Texas A&M University School of Law
Texas A&M Law Scholarship

Faculty Scholarship

6-2022

Laws and norms with (un)observable actions

Claude Fluet

Murat C. Mungan
Texas A&M University School of Law, mmungan@gmu.edu

Follow this and additional works at: https://scholarship.law.tamu.edu/facscholar

 Part of the Law and Economics Commons, and the Law and Society Commons

Recommended Citation

Claude Fluet & Murat C. Mungan, *Laws and norms with (un)observable actions*, 145 Eur. Econ. Rev. 104129 (2022).
Available at: https://scholarship.law.tamu.edu/facscholar/1822

This Article is brought to you for free and open access by Texas A&M Law Scholarship. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of Texas A&M Law Scholarship. For more information, please contact aretteen@law.tamu.edu.

European Economic Review 145 (2022) 104129

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

Laws and norms with (un)observable actions☆

Claude Fluet [a], Murat C. Mungan [b,*]

[a] Université Laval, Canada
[b] George Mason University, United States of America

ARTICLE INFO

Keywords:
Norms
Social concerns
Reputation
Esteem
Stigma
Signaling
Regulation

ABSTRACT

We analyze the interactions between social norms, the prevalence of acts, and policies when people cannot directly observe actors' behavior and must rely on noisy proxies. Norms provide ineffective incentives 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 leads to the opposite positive and normative conclusions with even modest errors. The review process through which public signals are provided is then an additional policy variable. When the cost of financing material rewards is high, it is optimal to rely solely on 'symbolic rewards' coupled with review standards that maximize reputational incentives, implying stricter criteria when there are weaker social norms. When material rewards are also used, review standards are stricter than that which would maximize reputational incentives.

1. Introduction

While making decisions, people are often motivated by how their behavior may impact their social standing. The decisions 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 one's material well-being. It is not surprising, therefore, that economists have formalized concepts like norms, reputation, esteem, and status through signaling models.[1] 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, which is a phenomenon at the heart of this literature.

An observation that emerged in prior work is that direct material incentives (whether provided through the law or private arrangements) and reputational incentives 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. 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. However, determining the optimal adjustments is not straightforward. This is because increasing direct material incentives may either increase

☆ For valuable comments and suggestions, we thank two anonymous referees, Jesse Bull, Ezra Friedman, Andreea Cosnita-Langlais, Carlos Oyarzun, Jennifer Reinganum, Sarath Sanga, Alexander Stremitzer, Bruno Strulovici, Abraham Wickelgren, and the participants of the 7th Annual Law and Economic Theory Conference, the 2021 ETH Zurich and University of Zurich Joint Law and Economics Seminar, the 2021 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. Claude Fluet acknowledges financial support from SSHRC Canada (grant 435-2013-1671).

* Corresponding author.
E-mail addresses: claude.fluet@fsa.ulaval.ca (C. Fluet), mmungan@gmu.edu (M.C. Mungan).

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

https://doi.org/10.1016/j.euroecorev.2022.104129
Received 30 May 2021; Received in revised form 1 March 2022; Accepted 6 March 2022
Available online 19 April 2022
0014-2921/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

or decrease reputational incentives, depending on whether behavior in the reputation game displays strategic complementarity or substitutability, which is, *a priori*, ambiguous.

Despite this ambiguity, a predictable relationship emerges between formal incentives and the equilibrium reputational incentives, when the propensities of individuals to engage in prosocial behavior are well-behaved, i.e., have a single-peaked distribution. This relationship is formalized within the honor-stigma model (Bénabou and Tirole, 2006, 2011; Adriani and Sonderegger, 2019) wherein third parties attach large stigma to rarely committed bad acts and honor good deeds when only few people have the courage or willingness to engage in them. 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 not directly observable by third parties, who need to rely on a public signal to form beliefs about actors. Many situations share this feature. Simple examples include courts rendering verdicts regarding defendants' behavior, certification agencies deciding whether to issue a certificate, academic committees determining whether to promote a scholar, and disciplinary committees deciding whether to reprimand an individual. In some of these cases the reviewing body has the ability to provide formal incentives (bonuses, an increase in salary, etc.), but in others they announce their decisions unaccompanied by a meaningful material reward or punishment (e.g., employee of the month awards, nominal damages of $1, or even the Légion d'honneur). A commonality across these contexts is that the public signal is noisy, because the reviewer has to determine, based on imperfect evidence, whether the agent is worthy of a reward or punishment. Sometimes the reviewer is free to choose the standard that it uses to make these determinations (e.g., an academic committee with broad discretion), but in other contexts the standard is pre-determined by a higher authority (e.g., courts bound to use a specific standard of proof).

We construct a unifying model which accommodates all the possibilities mentioned above. Specifically, we consider a setting wherein agents decide whether to engage in prosocial behavior. Third parties do not directly observe this behavior but only whether an agent received a reward, a public signal which they use as a proxy for his behavior. These proxies may be perfectly informative or noisy indicators of agents' behavior. Moreover, they may be produced through an exogenous process or through a review process determined by a principal. Third parties use this information to form opinions about agents, which is what ultimately generates reputational incentives. We use this model to both revisit the findings of prior results in the literature, and to investigate the properties of optimal review standards and formal incentives when the principal is free to choose its standard.

A number of key results emerge from our analysis. As a preliminary matter, we show that the positive as well as the normative conclusions of the honor-stigma model do not hold up when acts are unobservable, and the unobservability of acts often implies the opposite of these conclusions. For instance, the honor-stigma model's implication that extreme (i.e. very high or low) participation enhances reputational incentives is overturned when inferences about acts rely on noisy signals, and this has similar implications regarding the relationship between the size of optimal rewards and the frequency of acts.

Our analysis also provides a rationale for symbolic rewards. When reputational concerns matter, increasing people's incentives to engage in prosocial behavior through formal incentives is not cost justified when it involves significant financing costs. In these cases, it becomes optimal to use review standards geared towards maximizing reputational incentives. By contrast, when the cost of financing formal incentives is not too large, both formal and reputational incentives are used to encourage prosocial behavior. The bonus and review standard are chosen so as to minimize the social cost of the achieved level of prosocial behavior. The optimal review standard is then such that relaxing the standard would increase reputational incentives, i.e., a weaker review standard would increase participation in the socially valuable activity. Reputational incentives are not maximized because the optimal policy takes into account the fact that increasing participation is socially costly due to the cost of financing the material rewards. Typically, the optimal review standard is then a significantly demanding one, except possibly when engaging in the socially valuable activity is a common phenomenon, i.e., rewarding a majority of agents turns out to be the best policy.

Finally, shifts in the populations' intrinsic motivations often cause the optimal formal rewards to move in opposite directions when acts are observable versus unobservable. For instance, when rewards are noisy signals and participation in the prosocial activity is large, an improvement in intrinsic motivation reduces reputational incentives, which needs to be compensated by larger monetary incentives. When rewards are purely symbolic, the optimal review standards are less demanding when people are intrinsically more motivated, i.e., when there are stronger prosocial norms. This is because laxer review standards and higher participation are complements in maximizing reputational incentives; a complementarity which has important consequences. It implies, for instance, that an academic committee ought to apply stricter promotion criteria as the institution's research culture becomes weaker.[2]

We relegate more detailed descriptions and discussions of these results to the remaining parts, which are organized as follows. In the next section, we briefly discuss how this article relates to the prior literature on laws and norms, the principal–agent literature with unobservable actions, as well as the related experimental literature. In Section 3, we present a principal–agent model wherein agents commit acts that emit exogenously determined signals received by third parties, who adjust their interactions with agents based on these signals. We compare how reputational incentives and the equilibrium participation in the prosocial act interact with each other when signals received by third parties are perfect and imperfect, respectively. In Section 4, we derive optimal rewards in this setting. In Section 5, we formalize the processes through which evidence observed by the principal is generated, and explain its relationship to the review processes implementable by the principal. In Section 6, we allow the principal to choose the review standard it uses to provide formal incentives and derive optimal review standards and rewards. We analyze shifts in intrinsic motivations in Section 7, and conclude in Section 8. All proofs are in the Appendix.

[2] We thank an anonymous referee for suggesting this example.

2. Literature review

Our article builds on the honor-stigma model of Bénabou and Tirole (2006, 2010, 2011) wherein people make inferences about others based on their behavior. Prosocial behavior gives rise to positive inferences, and thus is associated with an 'honor' conveyed by third parties onto the actor, and conversely shirking from prosocial behavior carries 'stigma'. Thus, one consideration that affects a person's decision to act prosocially is the reputational net benefit from doing so, which is the difference between the values one attaches to the honor and the stigma, respectively. How these reputational incentives interact with formal incentives as well as social norms has played a central role in the literature (Rasmusen, 1996; Bénabou and Tirole, 2006, 2010, 2011; Iacobucci, 2014; Mungan, 2016; Mazyaki and van der Weele, 2019; Ali and Bénabou, 2020). This question is important, because optimal formal sanctions may require adjustments based on the size of reputational incentives (e.g., damages in torts or contracts cases as in Cooter and Porat (2001).

A change in social norms as well as formal incentives affect reputational considerations, because they change the composition, or average characteristics, of people who engage in prosocial behavior versus those who shirk, and thereby lead to a change in the value of the honor and stigma associated with each behavior, respectively. The responses in reputational incentives are driven by how the size (or measure) and characteristics of prosocial actors compare to those of shirkers. As emphasized by Adriani and Sonderegger (2019), the discrete nature of feasible acts plays an essential role in this respect, because some degree of pooling always arises at equilibrium.[3] Thus, an important factor that affects responses in reputational incentives is the distribution of individuals' propensities to commit prosocial acts, i.e., their intrinsic motivations. As shown in Bénabou and Tirole (2011), which builds on Jewitt (2004), when this distribution is single-peaked reputational incentives are U-shaped in the prevalence of the prosocial act. This allows Bénabou and Tirole to categorize acts based on their typical prevalence (e.g., "normal", "modal", and "heroic") and interpret the reputational incentives that accompany acts based on their categories (see Section 3.1 for a more detailed discussion). Adriani and Sonderegger (2019) extend Bénabou and Tirole (2011), and note that the opposite conclusions hold when the distribution of individuals' propensities are U-shaped. These observations also extend to comparative statics regarding individuals' behavior: a change in social norms causes the equilibrium threshold to move in opposite directions when the type distribution is U-shaped and inverse U-shaped, respectively.[4]

Quite importantly, the honor-stigma model considers cases where behavior is observable. This is somewhat surprising, given Rasmusen's 1996 prior formalization of stigma in a law enforcement context where crimes are imperfectly detected and sanctioned. Thus, Rasmusen's model incorporates one type of enforcement error (wrongful acquittal) while ignoring the other type (wrongful conviction), and shows how multiple equilibria may emerge as a result of the interdependency of stigma and expectations regarding crime rates.[5] However, the strand of the law enforcement literature following Rasmusen (1996) does not systematically study interactions between reputational incentives and equilibrium participation as in Bénabou and Tirole (2011) and Adriani and Sonderegger (2019).

The supplemental material in Adriani and Sonderegger (2019) is the only work that we are aware of which discusses unobservable actions as we do in this paper. We complement Adriani and Sonderegger (2019) in several ways. We show that when monitoring errors are not too small, reputational incentives are always inverse U-shaped, regardless of the underlying type distribution. Additionally, reputational incentives are always negligible with extreme participation rates, and become larger as the participation in prosocial behavior moves towards more moderate levels. Moreover, we show that unobservability of actions not only cuts against the positive implications of Bénabou and Tirole (2011) but against their normative conclusions as well; for instance, Pigouvian rewards (or subsidies) move in opposite directions when acts are observable and unobservable, respectively (see Section 4).

We obtain these results by considering exogenously determined signals received by third parties regarding the behavior of individuals. Subsequently, we endogenize the errors associated with these signals by building on the literature on principal–agent problems with unobservable actions as well as the law enforcement literature on standards of proof. Our analysis is related to the risk-neutral principal–agent model when the agent's liability is limited. It is well known that the most efficient contractual form is then a dichotomic bonus-based contract with respect to the signal obtained about the agent's action, e.g., a noisy performance score (Park, 1995; Kim, 1997). In particular, when the limited liability constraint is binding, the agent should be penalized in the sense of not receiving the bonus whenever a more favorable signal realization could have been observed (Demougin and Fluet, 1998). This typically leads to a large bonus, infrequently awarded. We show that this result must be substantially qualified when reputational considerations are introduced. Our analysis also relates to the law enforcement literature. The legal literature has discussed at length the risk of legal error and regulatory mistakes. In the economics or law and economics literature, the focus has often been on how optimal standards may deviate from the strong 'beyond a reasonable doubt' or the weaker 'preponderance of the evidence', the typical standard of proof used in civil cases (Posner, 2007; Demougin and Fluet, 2005, 2006; Kaplow, 2011; Rizzolli and Saraceno, 2013; Mungan, 2011, 2020). However, that literature does not discuss the choice of standards of proof in terms of the reputational consequences of convictions.

We also relate to existing work on the expressive role of the law (e.g., Sunstein, 1996; Cooter, 1998; Posner, 2000; McAdams, 2000). This literature points out that people respond not only to the size of the sanction imposed for violations of the law, but also to the information conveyed through the illegalization of the conduct in question. Expressive law theories are centered on

[3] This contrast with environments with a continuous action space, which may allow for full separation of types.

[4] This is a simplified summary. See Proposition 1 in Adriani and Sonderegger (2019) for a more precise description of this result.

[5] Later, Mungan (2017) extends Rasmusen's model to incorporate wrongful convictions to show that they reduce deterrence.

the observation that people not only care about formal sanctions, but also the degree to which others approve or disapprove of their behavior. Some of these theories, like McAdams' (2000) attitudinal theory of expressive law, also notes that the legislative process through which laws are made can inform people of what types of behavior are likely to gain the approval and disapproval of others (see also Bénabou and Tirole, 2011). Some experimental evidence provides support for this type of dynamic. Galbiati et al. (2010), for instance, provides evidence that an experimenter's choice to sanction non-cooperation may act as a signal of poor cooperation among other participants. This type of dynamic is absent in honor-stigma models, like ours, wherein subjects make inferences regarding shirkers' and cooperators' characteristics based on a known distribution of intrinsic motivations. Nevertheless, our article adds to this literature by formalizing the idea that courts and principals can supply additional information to third parties by rendering verdicts (or other judgments in different contexts) regarding people's otherwise unobservable behavior.

A related prior literature focuses on how formal sanctions may crowd out intrinsic or other motivations, and may in some cases even lead to the opposite of their intended consequences (e.g., (Gneezy and Rustichini, 2000; Bénabou and Tirole, 2006; Fehr and Schmidt, 2007), (Herold, 2008). Some theoretical explanations for this type of negative interdependence is similar to what we focus on here: an increase in formal incentives can reduce the difference in the average characteristics of people who shirk and who act prosocially (e.g. Bénabou and Tirole, 2006, 2011). Other explanations have also been offered. Herold (2008), for instance, notes that a principal who trusts her agent, i.e., holds a strong belief that he is intrinsically motivated for the task, may have to refrain from using formal incentives to signal that she trusts the agent to boost their productivity. Like Herold (2008), our analysis also suggests that it may be inefficient to use formal incentives, but for a different reason. In our context, this result is driven by cost-minimization considerations: because reputational considerations provide incentives that deter shirking to some extent, the additional gains from using formal incentives may not be cost-justified.

3. Model

Agents choose a discrete action $a \in \{0,1\}$ where 1 has social value in the sense that it improves welfare relative to action 0. The prosocial act involves a resource cost of c for the agent and generates a positive externality e to be shared equally by all agents. In addition, the agent has intrinsic motivations, such that he receives a pay-off of v from the act. Agents differ 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 support $[v_{\min}, v_{\max}]$ and continuously differentiable density $g(v) > 0$. The mean is denoted $\bar{v}$.

The principal incentivizes the agents by paying a reward or bonus of y for the prosocial act, based on a decision rule which leads it to incorrectly offer the reward with probability α (type-1 error or 'false positives') and to incorrectly refrain from rewarding with probability $1-\beta$ (type-2 error or 'false negative'), with $\beta > \alpha$. Thus, an agent who engages in the act receives the reward with probability β; otherwise, the reward is received with probability α. In this section and Section 4, we take the error rates as given, and we endogenize them in Sections 5–7.

The final component which affects an agent's incentives reflects his reputational concerns. Third parties value intrinsic motivations and make inferences regarding the agents' motivations by observing whether or not they 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, $E[\widetilde{v}|b]$ denotes third parties' expectations of the agent's type, conditional on whether or not a bonus was awarded. From the agent's perspective, engaging in act a causes this estimate to have an expectation of $E[E[\widetilde{v}|b]|a]$. Below, we explain how this expected value is related to the principal's and other agents' actions.

Given this set-up, the preferences of a type-v agent are represented by

$$U = (v + \beta y - c)a + \alpha y(1-a) + e\bar{a} + \mu E[E[\widetilde{v}|b]|a], \quad a \in \{0,1\} \tag{1}$$

where $\bar{a}$ is the proportion of agents engaging in the socially valuable action and μ is a positive parameter denoting the importance of reputation relative to other considerations (i.e., v, y and c).

3.1. Reputational incentives

As a first step, we derive explicit expressions for reputational gains or losses. From (1), the utility from act $a = 1$ is increasing in v. Thus, in equilibrium, individuals with intrinsic valuations above some threshold v^* engage in the act and those with lower valuations do not.

Given this threshold strategy, the measure of individuals who receive a reward is

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

while $1-\psi$ agents are not rewarded. The conditional expected type below and above the cut-off is denoted by

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

Using Bayes' rule, we then obtain

$$E[\widetilde{v}|b=1] \equiv \frac{\alpha G(v^*)\mathcal{M}^-(v^*) + \beta(1-G(v^*))\mathcal{M}^+(v^*)}{\psi(v^*,\alpha,\beta)}$$

Similarly,

$$E[\widetilde{v}|b=0] = \frac{(1-\alpha)G(v^*)\mathcal{M}^-(v^*) + (1-\beta)(1-G(v^*))\mathcal{M}^+(v^*)}{1-\psi(v^*,\alpha,\beta)}$$

The difference between the conditional expectations will play a key role and is denoted

$$\Lambda \equiv E[\widetilde{v}|b=1] - E[\widetilde{v}|b=0]$$

Substituting for the terms on the right-hand side then yields

$$\Lambda(v^*, \alpha, \beta) = \frac{(\beta - \alpha)(1 - G(v^*))G(v^*)}{\psi(v^*, \alpha, \beta)(1 - \psi(v^*, \alpha, \beta))} \Delta(v^*) \tag{3}$$

where

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

is the difference in the expected type conditional on engaging in the act versus not.

We note that Λ multiplied by the importance of reputation, that is $\mu\Lambda$, is the reputational benefit associated with receiving a reward versus not. If rewards were conferred without error, i.e., $\alpha = 1 - \beta = 0$, the reputational benefit would be $\mu\Delta$.

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. 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)

As noted, Bénabou and Tirole make their categorization in a setting where acts are directly observable. Thus, there is no need for third parties to rely on a public signal to form opinions, which is equivalent to the case where $\alpha = 1 - \beta = 0$. This setting is presumably a good fit for analyzing situations where relevant third parties are close to the actors, or where the actor's behavior is widely observed and verifiable without error (e.g., participation trophies). However, in many circumstances the relevant third parties (e.g., society at large) cannot directly observe acts and must rely on noisy public signals; for instance, Salvage rewards in maritime law conferred to people who place themselves at risk to save others or Good Samaritan awards offered by many organizations to recognize acts of kindness. Many other examples can be formulated when incentives are provided through punishment instead of rewards (e.g., criminal convictions, public reprimands), although our focus is on rewards throughout our analysis.

When third parties need to rely on noisy public signals, rewards do not perfectly indicate whether the recipient has intrinsic motivations above or below the equilibrium cut-off. For instance, if a large fraction of agents participates in the prosocial action, third parties hold strong priors that any given individual must have engaged in the prosocial act. The non-conferral of a reward works against this prior, but because it is subject to error, it may not significantly overturn it. The greater the participation rate, the stronger are third parties' priors, and thus the lower the stigma generated through the non-conferral of a reward. Similarly, when few agents participate in the prosocial action, rewards will be infrequent but it no longer follows that receipt of a reward confers much honor.

To disentangle the different effects, we rewrite (3) as

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

where

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

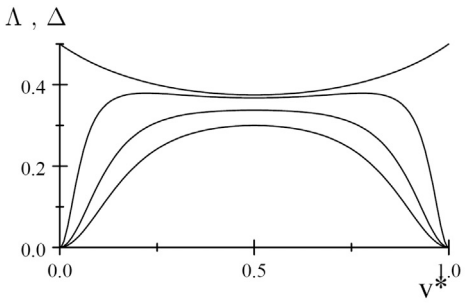
is a measure of the predictive value of rewards regarding the agents' behavior. The predictive value depends on the discriminatory power of the reward process, as captured by the type-1 and 2 errors, and on the prevalence of the prosocial act. The value is between zero and unity, with $\delta(v^*, 0, 1) = 1$ when conferral and non-conferral of a reward are perfectly informative about the agent's action.

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 $v_{\max}$.*

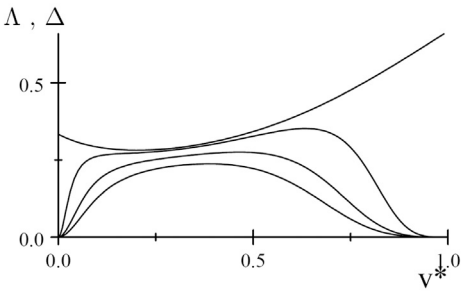
It follows that the reputational gain as defined in (3) is the product of two functions, and these have opposite properties when the distribution of types is strictly unimodal with an interior mode. The following proposition summarizes the implications.

Proposition 1. *(i) When the distribution of types is strictly unimodal with an interior mode and rewards are conferred without error (i.e., $\alpha = 1 - \beta = 0$), the reputational benefit equals the difference in the average intrinsic value of actors and non-actors, i.e., $\mu\Lambda = \mu\Delta$, and is quasiconvex in v^* with a unique interior minimum.*

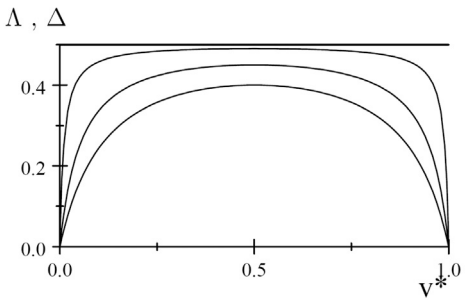
(ii) When rewards serve as noisy signals of acts (i.e., $\alpha > 0$ and $\beta < 1$), the reputational benefit $\mu\Lambda(v^*, \alpha, \beta) < \mu\Delta(v^*)$ and never has an interior minimum. When $\beta - \alpha$ is not too large, Λ is quasiconcave in v^* with a unique interior maximum.



a: Symmetric uni-modal $m = n = 2$.



b: Right-skewed uni-modal $m = 2$ and $n = 4$.



c: Uniform $m = n = 1$.

Fig. 1a-c. $\Delta(v^*)$ and $\Lambda(v^*, \alpha, \beta)$ under different Beta distributions.

The results reveal the contrast between the way reputational benefits interact with the prevalence of the act when acts are observable or only imperfectly so. The opposite of the interactions described in the literature is obtained when rewards are sufficiently noisy signals of the agents' behavior, with noisiness expressed in terms of the probability differential $\beta - \alpha$, a standard measure of accuracy in dichotomous discrimination tests.[6] Moreover, results obtained with noisy signals hold regardless of the type distribution.

Lemma 1 implies that when participation rates are extreme, a move towards even more extreme participation rates always reduces reputational incentives, as opposed to increasing them as in the observable acts case. This causes the *respectable/admirable/modal* categorization quoted above to lose its accuracy. Second, when bonuses are not awarded very accurately, the relationship between the commonality of the act and reputational incentives has the opposite characteristics compared to the observable acts case.

This latter observation raises the question of what happens for a reasonable range of non-extreme v^* and for reasonably small type-1 and type-2 errors. We provide numerical examples by plotting reputational effects as a function of v^* when types belong to the unit interval and follow a Beta distribution with parameters $m, n \geq 1$. We consider three illustrative cases: the symmetric unimodal ($m = n = 2$), the right-skewed unimodal ($m = 2 < n = 4$), and the uniform distribution ($m = n = 1$). 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 act.[7]

The Figs. 1a to 1c plot $\Delta(v^*)$ and $\Lambda(v^*, \alpha, \beta)$ where the latter is computed with symmetric type-1 and 2 errors, i.e., $\alpha = 1 - \beta = \varepsilon$ where $\varepsilon \in \{0.01, 0.05, 0.1\}$. These values are chosen because they either correspond to standard significance levels or would be considered acceptable in a dichotomous diagnostic test. The exercise reveals that even small error rates (5% is sufficient in all examples below) cause the relation between reputational benefits and the commonality of the act to be quasi-concave, the opposite of what would occur with perfectly observable acts.[8] These observations naturally have normative implications (e.g., vis-à-vis optimal subsidies) which we discuss after characterizing the equilibrium.

[6] Accuracy can also be written as $1 - (\alpha + 1 - \beta)$, i.e., as one minus the sum of type-1 and 2 errors. See Youden (1950).

[7] With a uniform distribution on the unit interval, $\Delta(v^*) = \frac{1}{2}$ for all v^*.

[8] In the figures 1a and 1b, $\Lambda(v^*, \alpha, \beta)$ is not quasiconcave with an error rate of 1%. In 1a, it has two local maxima.

3.2. Equilibrium

To characterize the equilibrium threshold, we start with the best response of an agent to the other agents' behavior profile as described by v^*. Thus, we express the type-v agent's utility as $U = U(a, v^*, v)$ and note that the prosocial act is weakly preferred if $U(1, v^*, v) \geq U(0, v^*, v)$, which corresponds to the condition:

$$v + (\beta - \alpha)(y + \mu\Lambda(v^*, \alpha, \beta)) - c \geq 0 \tag{5}$$

Engaging in the act increases the probability of receiving a reward from α to β. In (5), the probability differential multiplies the reward as well as the reputational benefit.

Given (5), a perfect Bayesian equilibrium is characterized by a cut-off intrinsic value v^* solving

$$\varphi(v, y, \alpha, \beta) \equiv v + (\beta - \alpha)(y + \mu\Lambda(v, \alpha, \beta)) - c = 0 \tag{6}$$

An interior solution $v^* \in (v_{\min}, v_{\max})$ exists if

$$\varphi(v_{\min}, y, \alpha, \beta) < 0 < \varphi(v_{\max}, y, \alpha, \beta) \tag{7}$$

When $\alpha > 0$ and $\beta < 1$, the reputational gain vanishes when v^* approaches the boundaries of the support of types, so that (7) is equivalent to

$$v_{\min} < c - (\beta - \alpha)y < v_{\max} \tag{8}$$

We assume this condition holds. In our subsequent analysis, the reward and the type-1 and 2 errors will be optimally chosen and (8) will always hold. To ensure that the equilibrium is unique, we impose[9]:

Assumption 1. $\varphi_v(v^*, y, \alpha, \beta) = 1 + (\beta - \alpha)\mu\Lambda_v(v^*, \alpha, \beta) > 0$ for all $v^* \in [v_{\min}, v_{\max}]$.

The foregoing condition holds, for instance, when the reputation parameter μ is not too large.

An important question is whether the incentives to engage in prosocial acts increase or decrease as more people participate. Totally differentiating (6),

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

A unit increase in y increases by $\beta - \alpha$ the expected material reward from engaging in act. Taking others' behavior as given, an individual's best-response threshold would therefore decrease by the same amount. However, as a result of social interactions, the total effect on everyone's behavior is (minus) the change in expected reward multiplied by the social multiplier[10]

$$M(v^*, \alpha, \beta) \equiv \frac{1}{1 + (\beta - \alpha)\mu\Lambda_v(v^*, \alpha, \beta)} \tag{9}$$

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 benefits and formal incentives reinforce or mitigate each other depends on how reputational benefits change with the participation rate.

Proposition 1 describes how these interactions depend on whether or not acts are perfectly observable. Figs. 1a-c depict a wide range of high participation rates (small values of v^*) for which $\Lambda_v > 0$ when acts are inferred from noisy rewards, but where $\Lambda_v = \Delta_v < 0$ when rewards reveal acts perfectly (or when acts are directly observable). In these examples, for high participation rates the social multiplier with noisy rewards is smaller than what would emerge absent reputational concerns, and the opposite is true when rewards are non noisy. Similar observations can be made with respect to low participation rates.

4. Optimal rewards with exogenous errors

To complete the preceding section, we describe how rewards should be set. Welfare is defined as the sum of agents' equilibrium expected utilities net of the cost of financing the bonuses. This is expressed as

$$W = \bar{U} - (1 + \lambda)y\psi \tag{10}$$

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 (10),

$$W = \int_{v^*}^{v_{\max}} (v + e - c)g(v)dv - \lambda y\psi(v^*, \alpha, \beta) + \mu\bar{v} \tag{11}$$

[9] We use subscripts, i.e., φ_v and Λ_v, and similarly in other expressions to denote partial derivatives.

[10] To quote (Scheinkman, 2018, 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."

where the equilibrium threshold satisfies (6), i.e., $v^* = v^*(y)$. The first term in the right-hand side of (11) is the net direct benefit from the prosocial 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. Specifically, applying Bayes' rule,

$$\begin{aligned}&\psi(v^*,\alpha,\beta)E[\widetilde{v}|\ \ b=1\ \ ]+(1-\psi(v^*,\alpha,\beta))E[\widetilde{v}|\ \ b=1\ \ ]\\&=G(v^*)\mathcal{M}^-(v^*)+(1-G(v^*))\mathcal{M}^+(v^*)=\bar{v}\end{aligned}$$

That the signaling of types is a zero-sum game follows from the assumption that reputational benefits are linear in third parties' beliefs.[11]

In a first-best world agents should participate in the prosocial 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

$$v_{FB}=c-e \tag{12}$$

We make the following assumptions, to ensure that the welfare maximization problem (and subsequent versions of this problem) is well behaved.

Assumption 2. $v_{\min}+e<c<v_{\max}$ and $e>\max_v \mu\Delta(v)$.

Some agents participate in the prosocial act even without material or reputational incentives (those with $v>c$). Moreover, the assumption implies that $v_{FB}\in(v_{\min},v_{\max})$, i.e., in the first best some agents participate in the prosocial act but not all of them. Positive bonuses are required to implement the first best, in the sense that relying solely on reputational incentives is not sufficient. Because $\Lambda(v)\leq\Delta(v)$, the same is true when rewards are noisy. We maintain Assumption 1 in the remainder of our analysis.

Consider first the special case where rewarding agents involves no social cost, i.e., $\lambda=0$. The second term in (11) then vanishes and the optimal reward solves $v^*(y)=c-e=v_{FB}$. Substituting (12) into (6) yields

$$y=\frac{e}{\beta-\alpha}-\mu\Lambda(v_{FB},\alpha,\beta) \tag{13}$$

Welfare is at the first-best level

$$W_{FB}\equiv\int_{v_{FB}}^{v_{\max}}(v+e-c)g(v)dv+\mu\bar{v}$$

When rewards involve no social costs, errors in rewarding agents do not impact the level of welfare obtainable. However, they dilute reputational incentives as well as the impact of formal incentives, which causes the optimal reward to be positively related to both types of errors. Thus, Pigouvian subsidies are greater when acts are unobservable. Moreover, with substantial errors, the relationship between the first-best threshold, v_{FB}, and the optimal reward is the opposite as that obtained when acts are observable per proposition 1. Noting that agents' material cost of acting, c, increases the first-best threshold without affecting anything else, we plot the optimal reward against c to provide an illustration of these facts.[12]

In Fig. 2 optimal subsidies with unobservable and observable acts are depicted by the thick and thin lines, respectively. The former lies above the latter reflecting that optimal subsidies in the unobservable case are greater. The figure also illustrates that as the target level of participation in the act moves towards the extremes, the optimal subsidy is decreasing when acts are observable, and the opposite is true when acts are signaled with significant errors. This is a direct corollary of Lemma 1 and Proposition 1: extreme participation rates translate into larger (resp. smaller) reputational incentives when acts are observable (resp. unobservable), and thus smaller (resp. larger) subsidies are needed to get the first-best result.

When $\lambda>0$,

$$\frac{\partial W}{\partial y}=[e+v^*-c-\lambda(\beta-\alpha)y]g(v^*)\left(\frac{-\partial v^*}{\partial y}\right)-\lambda\psi \tag{14}$$

where v^* is short-hand for the equilibrium threshold $v^*(y)$. The first term on the right-hand side of (14) 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 a fraction β of inframarginal agents (those with $v>v^*(y)$) as well as the bonuses erroneously paid to a fraction α of non-participating agents (those with $v<v^*(y)$).

When λ is not too large, an interior solution $y>0$ obtains satisfying the first-order condition $\partial W/\partial y=0$. Welfare is less than the first best because participation in the prosocial act is reduced, i.e. $v^*(y)>v_{FB}$ which follows from the first-order condition, and also because of the deadweight loss of financing rewards. Errors in the reward process reduce maximum welfare. With smaller errors, rewards can be more precisely targeted; in particular, smaller errors will yield larger reputational consequences. Thus, for implementing any target v^*, greater reputational incentives can substitute for socially costly rewards.

[11] This has been the standard assumption in the literature, e.g., Daughety and Reinganum (2010), Bénabou and Tirole (2006, 2011). Rasmusen (1996) shows that the zero-sum structure emerges when wages set in a competitive market (in later interactions with third parties) depend on the individuals' reputation. See Deffains and Fluet (2020) for an alternative framework where reputation has social value.

[12] We use examples where intrinsic motivations follow the Beta distribution with parameters $m=n=2$, $e=1$, $\mu=0.5$, and $\beta=0.9$, $\alpha=0.1$ in the unobservable act case.

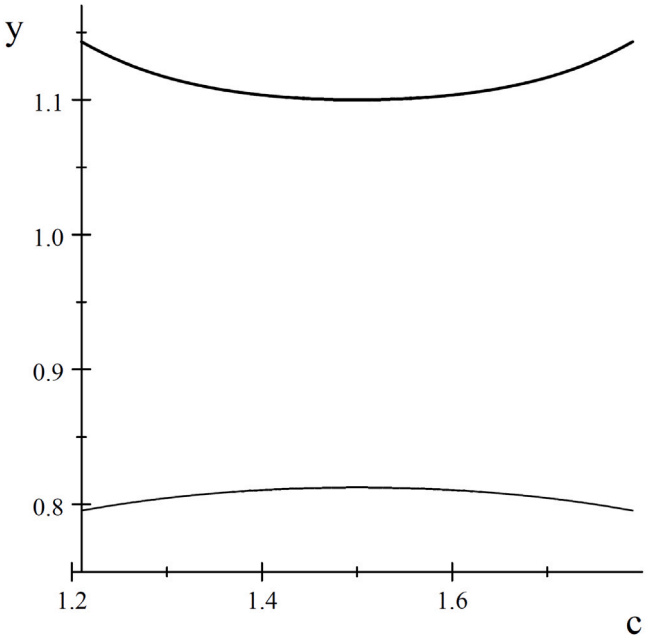


Fig. 2. Pigouvian subsidies.

5. Review process

We now assume that the principal is able to choose the type 1 and 2 error pair, subject to some informational constraints. In this section we describe the feasible set of error pairs as a function of these informational constraints, and how these relate to review standards and agents' incentives. In the next section, we use these observations to discuss the properties of optimal review standards.

5.1. Decision rule

Various reports and assessments are received about an individual's actions. We summarize this information by the random variable $\widetilde{x}$ (not necessarily scalar-valued) with continuous density functions $p_a(x)$, $a \in \{0,1\}$, positive over the same support. On this basis, the principal makes a binary decision, whether to reward the individual or not. When rewards are socially costly, the principal wants the type 2 error to be as small as possible for any level of type 1 error. An efficient decision rule is then for the principal to award a bonus if the signals about the individual satisfy $p_1(x) > \kappa p_0(x)$ and not to award any bonus if $p_1(x) < \kappa p_0(x)$, where κ is a critical value that depends on the size of type 1 error that the principal is willing to tolerate. This leads to a maximum value of β for any chosen α, which we write as the function $\beta(\alpha)$.[13] This function is increasing and concave, with $\beta(0) = 0$, $\beta(1) = 1$, and $\beta(\alpha) > \alpha$ otherwise.[14] Because the density functions have the same support, $\beta(\alpha) < 1$ for all $\alpha \in (0,1)$. For tractability we assume that $\beta(\alpha)$ is strictly concave and twice differentiable unless stated otherwise.[15]

To illustrate, the set of review policies that the principal may choose from is summarized by the error pairs along an increasing and concave curve such as the ones depicted in Fig. 3. Each curve plots the 'true positive rate' against the 'false positive rate', given the informativeness of reports and assessments. In the figure, $\beta(\alpha) = \alpha^{1-\gamma}$ for three different values of $\gamma \in (0,1)$, which are chosen so that the middle and top curves are such that symmetric type-1 and 2 errors[16] obtain when $\alpha = 0.1$ and $\alpha = 0.05$ respectively, which connects with the numerical examples in Fig. 1; for the bottom curve, corresponding to the least informative review process, symmetric errors obtain at $\alpha = 0.3$. As a useful typology (see, for instance, Hosmer and S. Lemeshow (2000), a dichotomous test would be considered as 'acceptable' if the area under the curve is between 0.7 and 0.8, as 'good' if the area is between 0.8 and 0.9, and as excellent if it is over 0.9. In this view, our examples range from good for the bottom curve to excellent for the two top ones.

[13] In principle, one can implement a type 1 (resp. type 2) error greater than α (resp. $1-\beta$) given β (resp. α). This would amount to disregarding some information conveyed by the signal. However, as will become clear from our discussion in Section 4, both types of error reduce social welfare, and we therefore restrict our attention to policies summarized by $\beta(\alpha)$.

[14] The pair (α,β) is feasible if there exists a function $\phi(x) \in [0,1]$ such that $\alpha = \int \phi(x)p_0(x)\,dx$ and $\beta = \int \phi(x)p_1(x)\,dx$, where $\phi(x)$ is the probability of rewarding the agent conditional on the signal. The set of feasible pairs is then easily seen to be convex, implying that its upper boundary, denoted $\beta(\alpha)$, is a concave function. See, e.g., Lehmann and Romano (2005).

[15] The only case where we consider a β function which is not twice differentiable is to construct the example in Figs. 4c, below.

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

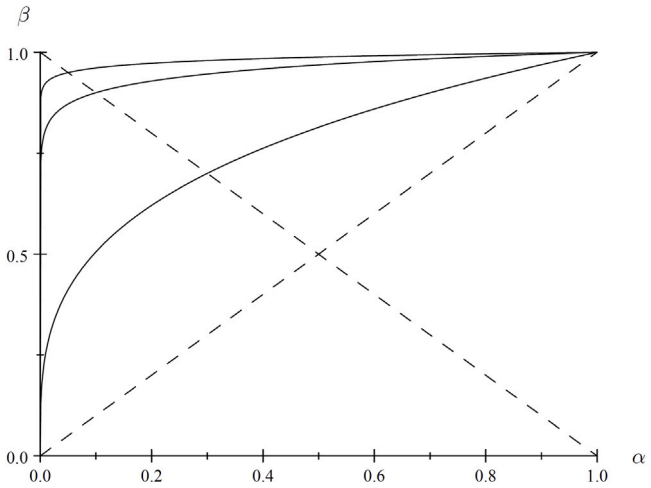


Fig. 3. β as a function of α.

5.2. Reputational and material 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 prosocial 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 by eliminating β as an argument of functions, e.g., we write $\Lambda(v^*,\alpha)$ and $\psi(v^*,\alpha)$.

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

$$\beta'(\overline{\alpha}) = 1 \tag{15}$$

The review process associated with $\overline{\alpha}$ is an important benchmark. In the notation of the preceding subsection, it corresponds to a critical κ equal to unity, i.e., the individual is rewarded when $p_1(x) > p_0(x)$, equivalently when the reports received are more likely to be produced about someone who acted prosocially rather than not.[17] 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

$$\mu\Lambda(v^*,\alpha) = \frac{\beta(\alpha)-\alpha}{\psi(v^*,\alpha)(1-\psi(v^*,\alpha))}\mu\tau(v^*) \tag{16}$$

where

$$\tau(v^*) \equiv (1-G(v^*))G(v^*)\Delta(v^*) \tag{17}$$

is the part of the reputational benefit that does not depend on α. Observe that the numerator of (16) 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 2. *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

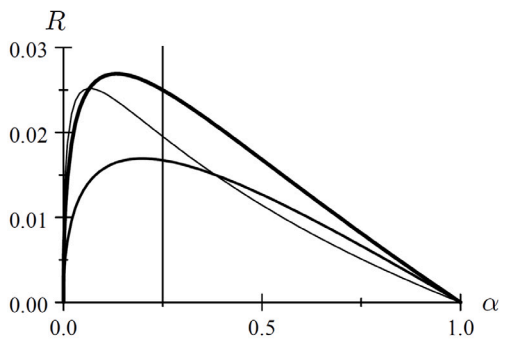
$$\beta = \alpha^{1-\gamma} \text{ and } \beta = 1-(1-\alpha)^{\frac{1}{1-\gamma}},\ \gamma \in (0,1)$$

In both cases, γ captures the level of informativeness of the review process.[18] In Figs. 4a and 4b, $\gamma = 0.5$ for both functions. We plot the expected reputational benefit

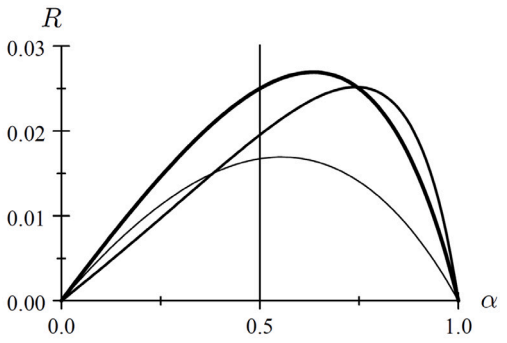
$$R(v^*,\alpha) \equiv (\beta(\alpha)-\alpha)\Lambda(v^*,\alpha) \tag{18}$$

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

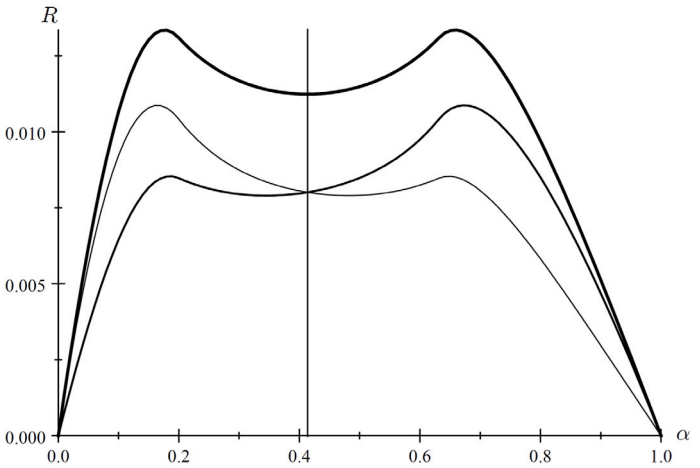
[18] For any level of the type-1 error, β is larger the larger the value of γ.



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



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



c: Expected reputational benefit and $\overline{\alpha}$ when β' is piecewise linear.

Fig. 4a-c. $R(v^*, \alpha)$ under different functional forms for β.

for the cases where v^* equals 0.3 (*medium*), 0.5 (*thick*), and 0.7 (*thin*). The distribution of types $G(v)$ is the unimodal symmetric Beta distribution with parameters $m = n = 2$. The vertical straight line identifies $\overline{\alpha}$.

These examples illustrate how reputational benefits respond to changes in the review standard, as captured by α, taking the participation rate as given. 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 prosocial activity. The converse holds in the right-hand side figures. Despite this difference, in both cases the review standard that maximizes the expected reputational benefit is stricter (a smaller α) when fewer agents participate in the prosocial act (a larger v^*).

The figures also illustrate that for a wide range of functional forms for β, expected reputational benefits are single peaked in the review standard. We illustrate that this is not a general property. In the next example, the β function is such that β' is piecewise linear with β'' small for intermediate values of α.[19] The expected reputational benefit curves then possess 'camel humps' as depicted in Fig. 4c, where v^* equals 0.3 (*medium*), 0.5 (*thick*), and 0.7 (*thin*) as in our previous examples.

In the middle range of type 1 errors, $\beta(\alpha) - \alpha$ is nearly constant. Hence, the curvature of the expected reputational benefit then depends primarily on the variance of the reward distribution, i.e., the denominator of (16), thereby producing a local minimum in

[19] Specifically, we use the following function:

$$\beta(\alpha) = \begin{cases} \beta_1(\alpha) \equiv m\alpha - 4\alpha^2 & \text{if} \quad \alpha \leq 0.2 \\ \beta_2(\alpha) \equiv 0.156 + (m - 1.56)\alpha - 0.1\alpha^2 & \text{if} \quad \alpha \in (0.2, \tilde{\alpha}] \\ \beta_3(\alpha) \equiv 1 - \beta_2^{-1}(1-\alpha) & \text{if} \quad \alpha \in (\tilde{\alpha}, 1 - \beta_1(0.2)] \\ \beta_4(\alpha) \equiv 1 - \beta_1^{-1}(1-\alpha) & \text{if} \quad \alpha > 1 - \beta_1(0.2) \end{cases}$$

where $m = \beta'(0) = \sqrt{4.3376}$ and $\tilde{\alpha} = 5(m - 2.56)$ are chosen such that $\beta'(\tilde{\alpha}) = 1$ and $\tilde{\alpha} = 1 - \beta(\tilde{\alpha})$. This causes $\beta'(\alpha)$ to not be differentiable at $\alpha \in \{0.2, \tilde{\alpha}, 1 - \beta(0.2)\}$.

this range. Since reputational benefits vanish as α approaches the edges, there are two local maxima. Nevertheless, stricter reviews standards maximize $R(v^*, \alpha)$ as fewer people participate in the prosocial act, even when jumps in the maximizing standards are possible. We formalize this result, as follows.

Proposition 2. *Let* $\alpha_1 \in \arg\max_\alpha R(v_1^*, \alpha)$ *and* $\alpha_2 \in \arg\max_\alpha R(v_2^*, \alpha)$*, where* $v_1^*, v_2^* \in (v_{\min}, v_{\max})$*. Then* $\alpha_2 < \alpha_1$ *if* $v_2^* > v_1^*$.

In the proof of the proposition, we show that $R_{v\alpha}(v^*, \alpha) < 0$ when $R_\alpha(v^*, \alpha) = 0$. When the expected reputational benefit is strictly quasiconcave as in Figs. 4a and 4b, the former condition is sufficient to ensure that the maximizing α is unique and strictly decreasing in v^*. That it is decreasing in v^*, however, holds more generally.[20] The intuition is that, in maximizing the expected reputational benefit, one trades off the direct incentive effect of a change in α (i.e., the change in the net probability $\beta(\alpha) - \alpha$ of getting the symbolic reward when one participates versus not) against the effect the change in α has on the noisiness of the reward/no reward signal, as captured by its variance $Y \equiv \psi(1 - \psi)$. The relative change in noisiness, Y_α/Y, is monotonically increasing in v^*. As a result, for maximizing the reputational impact of rewards, low participation in the prosocial act and demanding review standards go hand in hand. Whether or not an optimal policy should seek to maximize reputational incentives is of course another matter, which we discuss next.

6. Optimal review standards and rewards

We now extend our welfare analysis to the case where both rewards and review standards are determined endogenously, i.e., the optimal policy maximizes $W(y, \alpha)$. Writing the equilibrium threshold as $v^*(y, \alpha)$, the welfare effect of a change in the bonus is as in (14) and the effect of a change in the type-1 error is

$$\frac{\partial W}{\partial \alpha} = [e + v^* - c - \lambda(\beta - \alpha)y]g(v^*)\left(\frac{-\partial v^*}{\partial \alpha}\right) - \lambda y \psi_\alpha \tag{19}$$

where $\psi_\alpha(v^*, \alpha) = G(v^*) + (1 - G(v^*))\beta'(\alpha)$ and

$$-\frac{\partial v^*}{\partial \alpha} = \frac{(\beta' - 1)(y + \mu\Lambda) + (\beta - \alpha)\mu\Lambda_\alpha}{1 + (\beta - \alpha)\mu\Lambda_v} \tag{20}$$

The numerator in the right-hand side of (20) can be rewritten as

$$(\beta'(\alpha) - 1)y + \mu R_\alpha(v^*, \alpha)$$

The first term relates to formal incentives and is the change in the expected bonus for participating in the prosocial act, following a marginal increase in α. 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. The effect on the equilibrium threshold is (minus) the increase in total expected benefits times the social multiplier.

6.1. Exogenous bonus

As a start, to parallel Section 4, it is instructive to analyze the optimal review standard for an exogenously given material reward. For instance, a committee determines the standards for a scientific award but it has no say on the amount of the prize.

When $y = 0$, the award is purely symbolic. Therefore, the second term on the right hand side of (19) disappears, and by Assumption 1, the expression inside the square brackets in (19) is positive for all α. Therefore, α should be increased so long as this increases participation in the prosocial activity. The optimal review standard then solves

$$-\frac{\partial v^*}{\partial \alpha} = \left(\frac{1}{1 + (\beta - \alpha)\mu\Lambda_{v^*}}\right)\mu R_\alpha(v^*, \alpha) = 0$$

and maximizes the expected reputational benefit at the achieved participation threshold. For later use, denote by v_0 the participation threshold under the optimal symbolic reward.

When $y > 0$, increasing participation through a change in the review standard is socially costly because of the cost of financing rewards. When λ is not too large, we get an interior solution $\alpha > 0$ satisfying the first-order condition obtained by equating $\partial W/\partial \alpha$ as expressed in (19) to zero. In the solution, because $\psi_\alpha > 0$, participation must be increasing in α, i.e., marginally weakening the review standard would increase participation. This requires that the numerator in (20) is positive, equivalently

$$(\beta'(\alpha) - 1)y + \mu R_\alpha(v^*, \alpha) > 0$$

Thus, either $\alpha < \overline{\alpha}$ or the expected reputational benefit is increasing in α, or both. Whether the expected reputational benefit is itself increasing depends on the size of the exogenous bonus. We summarize our findings as follows.

[20] More formally, it follows from the proof that $R(v^*, \alpha)$ satisfies the single-crossing property: if $R(v^*, \alpha'') \geq R(v^*, \alpha')$ for $\alpha'' > \alpha'$, then $R(v^{**}, \alpha'') \geq R(v^{**}, \alpha')$ for all $v^{**} < v^*$. We thank Bruno Strulovici for this observation.

Proposition 3. *In the welfare maximizing policy with an exogenous y:*

(i) when $y = 0$, the optimal α maximizes the expected reputational benefit;

(ii) when $y > 0$ and λ is not too large, participation in the prosocial act is increasing in α; moreover, if $\partial W/\partial y \leq 0$ at the solution, the expected reputational benefit is also itself increasing in α; if $\partial W/\partial y > 0$, it may be either increasing or decreasing.

Part (i) of the proposition summarizes our observation that when rewards are symbolic, it is optimal to use a review standard that maximizes reputational incentives, since these are the only incentives available. If the reputational incentive is coupled with a bonus, we call it 'large' if an exogenous decrease in the bonus would increase welfare (i.e., when $\partial W/\partial y < 0$). In this case, the conferral of a bonus is partly wasteful in that it uses more resources than is desirable. In addition, the desired level of participation can be obtained with a relatively small α, which is a region where the expected reputational benefit is increasing. Conversely, when an exogenous increase in the bonus would increase welfare (i.e., $\partial W/\partial y > 0$), the given bonus is 'small'. When it is very small, the expected material reward will have little bite and the solution will be close to the zero bonus case where the review standard maximizes expected reputational incentives. Otherwise, there may be cases where the review standard is close to that which maximizes the expected material reward (i.e., α slightly less than $\overline{\alpha}$, see Eq. (15)), which is not inconsistent with a decreasing expected reputational benefit as illustrated in Fig. 4a.

6.2. Endogenous bonus and review standard

We start with some observations. First, the participation threshold v_0 under the optimal purely symbolic reward is achieved at no cost. Therefore, one cannot do worse than this policy when costly bonuses are allowed and can be optimally chosen.

Secondly, the optimal participation threshold will be in the interval $[v_{FB}, v_0]$. Achieving any $v^* < v_0$ requires a positive bonus. The bonus and the review standard are then chosen to minimize the cost of achieving v^*, which is the deadweight loss of financing bonuses defined as

$$C(v^*, y, \alpha) \equiv \lambda y \psi(v^*, \alpha)$$

where y and α satisfy the equilibrium condition

$$v^* + (\beta(\alpha) - \alpha)(y + \mu\Lambda(v^*, \alpha)) - c = 0 \tag{21}$$

Given that the required bonus is positive, we can substitute for y from Eq. (21) and write the deadweight loss as

$$C(v^*, \alpha) = \frac{\lambda(c - v^*)\psi(v^*, \alpha)}{\beta(\alpha) - \alpha} - \lambda\mu\Lambda(v^*, \alpha)\psi(v^*, \alpha) \tag{22}$$

The first term on the right-hand side is the deadweight cost of rewards that would arise when individuals have no reputational concerns. The second term is the cost savings due to such concerns. Both terms are strictly increasing in α,

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

and

$$\frac{\partial\left(\lambda\mu\Lambda(v^*, \alpha)\psi(v^*, \alpha)\right)}{\partial\alpha} = \frac{\lambda\mu\tau(v^*)[(1 - \alpha)\beta' - (1 - \beta)]}{(1 - \psi)^2} > 0$$

where the inequalities follow from the strict concavity of $\beta(\alpha)$.

Thirdly, the foregoing raises the possibility that minimizing costs requires an arbitrarily small α, which in turn (from the equilibrium condition (21)) implies an arbitrarily large y. This would arise when the savings due to reputational concerns, as the review standard is weakened, are always insufficient to compensate for the increase in the first term of (22). Loosely speaking, the optimal policy is then a 'no type-1 error' policy with a very large bonus.[21] In this policy, reputational incentives play no role because the second term in (22) vanishes.

In what follows, we abstract from the possibility that the optimal policy is 'no type-1 error' because it refers to situations where reputational concerns are irrelevant. The following assumption is sufficient.

Assumption 3. For all $v^* \in [v_{FB}, v_0)$,

$$\frac{\mu G(v^*)(1 - G(v^*))\Delta(v^*)}{c - v^*} > \frac{-\beta''(0)}{2(\beta'(0) - 1)^3} \tag{23}$$

The above condition is satisfied if $\beta'(\alpha)$ becomes sufficiently large as α goes to zero. In other words, the review process has enough discrimination power at very demanding standards. This ensures that $C_\alpha(v^*, 0) < 0$, so the cost minimization policy is characterized by some positive α; see the discussion in the Appendix. Henceforth, we maintain Assumption 3.

The optimal policy depends on the trade-off between the marginal costs and benefits from targeting a particular participation rate. This trade-off depends crucially on the shadow cost of material rewards, λ.

[21] Strictly speaking, the minimization problem then has no solution, but one could introduce an exogenous maximum feasible bonus assumed to be very large. This constraint would be binding and allow a very small α.

Proposition 4. *There exists a critical value $\bar{\lambda} > 0$ such that:*

(i) when $\lambda > \bar{\lambda}$ the optimal solution is a 'symbolic reward' policy with $y = 0$ where α maximizes the expected reputational benefit;

(ii) when $\lambda < \bar{\lambda}$ the optimal solution is an interior solution, i.e., $y > 0$ and $\alpha > 0$, such that the expected reputational benefit is strictly increasing in α. Moreover, α is smaller than that which would maximize the expected reputational benefit at the achieved threshold.

In either case, the optimal type-1 error is either smaller than $\bar{\alpha}$, or rewards are frequent, i.e., $\psi(v^*, \alpha) > \frac{1}{2}$.

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 negligible monetary values compared to the reputational rewards conferred) for which we have listed some examples in the introduction.

However, purely symbolic rewards are sub-optimal when the marginal cost of financing material rewards is sufficiently small. In the interior solution, y and α minimize the cost of the achieved participation threshold and satisfy[22]

$$\frac{\partial v^*(y,\alpha)/\partial \alpha}{\partial v^*(y,\alpha)/\partial y} = \frac{y\psi_\alpha(v^*,\alpha)}{\psi(v^*,\alpha)} \tag{24}$$

Expliciting the left-hand side, (24) is equivalent to

$$\frac{(\beta'-1)(y+\mu\Lambda)+(\beta-\alpha)\mu\Lambda_\alpha}{\beta-\alpha} = \frac{y\psi_\alpha(v^*,\alpha)}{\psi(v^*,\alpha)} \tag{25}$$

which can be simplified to yield

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

The right-hand side is positive due to the strict concavity of $\beta(\alpha)$; hence, the expected reputational benefit is increasing in α in an interior solution. This must be true, since otherwise one could lower costs by using a more demanding review standard (and possibly combine it with smaller bonuses) to implement the targeted participation threshold.

To illustrate, in Fig. 4a the expected reputational benefit is maximized at some $\alpha < \overline{\alpha}$ for various thresholds v^*. Therefore, the optimal review standard will also be below $\overline{\alpha}$: people are rewarded only if it is sufficiently more likely than not that they participated in the prosocial act. In Fig. 4b, by contrast, the expected reputational benefit is maximized at $\alpha > \overline{\alpha}$. In an interior solution, one cannot then exclude the possibility that the optimal policy involves a review standard laxer than $\overline{\alpha}$. If so, it corresponds to policies where rewards are frequent, in the sense that a majority of individuals are rewarded. People are then *not* rewarded only if there is sufficient evidence that they did *not* engage in the prosocial act.

The proposition also states that, in an interior solution, the review standard is always stricter than any standard that would maximize the expected reputational benefit. For instance, suppose the reputational benefit curve at the achieved participation threshold is the thick curve of Fig. 4c with two global maxima. Then, the optimal α is smaller than that which corresponds to the first global maximum.

7. Shifts in norms

So far, we analyzed optimal policies given the distribution of intrinsic motivations. These policies naturally need to be adjusted when the characteristics of the population change. For instance, when citizens acquire a greater sense of moral responsibility towards protecting the environment, should environmentally friendly behavior be rewarded more or less? Should the review standard be more or less demanding?

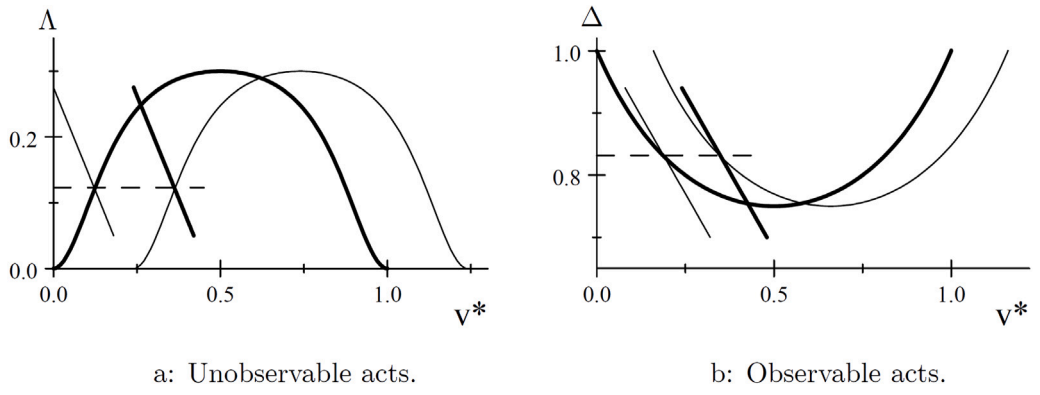
7.1. Equilibrium effects

We first describe the effects on the equilibrium participation threshold, taking the policy as given. The changes that we consider are uniform shifts in the distribution of types. A rightward shift means that individuals are on average more intrinsically motivated. Because dispersion does not change, such shifts have a simple interpretation in terms of exogenous changes in norms.[23] We let $G(v-\theta)$ be the original distribution shifted to the right when θ is positive. The reputational benefits are then given by the same functions as previously but taking the shift into account, i.e., the benefits are $\Delta(v^*-\theta)$ or $\Lambda(v^*-\theta)$ where in the latter we omit reference to type-1 and 2 errors.

An implication is that a positive shift θ has the same effect on the equilibrium reputational benefits as an identical increase in the expected bonus, i.e., a bonus increase equal to $\theta/(\beta-\alpha)$. We illustrate this graphically via Figs. 5a and 5b for the cases where acts are unobservable and observable, respectively.

[22] The equation follows directly from the first-order conditions $\partial W/\partial y = \partial W/\partial \alpha = 0$. It is also the condition for an interior solution to $\min_{y,\alpha} C(v^*, y, \alpha)$ subject to (21).

[23] See Bénabou and Tirole (2011) and in particular Adriani and Sonderegger (2019). The latter also considers other changes in the distribution of types, e.g., greater dispersion.



a: Unobservable acts. b: Observable acts.

Fig. 5a-b. Shifts in norms and an 'equivalent' change in the bonus.

In both figures, the thicker curves are associated with the 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.[24] 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 and as a function of the reputational benefit Λ. The initial equilibrium is the intersection of both curves. The thin reputational curve is $\Lambda = \Lambda(v^* - \theta)$. In Fig. 5a, the shift increases v^* and yields a smaller reputational benefit at equilibrium. Starting from the initial situation, an increase in the bonus to $y' = y + \theta/(\beta - \alpha)$ corresponds to a leftward 'equivalent' shift in the best response line. This reduces v^* and results in the same drop in the equilibrium reputational benefit as the shift in the distribution of types. In either case, the *proportion* of agents participating in the prosocial act is the same, which explains why reputational benefits are the same.

To formalize, we modify the equilibrium condition in (6) to incorporate the possibility of a shift in the distribution of types,

$$\varphi(v^*, y, \theta) \equiv v^* + (\beta - \alpha)[y + \mu\Lambda(v^* - \theta)] - c = 0 \tag{26}$$

such that the equilibrium threshold can be expressed as $v^*(y, \theta)$ solving (26). Our previous observations can now be formalized as follows.

Lemma 3. $v^*(y, \theta) = \theta + v^*\left(y + \frac{\theta}{\beta - \alpha}, 0\right)$.

We then have

$$\left.\frac{\partial v^*(y,\theta)}{\partial \theta}\right|_{\theta=0} = 1 + \frac{\partial v^*(y,0)}{\partial y}\frac{1}{\beta - \alpha} = 1 - M(v^*(y,0)) \tag{27}$$

where $M(v^*)$ is the social multiplier as defined in (9). As noted in Section 3, whether the social multiplier is larger or smaller than unity depends on whether acts are observable and whether the equilibrium threshold is small or large. Thus, an immediate implication of (27) is that the same factors also determine the direction of the impact of changes in social norms on the equilibrium threshold.

Proposition 5. *Let v^*_{in} be the initial equilibrium. Following a small shift $G(v - \theta)$:*

(i) v^* decreases (resp. increases) if $v^*_{in} < \arg\min \Delta(v^*)$ (resp. $>$), when acts are observable and the distribution of types is strictly unimodal; and

(ii) v^* increases (resp. decreases) if $v^*_{in} < \arg\max \Lambda(v^*)$ (resp. $>$), when acts are unobservable and $\beta - \alpha$ is not too large.

The proposition highlights the contrast between observable and unobservable acts: the equilibrium participation threshold moves in opposite directions for small as well as large participation rates, provided the type-1 and 2 errors are not too large.

7.2. Policy effects

When the type-1 and 2 errors are exogenous, the Pigouvian subsidy (for the case where $\lambda = 0$) is given by Eq. (13), which we rewrite as

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

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

Thus,

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

The change in the optimal bonus results from a pure substitution effect between formal and reputational incentives in implementing the same target threshold v_{FB}. Therefore, the Pigouvian subsidy and the equilibrium threshold move in the same direction. We note this as follows.

Corollary 1. *For given type-1 and 2 errors, the Pigouvian subsidy moves in the same direction as the impact of the shift in norms on the equilibrium threshold, i.e., $\partial v^*(y,0)/\partial\theta$, described in Proposition 5.*

Proposition 5 and its corollary together highlight the pivotal role that the observability of acts plays in determining how optimal subsidies respond to changes in norms. For instance, with observable acts, starting with a very large participation rate, the optimal subsidy ought to be decreased in response to people becoming intrinsically more motivated. The opposite is true, meaning that the subsidy ought to be increased, when acts can be inferred only with substantial noise.

When $\lambda > 0$, the change in the optimal subsidy depends on the 'reputation effect' of the shift in norms as described above, but it will also be affected by possibly countervailing factors. Rewriting (14), the first-order condition in the initial situation is

$$[e + v^*_{in} - c - \lambda(\beta - \alpha)y_{in}]g(v^*_{in})\left(\frac{-\partial v^*}{\partial y}\right) = \lambda\psi(v^*_{in})$$

The expression in the square brackets is the net social benefit of the marginal prosocial act. This increases or decreases depending on the sign of $\partial v^*/\partial\theta$. However, whether the bonus should be increased or reduced also depends on how the shift in norms affects ψ/g, which tracks the size of the population receiving bonuses relative to the measure of the individuals who are just on the margin, a 'hazard rate effect'.[25] The latter may reinforce or work against the pure reputation effect. For the case where acts are observable, Bénabou and Tirole (2011) show that the reputation effect dominates if $\left|\Delta(v^*_{in})\right|$ is bounded away from zero and λ is sufficiently small. The same argument applies with respect to $\left|\Lambda_v(v^*_{in})\right|$ when rewards are noisy signals. The effect of a shift in norms on the optimal bonus is then as in corollary 1.

Next, we discuss two results that do not depend on λ being small. Recall that the optimal policy is a purely symbolic reward when λ is sufficiently large. The optimal response to changes in norms is then obtained through the following corollary to Propositions 2 and 3.

Corollary 2. *When the initial optimal policy is a purely symbolic reward, the optimal policy change following a small positive shift $G(v-\theta)$ involves an increase in α.*

Shifts in norms then have straightforward effects because they impact a single policy tool and the effect does not depend on the sign of Λ_v at the initial solution. An increase in the population's sentiments towards the prosocial act causes the optimal review standard policy to become 'laxer'. This is because, as noted in Proposition 3, the optimal policy maximizes the expected reputational benefit, and as noted in Proposition 2, the corresponding review standards is laxer when there is more participation in the prosocial act. Thus, intrinsic motivations and symbolic rewards act as complements, and an increase in the former causes the latter to be utilized more generously.

When the initial optimal policy involves a positive bonus, the sign of Λ_v at the initial solution determines whether incentives would need to be increased or reduced to maintain the same participation threshold. Because costs are minimized in the optimal solution, we know that incentives are marginally increasing in both y and α. An interesting question is then whether both instruments move in the same direction following a shift in norms. We show that this is not necessarily true by considering a case where the cost minimizing pair to achieve an initial equilibrium threshold v^*_{in} is unique, regular, and is denoted by $\hat{y}(v^*_{in},\theta) = y_{in}$ and $\hat{\alpha}(v^*_{in},\theta) = \alpha_{in}$. Then we have the following.

Corollary 3. *If $\Lambda_v(v^*_{in},\alpha_{in}) \leq 0$, then $\partial\hat{y}(v^*_{in},0)/\partial\theta < 0 < \partial\hat{\alpha}(v^*_{in},0)/\partial\theta$.*

The intuition behind this corollary becomes more apparent by focusing on a case where initially there is little participation, Λ is single peaked, and an increase in intrinsic motivations increases reputational incentives, since $\Lambda_v \leq 0$. Thus, the increase in intrinsic motivations reduces the relative cost of providing incentives through a larger α compared to a larger y (the right hand side of the cost minimization condition (25)) while also increasing the relative efficiency of α versus y (the left hand side of the cost minimization condition). This can be interpreted as bonuses becoming a more blunt tool compared to the review standard in affecting incentives. Effects similar to that in Corollary 3 may also arise in the optimal policy with endogenous participation threshold; for instance, when $\Lambda_v \leq 0$ and the optimal threshold is relatively unresponsive to the change in social norms. It may be remarked that similar effects also arise following an exogenous increase in reputational concerns, i.e., an increase in the reputation parameter μ. Borrowing from the proof of Corollary 3, it is easily seen that $\partial\hat{y}/\partial\mu < 0 < \partial\hat{\alpha}/\partial\mu$.

[25] $\psi/g = \alpha(G/g) + \beta[(1-G)/g]$ is a weighted sum of the reciprocals of hazard rates.

8. Concluding remarks

People are often driven by reputational considerations when deciding whether to commit acts that benefit others. How these reputational concerns are related to the prevalence of the act is an important question which has normative implications. We have demonstrated here that the nature of these interactions is quite sensitive to how third parties form the beliefs that they use to honor or stigmatize actors. When third parties can observe acts and form beliefs about others based on these acts, as shown in Bénabou and Tirole (2011), norms may emerge which attach large reputational consequences to acts that are committed by a very large or a very small fraction of actors. On the other hand, when third parties must rely on noisy public signals to form opinions, we have shown that the opposite result is obtained. Similarly, 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, which naturally affects normative conclusions (e.g., the relationship between Pigouvian subsidies and external factors). These results highlight the importance of correctly identifying the source of the information used by people in forming opinions about others (e.g., observable acts or bonuses/penalties) when thinking about reputational incentives.

When third parties rely on rewards whose conveyance is subject to errors, the review standard used to determine reward recipients becomes an important policy variable, because they affect the size of both formal and reputational incentives. The review standard that maximizes accuracy (i.e., minimizes the sum of type-1 and type-2 errors) also maximizes formal incentives, and is thus independent of the prevalence of the act. The same is not true for the review standard that maximizes reputational incentives. In maximizing reputational incentives the prevalence of the act and the review standard act as complements: the more common the act the laxer is the review standard that maximizes reputational incentives.

The latter observation plays an important role for the conferral of symbolic rewards, which becomes the optimal policy when the shadow cost of funds or reputational concerns are large. As the population embraces a prosocial act more, it becomes preferable to convey symbolic rewards more frequently by using a laxer review standard. On the other hand, when formal rewards need to be used to bolster incentives, using stricter review standards reduces the cost of conveying rewards. This causes the review standard to be stricter than that which maximizes accuracy, unless the conferral of rewards is the norm. Thus, the optimal review standard depends on multiple factors, including the importance of reputational concerns and the shadow cost of funds, which determines whether reputational incentives ought to be maximized or whether stricter standards ought to be employed.

In closing, we note a few important dynamics highlighted by our analysis, which can be studied further in future research.

First, we noted the crucial relationship between signal generating processes and review standards that maximize reputational incentives. The properties of these processes have not been studied in the literature. Our preliminary observations (see Fig. 4a-c and accompanying text) suggest that there may be an intuitive link between the skewness of signal generating processes and the stringency of standards that maximize reputational incentives.

Second, throughout our analysis we focused on the case where the principal was committed to using carrots, and not sticks. Although there is symmetry in the incentive effects of rewarding good behavior and punishing bad behavior, the cost of using the stick is greater in many circumstances than the cost of using the carrot. Therefore, in a broader context where the choice between rewards versus sanctions is endogenous, the government may choose to use the stick only when targeting minoritarian participation in bad behavior. Thus, formally studying the case wherein the choice between carrots and sticks is endogenous can provide interesting insights regarding the asymmetry of review standards typically used in these two contexts.

Third, in our current analysis we considered third parties who all receive the same public information in order to form opinions about actors. However, people often interact with different kinds of third parties who may receive their information from different sources. It is possible, for instance, for family members and friends to observe the actor's behavior with less noise than the publicly provided signal, while other third parties need rely on the publicly provided signal to form opinions. Analyzing this possibility may highlight additional insights regarding the importance of public signals in shaping behavior as a function of the importance of the actor's behavior for close third parties versus distant third parties.

Our framework can readily be extended to study these and related questions in future work.

Appendix

Proof of Lemma 1. From (2), $\psi(\upsilon^*_{\min},\alpha,\beta)=\beta$ and $\psi(\upsilon^*_{\max},\alpha,\beta)=\alpha$. Therefore, given $\alpha>0$ and $\beta<1$, the denominator in (4) is always positive, so that $\delta(\upsilon^*_{\min},\alpha,\beta)=\delta(\upsilon^*_{\max},\alpha,\beta)=0$. 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 (4), $\partial\delta/\partial\upsilon^*>0$ is equivalent to

$$g(1-2G)\psi(1-\psi)-\frac{\partial\psi}{\partial\upsilon^*}(1-2\psi)(1-G)G>0 \tag{28}$$

Substituting $\partial\psi/\partial\upsilon^*=-g(\beta-\alpha)$, 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 canceling terms then yields

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

Thus, (28) 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

$$1-\beta+\alpha > N[2(\beta-\alpha)G-\beta] \tag{29}$$

If $N\geq 0$, the right hand side is non decreasing in G. A sufficient condition for (29) 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 (29) 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 Bénabou and Tirole (2011) or Adriani and Sonderegger (2019).

Part (ii): That $\mu\Lambda(v^*,\alpha,\beta)<\mu\Delta(v^*)$ follows from the fact that $\delta(v^*,\alpha,\beta)<1$ for all $\alpha>0$ and $\beta<1$, as noted in Lemma 1.

That $\mu\Lambda(v^*,\alpha,\beta)$ never has an interior minimum follows from the facts that $\delta(v^*,\alpha,\beta)=0$ when $v^*\in\{v_{\min},v_{\max}\}$, as noted in Lemma 1, and that $\delta(v^*,\alpha,\beta),\Delta(v^*)>0$ when $v^*\in(v_{\min},v_{\max})$.

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 $q\in(0,1)$,

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

which reveals that

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

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 $t\in(0,1)$,

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

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. 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 (16) and (17), 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 Proposition 2. First, note that $R(v^*,\alpha) > 0$ for any interior v^* and α, while $R(v^*,0) \equiv \lim_{\alpha\to 0} R(v^*,\alpha) = 0$ and $R(v^*,1) \equiv \lim_{\alpha\to 1} R(v^*,\alpha) = 0$. Hence, a maximizer is interior. Secondly, it is easily shown that

$$\frac{R_v(v^*,\alpha)}{R(v^*,\alpha)} = \frac{\tau'(v^*)}{\tau(v^*)} - \frac{Y_v(v^*,\alpha)}{Y(v^*,\alpha)}$$

where $Y(v^*,\alpha) \equiv \psi(v^*,\alpha)(1-\psi(v^*,\alpha))$. Therefore,

$$\frac{\partial(R_v/R)}{\partial\alpha} = -\frac{Y_{v\alpha}Y - Y_vY_\alpha}{Y^2} \tag{30}$$

It can be shown that

$$Y_{v\alpha}Y - Y_vY_\alpha = \{[\beta-\alpha\beta'](1-\psi)^2 + [\beta'(1-\alpha)-(1-\beta)]\psi^2\}g > 0. \tag{31}$$

where the inequality follows from the fact that the terms in both square brackets are positive owing to the concavity of β. Finally, we note that maximizing $R(v^*,\alpha)$ is equivalent to maximizing $\ln R(v^*,\alpha)$. From (30) and (31),

$$\frac{\partial^2 \ln R(v^*,\alpha)}{\partial v^*\partial\alpha} = -\frac{Y_{v\alpha}Y - Y_vY_\alpha}{Y^2} < 0 \text{ for all } v^* \text{ and } \alpha$$

The proposition then follows directly from theorem 1 in Edlin and Shannon (1998). ■

Proof of Proposition 3. We consider only the case where $y > 0$. Substituting from $\frac{\partial W}{\partial\alpha} = 0$ (see (19)) into (14), we get

$$\begin{aligned}\frac{\partial W}{\partial y} &= \lambda y\psi_\alpha\left(\frac{-\partial v^*/\partial y}{-\partial v^*/\partial\alpha}\right) - \lambda\psi \\ &= \lambda y\psi_\alpha\left[\frac{\beta-\alpha}{(\beta'-1)(y+\mu\Lambda)+(\beta-\alpha)\mu\Lambda_\alpha} - \frac{\psi}{y\psi_\alpha}\right]\end{aligned}$$

It follows that $\partial W/\partial y \le 0$ is equivalent to

$$\frac{(\beta'-1)(y+\mu\Lambda)+(\beta-\alpha)\mu\Lambda_\alpha}{\beta-\alpha} - \frac{y\psi_\alpha}{\psi} \ge 0 \tag{32}$$

Straightforward substitutions show that (32) is equivalent to

$$\mu R_\alpha(v^*,\alpha) - \frac{(\beta-\alpha\beta')y}{\psi} \ge 0 \tag{33}$$

The concavity of $\beta(\alpha)$ implies that $\beta > \alpha\beta'$. The first term in (33) must therefore be positive. When $\partial W/\partial y > 0$, the reverse inequality holds strictly in (33), which is consistent with the expected reputational benefit being either increasing or decreasing. ■

Discussion of Assumption 3: We show that the assumption implies $C_\alpha(v^*,0) < 0$. Substituting from (16) in the cost function (22) and defining $K(v^*,\alpha) \equiv C(v^*,\alpha)/\lambda$,

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

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, $C_\alpha(v^*,\alpha) < 0$ 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 (23). ■

Proof of Proposition 4. Write welfare as

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

where $v^* = v^*(y,\alpha)$. Welfare is maximized subject to $y \ge 0$.

We first show the existence of $\overline{\lambda} > 0$ as stated. The policy with $y = 0$ is obviously optimal for λ sufficiently large. So suppose it is optimal for some λ_0 and let α_0 denote the optimal standard, so that $W(0,\alpha_0,\lambda_0) \ge W(y,\alpha,\lambda_0)$ for all y and α. This policy remains optimal for any $\lambda > \lambda_0$ because

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

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 we must have

$$\begin{aligned} W_y(0,\alpha_0,\lambda_0) &= (e + v^*(0,\alpha_0) - c)g(v^*(0,\alpha_0))\left(-v_y^*(0,\alpha_0)\right) \\ &\quad - \lambda_0\psi(v^*(0,\alpha_0),\alpha) \leq 0 \end{aligned}$$

Because the first term in the middle expression is positive, there exists a positive $\overline{\lambda} \leq \lambda_0$ satisfying $W_y(0,\alpha_0,\overline{\lambda}) = 0$ and such that $W_y(0,\alpha_0,\lambda) > 0$ for all $\lambda < \overline{\lambda}$. Therefore, the optimal policy is no-bonus if $\lambda \geq \overline{\lambda}$ and otherwise is an interior solution with a positive bonus.

Next, we show that, in the no-bonus policy, α maximizes the expected reputational benefit. By Assumption 1, the policy minimizes v^* subject to (v^*,α) satisfying:

$$\varphi(v^*,\alpha) \equiv v^* + \mu R(v^*,\alpha) - c = 0$$

Let $\hat{v}^*$ be the smallest feasible threshold and suppose it is achieved with $\hat{\alpha}$. We show that $R(\hat{v}^*,\hat{\alpha}) \geq R(\hat{v}^*,\alpha)$ for all α. Suppose, to the contrary, that there exists α such $R(\hat{v}^*,\hat{\alpha}) < R(\hat{v}^*,\alpha)$. Then,

$$\varphi(\hat{v}^*,\alpha) = \hat{v}^* + \mu R(\hat{v}^*,\alpha) - c > 0$$

Now, by assumption, $\varphi_v(v^*,\alpha) = 1 + \mu R_v(v^*,\alpha) > 0$ for all v^*. Hence, the equilibrium threshold with α is some $v^* < \hat{v}^*$, a contradiction.

Consider now the interior solution which we denote $\hat{y}$ and $\hat{\alpha}$. Then,

$$W_\alpha(\hat{y},\hat{\alpha},\lambda) = W_y(\hat{y},\hat{\alpha},\lambda) = 0$$

Let v^* be the achieved threshold. From Proposition 3, $R_\alpha(v^*,\hat{\alpha}) > 0$. Also, $(\hat{y},\hat{\alpha})$ minimizes the cost of achieving v^*; that is, it minimizes $C(v^*,y,\alpha) \equiv \lambda y\psi(v^*,\alpha)$ where y and α satisfy the equilibrium condition

$$v^* + (\beta(\alpha) - \alpha)y + \mu R(v^*,\alpha) - c = 0$$

We show that $\hat{\alpha} < \alpha^{**}$ where α^{**} is the smallest element in $\arg\max_\alpha R(v^*,\alpha)$. The proof is by contradiction. Let $\overline{R}(v^*,\alpha)$ be defined as follows: for $\alpha < \alpha^{**}$, $\overline{R}(v^*,\alpha) = R(v^*,\alpha)$; for $\alpha \geq \alpha^{**}$, $\overline{R}(v^*,\alpha)$ is the concave closure of $R(v^*,\alpha)$, i.e., the smallest concave function weakly greater than $R(v^*,\alpha)$ over $[\alpha^{**},1]$. Then, $\overline{R}(v^*,\alpha) \geq R(v^*,\alpha)$; moreover, $\overline{R}(v^*,\alpha)$ has a global maximum at α^{**} and $\overline{R}_\alpha(v^*,\alpha) \leq 0$ for all $\alpha \geq \alpha^{**}$. We show that $\hat{\alpha} \geq \alpha^{**}$ yields a contradiction. Let $(\widetilde{y},\widetilde{\alpha})$ minimize $C(v^*,y,\alpha)$ subject to

$$v^* + (\beta(\alpha) - \alpha)y + \mu\overline{R}(v^*,\alpha) - c = 0 \tag{34}$$

Clearly, $C(v^*,\widetilde{y},\widetilde{\alpha}) \leq C(v^*,\hat{y},\hat{\alpha})$. But then, replicating the argument in the text, it must be that $\overline{R}_\alpha(v^*,\widetilde{\alpha}) > 0$, implying that $\widetilde{\alpha} < \alpha^{**}$; therefore, $(\widetilde{y},\widetilde{\alpha})$ also solves the original problem. Moreover, because $(\hat{y},\hat{\alpha})$ with $\hat{\alpha} \geq \alpha^{**}$ does not minimize $C(v^*,y,\alpha)$ subject to (34), it follows that $C(v^*,\widetilde{y},\widetilde{\alpha}) < C(v^*,\hat{y},\hat{\alpha})$, hence the contradiction because $(\hat{y},\hat{\alpha})$ then does not solve the original problem.

The last claim, that either $\alpha < \bar{\alpha}$ or $\psi(v^*,\alpha) > \frac{1}{2}$, follows from Lemma 2 when the optimal policy involves $y = 0$ and from a straightforward reformulation of the same argument when the optimal policy involves $y > 0$.■

Proof of Lemma 3. Let $v^*(y,\theta)$ solve

$$v^*(y,\theta) + (\beta - \alpha)\left(y + \mu\Lambda(v^*(y,\theta) - \theta)\right) = c \tag{35}$$

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

Proof of Proposition 5. From (27), 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 Corollary 1. See the argument in the text.

Proof of Corollary 2. By Assumption 1, the equilibrium condition

$$v^* + \mu R(v^* - \theta,\alpha) - c = 0 \tag{36}$$

has a unique solution, hereafter denoted by $\hat{v}(\theta,\alpha)$. Moreover,

$$\frac{\partial \hat{v}(\theta,\alpha)}{\partial \theta} = \frac{\mu R_v(\hat{v}(\theta,\alpha),\alpha)}{1+\mu R_v(\hat{v}(\theta,\alpha),\alpha)} < 1$$

The inequality holds irrespective of the sign of R_v because $1+\mu R_v > 0$. Therefore,

$$\text{if } \theta'' > \theta', \text{ then } \hat{v}(\theta'',\alpha) - \theta'' < \hat{v}(\theta',\alpha) - \theta' \text{ for all } \alpha \tag{37}$$

Now, the optimal policy minimizes v^* subject to (v^*,α) satisfying (36). By Proposition 4, the solution is $v^*(\theta) \equiv \hat{v}(\theta,\alpha(\theta))$ where

$$\alpha(\theta) \in \arg\max_{\alpha} R(\hat{v}(\theta,\alpha(\theta)) - \theta, \alpha)$$

Hence, $\hat{v}(\theta,\alpha(\theta)) \leq \hat{v}(\theta,\alpha)$ for all α. Combined with (37), if $\theta'' > \theta'$, then

$$\hat{v}(\theta'',\alpha(\theta'')) - \theta'' \leq \hat{v}(\theta'',\alpha(\theta')) - \theta'' < \hat{v}(\theta',\alpha(\theta')) - \theta'$$

Thus, $\theta'' > \theta'$ implies $v^*(\theta'') - \theta'' < v^*(\theta') - \theta'$, which implies $\alpha(\theta'') > \alpha(\theta')$ per Proposition 2. ■

Proof of Corollary 3. Let $C(v^*,\alpha)$ be the cost function as defined in (22), given a positive bonus. In the initial solution, the α_{in} solves

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

With a strict regular minimum with $C_{\alpha\alpha} > 0$,

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

Define

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

so that $C_{\alpha v}(v^*_{in},\alpha_{in}) = -\xi_v(v^*_{in},\alpha_{in})$. Now,

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

where

$$\begin{aligned}S &\equiv \left[\frac{\beta'-1}{\beta-\alpha}+\frac{\psi_\alpha}{1-\psi}\right]\psi\\ &= \left[\frac{\beta-\alpha\beta'+\beta'-1}{\beta-\alpha}\right]\frac{\psi}{1-\psi} > 0\end{aligned} \tag{38}$$

and

$$S_v = \left[\frac{\beta-\alpha\beta'+\beta'-1}{\beta-\alpha}\right]\frac{\psi_v}{(1-\psi)^2} < 0 \tag{39}$$

The expression in the square brackets in (38) and (39) is positive because of the strict concavity of β. The sign in (39) then follows from $\psi_v = -(\beta-\alpha)g < 0$.

Thus, when $\Lambda_v(v^*_{in},\alpha_{in}) \leq 0$,

$$\xi_v(v^*_{in},\alpha_{in}) = \lambda\mu\Lambda_v(v^*_{in},\alpha_{in})S(v^*_{in},\alpha_{in}) + \lambda\mu\Lambda(v^*_{in},\alpha_{in})S_v(v^*_{in},\alpha_{in}) < 0$$

implying that $\partial\hat{\alpha}/\partial\theta > 0$. Rewriting (21), the cost minimizing bonus $\hat{y}(v^*_{in},\theta)$ satisfies

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

Therefore

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

Using the cost-minimization condition (25), this reduces to

$$\frac{\partial\hat{y}}{\partial\theta} = \mu\Lambda_v(v^*_{in},\alpha_{in}) - \frac{y_{in}\psi_\alpha(v^*_{in},\alpha_{in})}{\psi(v^*_{in},\alpha_{in})}\frac{\partial\hat{\alpha}}{\partial\theta} < 0$$

which completes the proof. ■

References

Adriani, F., Sonderegger, S., 2019. A theory of esteem based peer pressure. Games Econom. Behav. 115, 314–335.
Ali, S.N., Bénabou, R., 2020. Image versus information: Changing societal norms and optimal privacy. Am. Econ. J. Microecon. 12, 1–49.
Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. Amer. Econ. Rev. 96, 1652–1678.
Bénabou, R., Tirole, J., 2010. Individual and corporate social responsibility. Economica 77, 1–19.
Bénabou, R., Tirole, J., 2011. Laws and Norms National Bureau of Economic. Research Working Paper 17579.
Bernheim, B.D., 1994. A theory of conformity. J. Polit. Econ. 102, 841–877.
Cooter, R., 1998. Expressive law and economics. J. Law Econ. 27, 585–607.
Cooter, R., Porat, A., 2001. Should courts deduct nonlegal sanctions from damages? J. Legal Stud. 30, 401–422.
Daughety, A., Reinganum, J., 2010. Public goods, social pressure, and the choice between privacy and publicity. Am. Econ. J. Microecon. 2, 191–221.
Deffains, B., Fluet, C., 2020. Social norms and legal design. J. Law, Econ. Organ. 36, 139–169.
Demougin, D., Fluet, C., 1998. Mechanism sufficient statistic in the risk-neutral agency problem. J. Inst. Theor. Econ. 154, 622–639.
Demougin, D., Fluet, C., 2005. Deterrence versus judicial error: A comparative view of standards of proof. J. Inst. Theor. Econ. 161, 193–206.
Demougin, D., Fluet, C., 2006. Preponderance of evidence. Eur. Econ. Rev. 50, 963–976.
Edlin, A.A., Shannon, C., 1998. Strict monotonicity in comparative statics. J. Econom. Theory 81, 201–219.
Ellingsen, T., Johannesson, M., 2008. Anticipated verbal feedback induces altruistic behavior. Evol. Hum. Behav. 29, 100–105.
Fehr, E., Schmidt, K., 2007. Adding a stick to the carrot? The interaction of bonuses and fines. Amer. Econ. Rev. 97, 177–181.
Galbiati, R., Schlag, K., van der Weele, J., 2010. Sanctions that signal: An experiment. J. Econ. Behav. Organ. 94, 34–51.
Glazer, A., Konrad, K., 1996. A signaling explanation for charity. Am. Econ. Rev. 86, 1019–1028.
Gneezy, U., Rustichini, A., 2000. A fine is a price. J. Legal Stud. 29, 1–18.
Herold, F., 2008. Contractual incompleteness as a signal of trust. Games Econom. Behav. 68, 180–191.
Hosmer, D.W., S. Lemeshow, S., 2000. Applied Logistic Regression, second ed. John Wiley and Sons, New York.
Iacobucci, E., 2014. On the interactions between legal and reputational sanctions. J. Legal Stud. 43, 189–207.
Ireland, N.J., 1994. On limiting the market for status signals. J. Public Econ. 53, 91–110.
Jewitt, I., 2004. Notes on the 'shapes' of distributions unpublished.
Kaplow, L., 2011. On the optimal burden of proof. J. Polit. Econ. 119, 1104–1140.
Kim, S.K., 1997. Limited liability and bonus contracts. J. Econ. Manag. Strategy 6, 899–913.
Lehmann, E., Romano, J., 2005. Testing statistical hypotheses. In: Springer Texts in Statistics. Springer, New York, NY.
Mazyaki, A., van der Weele, J., 2019. On esteem-based incentives. Int. Rev. Law Econ. 60, 105848.
McAdams, R., 2000. An Attitudinal Theory of Expressive Law Oregon Law Review. Vol. 79. pp. 339–390.
Mungan, M., 2011. A utilitarian justification for heightened standards of proof in criminal trials. J. Inst. Theor. Econ. 167, 352–370.
Mungan, M., 2016. A generalized model for reputational sanctions and the (Ir)relevance of the interactions between legal and reputational sanctions. Int. Rev. Law Econ. 46, 86–92.
Mungan, M., 2017. Wrongful convictions, deterrence, and stigma dilution. Supreme Court Econ. Rev. 25, 199–216.
Mungan, M., 2020. Justifications, excuses, and affirmative defenses. J. Law, Econ. Organ. 36, 343–377.
Park, E.S., 1995. Incentive contracting under limited liability. J. Econ. Manag. Strategy 4, 477–490.
Posner, E., 2000. Law and Social Norms. Harvard University Press, Cambridge, MA.
Posner, R.A., 2007. Economic Analysis of Law. Aspen Publishers, New York.
Rasmusen, E., 1996. Stigma and self-fulfilling expectations of criminality. J. Law Econ. 39, 519–543.
Rizzolli, M., Saraceno, M., 2013. Better that ten guilty persons escape: Punishment costs explain the standard of evidence. Public Choice 155, 395–411.
Scheinkman, J., 2018. Social interactions (theory) in Macmillan Publishers Ltd (eds). In: The New Palgrave Dictionary of Economics. Palgrave Macmillan, London.
Sunstein, C., 1996. On the expressive function of law. Univ. Pa. Law Rev. 144, 2021–2053.
Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35.