic Tests" 3 Cancer 32-35.