NBER WORKING PAPER SERIES

LAWS AND NORMS

Roland Benabou
Jean Tirole

Working Paper 17579
http://www.nber.org/papers/w17579

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2011

We are grateful to Daron Acemoglu, Tim Besley, Betsy Paluck, Torsten Persson, Aleh Tsyvinski, Glen Weyl, Yao Zeng and participants at many seminars,lectures and conferences for valuable comments, and to Andrei Rachkov and Edoardo Grillo for superb research assistance. Bénabou gratefully acknowledges support from the Canadian Institute for Advanced Research. Both authors gratefully acknowledge support from the European Research Council (Grant No. FP7/2007-2013 - 249429) and from IDEI's "Chaire Association Finance Durable et Investissement Responsable" (AFG). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Laws and Norms
Roland Benabou and Jean Tirole
NBER Working Paper No. 17579
November 2011
JEL No. D64,D82,H41,K1,K42,Z13

ABSTRACT

This paper analyzes how private decisions and public policies are shaped by personal and societal preferences ("values"), material or other explicit incentives ("laws") and social sanctions or rewards ("norms"). It first examines how honor, stigma and social norms arise from individuals' behaviors and inferences, and how they interact with material incentives. It then characterizes optimal incentive-setting in the presence of norms, deriving in particular appropriately modified versions of Pigou and Ramsey taxation.

Incorporating agents' imperfect knowledge of the distribution of preferences opens up to analysis several new questions. The first is social psychologists' practice of "norms-based interventions", namely campaigns and messages that seek to alter people's perceptions of what constitutes "normal" behavior or values among their peers. The model makes clear how such interventions operate but also how their effectiveness is limited by a credibility problem, particularly when the descriptive and prescriptive norms conflict.

The next main question is the expressive role of law. The choices of legislators and other principals naturally reflect their knowledge of societal preferences, and these same "community standards" are also what shapes social judgments and moral sentiments. Setting law thus means both imposing material incentives and sending a message about society's values, and hence about the norms that different behaviors are likely to encounter. The analysis, combining an informed principal with individually signaling agents, makes precise the notion of expressive law, determining in particular when a weakening or a strengthening of incentives is called for. Pushing further this logic, the paper also sheds light on why societies are often resistant to the message of economists, as well as on why they renounce certain policies, such as "cruel and unusual" punishments, irrespective of effectiveness considerations, in order to express their being "civilized".

Roland Benabou
Department of Economics
and Woodrow Wilson School
Princeton University
Princeton, NJ 08544
and NBER
rbenabou@princeton.edu

Jean Tirole
Institut d'Economie Industrielle
Bureau MF529 - Bat. F
21 allees de Brienne
31000 Toulouse
FRANCE
tirole@cict.fr

Introduction

To foster desired behaviors, economists emphasize (with a number of caveats) material incentives provided through contracts, markets or policy. While these often work very effectively, there also many puzzling cases where incentives fail to have the desired effects (e.g., crowding out) or, conversely minor ones have a disproportionately large impact (crowding in, shift in norms).[1] Societies also sometimes "insist" on what seem like inefficiently costly forms of incentives (e.g., prison rather than fines or reparations) or renounce others that might be quite cheap or effective (paying for organ donations, corporal punishments, public shaming).

Rather than incentives, psychologists emphasize persuasion and social influence, in particular through manipulations of collective identity, peer comparison and other interventions aimed at changing the "social meaning" of actions and shifting the norms that prevail in a population.[2] This body of work offers many valuable insights and a wide knowledge base of experimental regularities, but no clear analytical framework.

Legal scholars, finally, certainly agree on the importance of incentives, but many argue that the law is not merely a price system for bad and good behaviors –it also plays an important role in expressing and shaping the values of society. The spectrum of opinions ranges from pure "consequentialists" to pure "expressivists", but here also the underlying architecture –exactly how laws do or should convey societal values– remains elusive. Thus, the expressive content of law is sometimes invoked to call for harsher measures and sometimes for more lenient ones, or appealed to both for and against a given form of punishment.

These apparently disjoint approaches are in fact highly complementary and can be fruitfully brought together to shed new light on the determinants of compliance and the effects of incentives. To this effect, we develop in this paper a unifying framework to analyze how private decisions as well as public policies are shaped by personal and societal preferences ("values"), material or other explicit incentives ("laws") and social sanctions or rewards ("norms").

We first show how honor, stigma and social norms –a social multiplier, more generally– arise from individuals' behaviors and inferences, and when they are strengthened or undermined by the presence of material incentives. We then characterize optimal incentive-setting in the presence of norms, deriving appropriately modified versions of Pigou and Ramsey taxation that correct not just for standard externalities but also for the zero-sum aspect of image-seeking. In particular, this "reputation tax" makes the optimal incentive depend, nonmonotonically, on aggregate shifts in costs or preferences that affect the overall rate of compliance. For well-behaved (unimodal) distributions of individual values, the subsidy is lowest for behaviors with very high or very low participation rates (as these respectively induce maximal stigma and maximal honor), and highest

[1] Examples of such puzzles include e.g., Gneezy and Rustichini [2000a,b], Fehr and Gächter [2001], Knez and Simester [2001], Fehr and Falk [2002], Fehr and Rockenbach [2003], Falk and Kosfeld [2006], Karlan and List [2007], Ariely et al. [2009], Panagopoulos [2009], Funk [2010] and Fryer [2010]. See, e.g., Bowles [2008] and Bowles and Reyes [2009] for recent surveys of the empirical puzzles, and Gibbons [1997] and Prendergast [1999] for the more "classical" literature on incentives in organizations.

[2] See, e.g., Cialdini [1984], Cialdini et al. [2006], Prentice and Miller (1993), or Schultz et al. [2007].

for behaviors in the "grey zone" where compliance and noncompliance are both common behaviors (and social pressure is thus at its weakest).

Next, we incorporate into the model the idea that the distribution of preferences in society may be only imperfectly known by agents. Allowing for such aggregate uncertainty, in addition to the individual heterogeneity standard in signaling models, opens up to analysis a number of ideas and practices found mostly outside economics, but closely linked to the study of incentives.

The first is social psychologists' practice of "norms-based interventions", namely campaigns and messages that seek to alter people's perceptions of what constitutes "normal" behavior (or values) among their peers. We make clear how such interventions operate, but also how their effectiveness is limited by a credibility problem, particularly when the "descriptive" and "injunctive' norms (what most people do, versus what most approve of) are in conflict.

The next and central question we analyze is the expressive role of law. Whether intended to foster the common good or more narrow objectives, laws and other policies reflect the knowledge that decision-makers have about societal preferences. These same "community standards" are also what shapes social norms (conferring esteem or stigma) and moral sentiments (pride and shame). Thus, imposing a heavy sentence for some offense or a zero price on certain transactions means both setting material incentives and sending a message about society's values, and hence about the norms according to which different behaviors are likely to be judged. The analysis, combining an informed principal with individually signaling agents, makes precise the notion of expressive law, determining in particular when a weakening or a strengthening of incentives is called for.

Somewhat surprisingly, the answer turns out to be entirely independent of whether individual behaviors are complements (the usual understanding of a norm) or substitutes (search for distinction). Instead, it hinges on what specific variable the law signals –agents' general willingness to contribute to the public good, or the value to society of such contributions. The underlying intuition, and the main thread running through our analysis, is that the principal can use multiple "currencies" to shape agents' behavior. In the empirically relevant case where rewards and punishments are costly to implement, he will seek to economize on them by harnessing agents' other sources of motivation –intrinsic and reputational. Thus, when better informed about prevailing standards of behavior, he tries to signal that social sanctions or payoffs are large by lowering extrinsic incentives, at some cost in compliance. In contrast, when the asymmetric information concerns the magnitude of the externalities that agents impose on each other (and provided that they care more, the larger their social impact), the principal seeks to enhance their intrinsic motivation by convincing them that the externalities are large, and this now involves setting higher incentives than under symmetric information.

Pushing further this logic, we extend the model to also shed light on why societies are often resistant to economists' messages about the virtues of incentives, as well as on why they forego certain policies, such as "cruel and unusual" punishments, irrespective of effectiveness considerations, in order to express their being "civilized".

• **Related literature.** The need for an integrated analysis of law and social norms is stressed, among others, by Ellickson [1998], Lessig [1998] and McAdams and Rasmusen [2007], who provide an excellent survey. The expressive role of law is emphasized in particular by Sunstein [1996], Kahan [1997], Cooter [1998], Posner [1998, 2000a,b] and McAdams [1999]. Our signaling formalization is most closely related to the approach advocated by the last two authors.[3] Recent experimental evidence on the expressive effect of incentives is provided by Tyran and Feld [2006], Galbiati and Vertova [2008], Galbiati et al. [2010] and Bremzeny et al. [2011].

The interaction of incentives with other forms of motivation under symmetric information is studied by, among others, Frey [1997], Brekke and Nyborg [2003], Besley and Ghatak [2005] and Bénabou and Tirole [2006a], and we start by extending the model of prosocial behavior developed in this last paper to new settings (distributions of preferences, distortive taxation) and results. Kaplow and Shavell [2007] consider a social planner who, instead of incentives, has access to a costly "inculcation" technology for feelings of guilt and virtue (acting respectively as a tax and a subsidy) and characterize the optimal mix of these two instruments. Fischer and Huddart [2008] study the impact of incentives when agents engage in both desirable and undesirable behaviors (e.g., performance falsification) which the principal cannot tell apart, but which are subject to separate social norms among agents, giving rise to different social multipliers.

The informed-principal problem that arises with expressive law bears some relationship to those in Bénabou and Tirole [2003], Ellingsen and Johannesson [2008] and Herold [2010], but with the important difference that what agents now try to infer − the prevailing social standard− embodies everyone's equilibrium actions and beliefs. The idea that incentives convey information about the distribution of preferences is shared with Sliwka [2008] and van der Weele [2009], but the nature of normative influences is quite different. In the first paper, social complementarities operate through "conformist" types, whose preference is to mimic whatever action the majority chooses. In the second they involve "reciprocal altruists", whose taste for contributing to a public good rises with total contributions. Our model has no built-in complementarity; conformity or distinction effects arise endogenously, and we analyze expressive law in both cases.

The paper is organized as follows. Section 1 lays out the model and basic results concerning honor, stigma, and the social multiplier. Section 2 characterizes optimal incentives under common knowledge about societal values. Section 3 takes up norms-based interventions and Section 4 analyzes the expressive role of law. Section 5 discusses robustness, while Section 6 concludes with directions for further research.

[3] An alternative route for laws to affect social norms is an evolutionary process of preference adaptation; e.g., Huck [1998], Bohnet et al. [2001], Bar-Gill and Fershtman [2004], Tabellini [2008] and Guiso et al. [2008].

1 Model

1.1 Basic framework

• **Agents.** A continuum of agents each choose some discrete action $a \in \{0, 1\}$, with resource cost (time, effort, etc.) ca, $c > 0$. Each also receives from some principal an incentive of ya. In a firm or organization, a represents working rather than shirking, abstaining from opportunism, etc., and y a wage rate or performance-contingent bonus. In a public-goods context a is some prosocial action such as not polluting, voting, contributing, etc., with y representing a subsidy on the provision of the public good or, conversely, a penalty (tax, fine, prison) on undesirable behaviors (i.e., on $-a$).

To represent agent's preferences we use the simplest specification that encompasses the three key ingredients of intrinsic motivation, extrinsic incentives and (social or self) esteem concerns:

$$U = (v + y - c)a + e\bar{a} + \mu\ E(v \mid a). \tag{1}$$

We refer to v as the agent's *intrinsic motivation.* In a firm or organization it corresponds to liking and motivation for the task (sales, research), work ethic, etc. In a public-goods context, it represents the agent's degree of altruism or prosocial orientation, whether general or activity-specific (e.g., concern for the environment); each agent then also derives a benefit e (for "externality") from the aggregate supply, $\bar{a}$.[4, 5]

Decisions also carry reputational costs and benefits, reflecting the judgements and reactions of others as they assess an agent's intrinsic motivation, which is private information, in light of his actions. These concerns, represented by the last term in (1), can be purely hedonic (valuing social esteem per se) or instrumental.[6] Thus, in a labor-market, career concerns make it valuable to be seen by employers as having a strong work ethic, caring about the activity in question, etc. In the social sphere, people perceived as generous, public minded, good citizens, etc., are more likely to be chosen as mates, friends, or leaders. Reputational payoffs can also be reinterpreted as the value of *self*-image or moral sentiments, with each individual judging his "true character" by his own conduct: self-signaling works much like social signaling, with memorability or salience substituting

[4] It does not matter at this stage whether v reflects, in Andreoni's [1989] terminology: (i) pure altruism, meaning that the agent values his actual contribution to others' welfare via his impact on $\bar{a}$, and requiring group size to be small (in this case the v's are linked to e; see Section 4.3); (ii) impure altruism, arising from a "joy of giving" (social and self esteem are accounted for separately in (1)) or reflecting "Kantian"-type reasonings in which the agent overscales his real impact on $\bar{a}$ (e.g., Brekke et al. [2003]). On intrinsic motivation in firms or organizations, see also Besley and Ghatak [2005].

[5] One can easily allow for a differential impact of $\bar{a}$ across agents, but focussing on its average value is sufficient for most purposes. Also for simplicity, we abstract from decreasing or increasing returns in the value of $\bar{a}$.

[6] The evidence supporting the role of reputational concerns is extensive; see footnote 24 for some examples, based on manipulations of the act's visibility. On the modeling side, value functions derived from an explicit second-stage game may not be linear (e.g, Rotemberg [2008]), or may involve weights that vary with the agent's type. The reduced-form specification (1) greatly simplifies the analysis, while capturing key effects also present in more complicated cases. One can also relax it substantially, as discussed in Section 5.

for external visibility.[7]

To analyze most transparently the interplay of individual and aggregate uncertainty over agents' preferences we focus on a single source of heterogeneity, namely intrinsic motivation. Thus all share the same marginal valuation for money or other extrinsic incentives y, which is normalized to 1; they also care equally about social (or self) esteem, as reflected by a common μ.[8]

• **Principal.** Agents face a principal who sets the incentive y –wage, tax or subsidy, sentence, etc. This is "the law", whether that of the company or that of the land.[9] The principal may also be able to communicate with agents directly, whether by disclosing hard information or through cheap talk.[10] We focus the exposition on the case of a planner who maximizes social welfare

$$W = \bar{U} - (1+\lambda) y \bar{a}, \tag{2}$$

where $\bar{U}$ is the sum, in equilibrium, of all agents utilities' defined in (1), $\bar{a}$ their total contribution and $\lambda \geq 0$ the shadow cost of funds or other resources used as incentives –deadweight loss from taxes, enforcement costs, etc. This focus is without loss of generality, as a simple renormalization maps the objective function of less benevolent or even purely selfish principals into (2).[11]

1.2 The calculus of esteem and the social multiplier

Let $G(v)$ denote the distribution of agents' intrinsic motivations, with finite support $V \equiv [v_{\min}, v_{\max}]$, continuously differentiable density $g(v) > 0$, hazard rate $h(v) \equiv g(v)/\left(1 - G(v)\right)$ and

[7] See Smith [1759], Bem [1972], Bodner and Prelec [2003], Bénabou and Tirole [2004, 2011]

[8] The specification (1) is thus a special case of the more general model (Bénabou and Tirole [a])

$$U = (v_a + v_y y)\, a - C(a) + e\bar{a} + \mu_a\, E(v_a|a) - \mu_y\, E(v_y|a),$$

in which: (i) the action a can be discrete or continuous; (ii) agents can also have different marginal utilities of money v_y; (iii) it may be undesirable to be perceived as greedy or needy, as reflected in the last term; (iv) agents' reputational concerns can also differ, so that a type is a quadruplet (v_a, v_y, μ_a, μ_y). At least two-dimensional heterogeneity is required to generate a negative response to incentives (net crowding out), as well a related "overjustification effect" of publicity. We shall be abstracting here from these phenomena, to better focus on new ones, arising in particular from aggregate uncertainty.

[9] We assume costless observation of behaviors by the principal and a well-understood notion of what constitutes a good or bad action. Shavell (2002) argues that transaction costs and better local knowledge of situational factors can make social norms preferable to legal enforcement. See also Fisher and Hudddart (2008) for a model with norms and an informationally constrained principal.

[10] Another policy tool can be for the principal to affect the public visibility or memorability of agents' actions, thus scaling the reputational weight μ (more generally, all the (μ_a, μ_y)'s in footnote 8), by a factor $x \geq 0$, at some cost $\phi(x)$. Similarly, laws and public campaigns can also influence behavior by causing people to pay more attention to their actions and the image they will project. On the benefits and costs of visibility-based policies, see Prat [2005], Bénabou and Tirole [2006a], Daughety and Reinganum [2009], Bar-Isaac [2009] and Ali and Bénabou [2010].

[11] Let $W = \alpha\bar{U} + [B - (1+\lambda)y]\,\bar{a}$, where $0 \leq \alpha \leq 1$ and B represents any private benefits derived from agents' participation or effort. For the planner, $\alpha = 1$ and $B = 0$, while a firm selfishly maximizing profits has $\alpha = 0 < B$. Imperfectly benevolent principals –government agency, NGO, church, etc.– fall in-between. Clearly, by redefining agents' values as $v' \equiv \alpha v, C' \equiv \alpha C$, $e' = \alpha e + B$ and $\lambda' \equiv 1 - \alpha + \lambda \geq 0$, this more general problem reduces to (2).

mean $\bar{v}$. We also define two important conditional moments and their difference:

$$\mathcal{M}^{+}(v) \equiv E\left[\tilde{v} \mid \tilde{v} > v\right], \quad \mathcal{M}^{-}(v) = E\left[\tilde{v} \mid \tilde{v} < v\right], \tag{3}$$

$$\Delta(v) \equiv \mathcal{M}^{+}(v) - \mathcal{M}^{-}(v), \quad \text{for all } v \in V. \tag{4}$$

Given y, an agent chooses $a = 1$ if $v \geq c - y - \mu\left(E\left[\tilde{v} \mid a = 1\right] - E\left[\tilde{v} \mid a = 0\right]\right) \equiv v^{*}$, defining a cutoff rule. Conversely, in an interior equilibrium (on which we shall focus for simplicity),the two conditional expectations are given by $\mathcal{M}^{+}(v^{*})$, which governs the "honor" conferred by participation, and $\mathcal{M}^{-}(v^{*})$, which governs the "stigma" from abstention. In the self-image interpretation of the model, $\mathcal{M}^{+}(v^{*})$ and $\mathcal{M}^{-}(v^{*})$ correspond to virtue and guilt.

The net reputational incentive is $\mu\Delta(v^{*})$, and the cutoff v^{*} solves the fixed-point equation[12]

$$v^{*} - c + y + \mu\Delta(v^{*}) = 0. \tag{5}$$

• **The social multiplier.** When more people "do the right thing", or are thought to do so, does the pressure on individuals to also choose $a = 1$ rise or fall? As v^{*} decreases (see Figure 1a), honor declines but stigma worsens, since both $\mathcal{M}^{+}$and $\mathcal{M}^{-}$ are increasing functions. Depending on which effect dominates, the net social or moral pressure $\mu\Delta(v^{*})$ can decrease or increase. In the first case, $\Delta'(v^{*}) > 0$ and decisions are (locally) strategic substitutes. In the latter case, $\Delta'(v^{*}) < 0$, they are (locally) strategic complements, which corresponds to the usual definition of a norm.

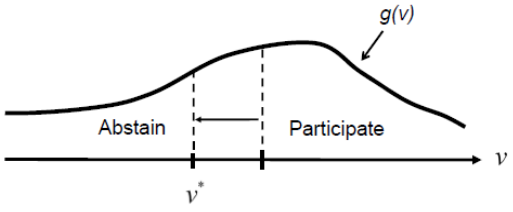

Figure 1a: preference distribution

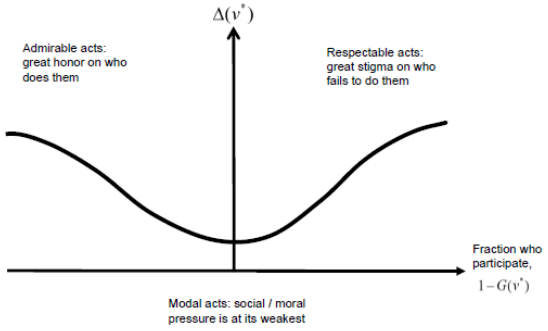

Figure 1b: reputational returns

If complementarity is strong enough and μ high enough, there can be multiple equilibria –that is, self-sustaining norms. In this paper, however, we ensure uniqueness by imposing $1+\mu\Delta'(v) > 0$, which holds for μ not too large.[13] The slope of aggregate supply $a(y) = 1 - G(v^{*}(y))$ is then $g(v^{*})$

[12] An interior equilibrium will be ensured by assuming (or, later on, ensuring that the optimal y satisfies) $(1-\mu)v_{\min} + \mu\bar{v} < c - y < (1+\mu)v_{\max} - \mu\bar{v}$, together with the condition stated below for monotonicity of $v + \mu\Delta(v)$.

[13] The fact that $|\Delta'|$ is bounded is shown in the Appendix. Bénabou and Tirole [2006a] provide sufficient conditions and explicit examples for the case of multiplicity, $1 + \mu\Delta' < 0$. Previous signaling models with a continuum of types and potentially multiple equilibria include Bernheim [1994] and Rasmusen [1996]. For a model with complementarities between non-reputational norms and incentives, see Weibull and Villa [2005].

times the *social multiplier,*

$$-\frac{\partial v^*}{\partial y} = \frac{1}{1+\mu\Delta'(v^*)}. \tag{6}$$

Intuition suggests that honor concerns will dominate when people who "do the right thing" ($a = 1$) are fairly rare, and stigma considerations prevail when only a few "deviants" fail to comply. This is only true, however, under appropriate restrictions on the distribution of agents' preferences.

Lemma 1 (Jewitt 2004) *If g is everywhere decreasing (increasing), then Δ is everywhere increasing (decreasing). If g has a unique interior maximum, then Δ has a unique interior minimum.*

We focus on the second case, which is more general as it allows for both strategic substitutability and complementarity; see Figure 1b. For technical reasons, we impose on Δ a slightly stronger version of the quasiconcavity in (2), by assuming that it is *strictly* decreasing everywhere to the left of its minimum, and *strictly* increasing everywhere to the right. Note that this minimum does *not* coincide (generically) with the mode of g.

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.[14]

It is worth noting that the model generates endogenously the two types of signaling motives which, in the previous literature, were taken as alternative assumptions: a desire to signal conformity (e.g., Bernheim [1994]), and a desire to signal distinction (e.g., Pesendorfer [1995]).[15]

Crowding in and out. Because they (partially) crowd out social esteem, material incentives, laws, fines and prizes are not very effective means to spur admirable, honor-driven behaviors such as military valor, or risking one's life to save someone else's : the multiplier is less than 1.[16] Incentives are much more effective (multiplier exceeding 1) for respectable behaviors, such as tax compliance, as they are amplified by the dynamics of stigma (partial crowding in). Where net costs

[14] Other factors affecting the relative strength of honor and stigma include nonlinearities in reputational payoffs, $E[\varphi(v) \mid a]$ (e.g., Corneo [1997]) and differential visibility of good and bad deeds (Bénabou and Tirole [2006a]).

[15] See also Brennan and Brooks [2007], who do not formulate a signaling model but postulate, based on intuition, that the interplay of esteem and disesteem should lead to a net reputational value that is U-shaped with respect to the rate of compliance. We prove such a result, which holds provided the distribution of types is well behaved (unimodal).

[16] Full crowding out (a negative supply response to incentives, or multiplier) requires multidimensional heterogeneity, as described in footnote 8. This phenomenon was investigated elsewhere and is therefore not our focus here.

are not too high (a moderately low v^*) and actions easily observable (a high μ), small variations in incentives can induce large changes in aggregate behavior.[17]

• **Shifts in societal preferences.** A key focus of the paper are situations in which there may be aggregate shifts in the distribution of agents' preferences. For any $\theta \in \mathbb{R}$, let

$$G_\theta(v) \equiv G(v - \theta) \tag{7}$$

be the original distribution shifted to the right by θ, with density $g_\theta(v) = g(v - \theta)$ and hazard rate h_θ defined similarly. We assume from here on that individuals' private valuations are distributed according to $G_\theta(v)$ on the support $V_\theta \equiv [v_{\min} + \theta, v_{\max} + \theta]$, where "community standards" θ may be known or a priori uncertain. Conditionally on θ, the reputational return to choosing $a = 1$ is easily seen to be $\Delta_\theta(v) \equiv \Delta(v - \theta)$. Without loss of generality we normalize the v's (adding a constant) so that the minimum of Δ occurs at $v = 0$, and that of Δ_θ therefore at $v = \theta$. To insure that the equilibrium cutoff is always interior (and, a bit more strictly, participation bounded away from 0 and 1), we then restrict the model's parameters to satisfy

$$v_{\min} + \theta + \varepsilon < c - e < v_{\max} + \theta - \varepsilon \tag{8}$$

for some fixed but arbitrarily small $\varepsilon > 0$. Equivalently the support of θ lies in $\Theta \equiv [c - e - v_{\max} + \varepsilon,\ c - e - v_{\min} - \varepsilon]$.[18]

For known θ the whole model remains unchanged, with all variables simply indexed by θ : the cutoff, $v^*_\theta(y)$, is still given by (5) and the multiplier by (6), with Δ_θ in place of Δ, implying

$$v^*_\theta(y) - \theta = v^*(y + \theta), \quad \forall\ y,\ \theta. \tag{9}$$

A known shift in societal preferences θ therefore has *the same effect* on (equilibrium) social norms $\Delta_\theta(v^*_\theta(y))$ and aggregate behavior $a_\theta(y)$ as an increase in material incentives y of the same magnitude. This equivalence already suggests that, for a principal, communicating about community standards, or a firm's "culture", can be an attractive substitute to costly rewards and punishments –provided he can achieve credibility.

[17] A good example is Ireland's 33¢ tax on plastic bags (instituted in conjunction with an awareness campaign): "Within weeks, plastic bag use dropped 94%. Within a year, nearly everyone had bought reusable cloth bags, keeping them in offices and in the backs of cars. Plastic bags were not outlawed, but carrying them became socially unacceptable –on a par with wearing a fur coat or not cleaning up after one's dog." (Rosenthal [2008]). Other examples include Continental Airlines' $50 bonus program based on company-wide performance (Knez and Simester [2001]) and the impact on voting turnout of (removing) "symbolic" fines for non-voters in Switzerland (Funk [2007]).

[18] Condition (8) means that it is socially inefficient (respectively, efficient) for the least (most) motivated agent, with types close to $v_{\min} + \theta$ (close to $v_{\max} + \theta$) to contribute. It will imply that for y close to the first-best optimum (which delivers $v^* = c - e$), the cutoff remains interior (i.e., the condition given in footnote 12 is satisfied).

2 Optimal incentives with norms: symmetric information

Consider the problem of a social planner who sets the incentive y to maximize total welfare, $W = \bar{U} - (1+\lambda)y\bar{a}$. Each individual who contributes, at a cost of c, values doing so at v and additionally generates an external benefit e for society. He also receives y, but this costs $(1+\lambda)y$ to provide, where $\lambda \geq 0$ is the shadow cost of public funds. Finally, the contributor reaps the reputational benefit $\Delta_\theta(v^*)$ but, image being a zero-sum game (or "positional good"), inflicts an equivalent reputational loss on non-contributors.[19] Thus, social welfare is equal to

$$W_\theta^{FI}(y) \equiv \int_{v_\theta^*(y)}^{+\infty} (e + v - c - \lambda y)\, g_\theta(v) dv + \mu\bar{v}. \tag{10}$$

In all that follows we assume this objective function to be strictly quasiconcave in y, for all θ; such is the case provided λ is small enough. The optimal incentive is then given as the solution to

$$[e + v_\theta^*(y) - c - \lambda y] \left(\frac{-\partial v_\theta^*(y)}{\partial y} \right) g_\theta(v_\theta^*(y)) = \lambda \left[1 - G_\theta(v_\theta^*(y)) \right]. \tag{11}$$

The interpretation is familiar from Ramsey taxation: the net social marginal benefit of raising y by one dollar (inducing $da_\theta = (-\partial v_\theta^*/\partial y)\, g_\theta$ new agents to participate) is equated to the deadweight loss from paying the extra subsidy to all inframarginal agents.[20] The condition can also be rewritten so as to make clear role of the *social multiplier* in the participation response,

$$\frac{e + v_\theta^*(y) - c - \lambda y}{1 + \mu\Delta_\theta'(v_\theta^*(y))} = \frac{\lambda}{h_\theta(v_\theta^*(y))}. \tag{12}$$

The *first-best* case of no distortion will prove to be an important benchmark under both symmetric and asymmetric information. With $\lambda = 0$, (12) simplifies to $e + v_\theta^*(y^{FB}(\theta)) = c$, which is the standard Samuelson condition equating the total social benefit and cost of a marginal contribution.

Proposition 1 (modified Pigou) *The first-best (symmetric information, $\lambda = 0$) incentive is*

$$y^{FB}(\theta) = e - \mu\Delta_\theta(c - e) = e - \mu\Delta(c - e - \theta). \tag{13}$$

When g is strictly unimodal in v, y^{FB} is single-peaked with respect to θ and c, and maximized at $\theta_0 \equiv c - e$.

[19] Computing agents' average utility $\bar{U}$, esteem and stigma sum up to $(1 - G(v^*))\mathcal{M}^+(v^*) + G(v^*)\mathcal{M}^-(v^*) = \bar{v}$. If reputational payoffs are nonlinear, or if μ varies with v, signaling can be a positive or negative-sum game, depending in particular on the curvature of probability functionals. The linear case serves as a natural and important benchmark, and also avoids "philosophical" debates on whether or not esteem and stigma, or pride and shame, should be counted as part of social welfare.

[20] Defining the elasticity $\varepsilon_\theta(y) \equiv y a_\theta'(y)/a_\theta(y)$, (12) can be rewritten in the Lerner-Ramsey form $(\hat{e} - y)/y = (\lambda/1+\lambda)\,(1/\varepsilon_\theta(y))$, where $\hat{e} \equiv (e - \mu\Delta)/(1+\lambda)$ is the net externality, scaled in monetary units.

One must subtract from the standard Pigovian subsidy, e, the reputational rent extracted by a marginal contributor from the rest of society. Otherwise choosing $a = 1$ would be overcompensated, and conversely noncompliers would suffer an excessive "double whammy". This *reputation tax* makes y^{FB} dependent, in a nonmonotonic way, on θ and c; see the top curve in Figure 2. When θ is low or c high, most people do not contribute, so the few who do reap significant honor. Conversely, when θ is high relative to c, the few "bad apples" who fail to participate incur strong stigma. At both ends there is thus a strong reputational incentive, making a low y optimal. When θ is close to θ_0, on the other hand, social pressure is at its weakest –contributing and abstaining are both common– requiring higher incentives.[21]

• *Implications.*

(i) The tax deduction rate for charitable donations should be lower than the standard Pigovian level and, most importantly, vary inversely with the publicity or image value inherent to the gift. While implementing such a scheme in practice may not be easy, there are, for instance, well established "market prices" for naming rights to a university or hospital building, an endowed chair, etc. Similarly, agencies rating corporations on their social responsibility could aim to incorporate a "publicity discount" in their scores. Indeed, the purpose of such evaluations is to measure true contributions to social welfare and even, through the response of some market participants, reward or punish corporations accordingly.[22]

(ii) Similar distortions toward the more visible (high μ) occur on the consumer side: the premium paid for "fair trade" or "green" products also buys social and self image, the flip side of which is the stigma (or bad conscience) shifted to others. As a result, too many dollars likely flow toward hybrid cars and solar panels relative to housing insulation and efficient furnaces (Ariely [2008]), or toward free-trade coffee compared to food kitchens.

(iii) Consider a new environment-friendly technology that diffuses more widely as its cost c falls, due to technological progress. The optimal subsidy rate should first rise, then fall over time.

In general, y^{FB} could be positive or negative (taxing image-seeking behaviors with low or negative social value). We shall assume from now on that

$$e > \max\left\{\mu\Delta(v_{\min}); \mu\Delta(v_{\max})\right\} = \mu \max\left\{\bar{v} - v_{\min}, v_{\max} - \bar{v}\right\}, \tag{14}$$

[21] This result has interesting parallels with Kaplow and Shavell [2007], who relate the optimal use of guilt and virtue to the frequency of good or bad behavior. In their model, society has a costly "inculcation" technology for feelings of guilt and virtue, which can be manipulated separately. In our model, guilt and virtue ($\mathcal{M}^-$ and $\mathcal{M}^+$) arise in equilibrium from everyone's actions and inferences. This makes them interdependent, and vary with ("control for") the level of material incentives. The principal also has more limited ways of affecting these feelings (or social payoffs): publicity x (which amplifies them), or incentives y and messages about some society-wide variable (θ or e), both of which alter equilibrium inferences.

[22] One can think of agencies as reporting, and "ethical" consumers or investors's willingness to pay depending on, the net social impact $a\left[e - \mu\Delta(v^*(y)) - y(1+\lambda)\right]$ or a firm's actions, Alternatively, the information communicated may be what is leaned about the firm's (or its management's) intrinsic "goodness" v, namely that it is above or below a cutoff $v^*(y) = c - y - \mu\Delta(v^*(y))$. In that case proper adjustment must also be made (when computing $v^*(y)$) for the reputational benefits reaped by the firm (which are no more observable to individuals than a or e).

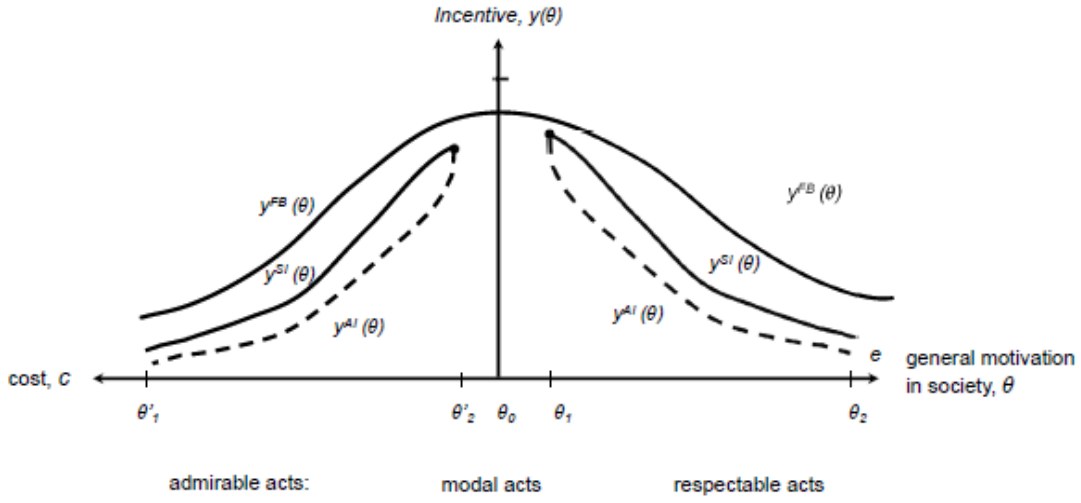

Figure 2: law and societal values

which ensures that $y^{FB}(\theta) > 0$ for all θ, since the function Δ is quasiconvex.

Turning back now to the more realistic *second-best* case where $\lambda > 0$, but still with full information about θ, (12) implies

$$e + v_\theta^*(y^{SI}(\theta)) > e + v_\theta^*(y^{SI}(\theta)) - \lambda y^{SI}(\theta) > c. \tag{15}$$

The social benefit from the marginal contribution exceeds its social cost, implying that $y^{SI}(\theta) < y^{FB}(\theta)$; see Figure 3, middle curve. Showing these results rigorously, however, requires solving a system of implicit equations ((5) and (12)) in $y^{SI}(\theta)$ and v_θ^* that can have singularity points. By excluding some (arbitrarily small) interval around $\theta_0 = c - e$ and restricting attention to values of λ that are not too large, however, we can show:

Proposition 2 (modified Ramsey) *Let (θ_1, θ_2) be any subinterval of Θ not containing θ_0. For all λ below some $\bar{\lambda} > 0$:*
(i) The symmetric-information policy $y^{SI}(\theta)$ is uniquely defined on (θ_1, θ_2) by (12) and it satisfies $0 < y^{SI}(\theta) < y^{FB}(\theta)$. Compliance is thus lower than in the first-best case.
(ii) The function $y^{SI}(\theta)$ is strictly increasing in θ when $\theta_2 < \theta_0$, and strictly decreasing when $\theta_1 > \theta_0$.

These results demonstrate the robustness of the insights from the first-best case –reputation tax and bell-shape of the optimal policy. Where the shadow cost λ (economic, political) of providing material incentives is high, however, compliance will fall substantially short of the first best.[23] In such cases, other instruments may be more effective.

[23] We show in Proposition 9. that $y^{FI}(\theta)$ is indeed decreasing in λ, under an additional technical assumption.

3 Persuasion and norms-based interventions

• **Empirical evidence.** While straight public appeals to good citizenship or generosity sometimes work fairly well (e.g., Reiss and White [2008] on electricity conservation during the California energy crisis of 2000-2001), substantial evidence shows that they are much more effective when leveraged by *social comparisons.* The first form this can take is making individual behavior more observable to everyone (increasing μ).[24] The second, on which we shall focus here, consists in altering people's perceptions of what constitutes "normal" behavior or values in their reference group (e.g., Cialdini [1984], Cialdini et al. [2006]). Two types of norms are distinguished in the social-psychology literature and may be targeted in these interventions –sometimes also called "norms marketing campaigns". The first is the *descriptive* norm (the norm of "is"), namely what most other people in your community actually do. The second is the *prescriptive* or *injunctive* norm (the norm of "ought"): what most other people in your community approve of.

Schultz et al. [2007] thus monitored the electricity meters of 290 households, hanging each week on their doors a visible feedback form with either: (i) the household's own electricity consumption and the average for comparable ones in their neighborhood; (ii) the same information, plus a smiley face if they were below average, a frowning one if they were above. The "descriptive" condition (i) led to a convergence toward the mean: high-consuming households adjusted down but low-consuming ones adjusted up. In the "descriptive + prescriptive" condition (ii), the latter effect disappeared, leading to a reduction in total consumption. Several electricity districts have now adopted similar programs on a much larger scale. Ayres et al. [2010] study those of Sacramento and Puget Sound (each with samples sizes of about $85,000$); implementing on consumers' monthly statements a scheme broadly similar to (ii) led to a rapid and lasting reduction in average consumption, of about 2.5%.

Psychologists' view of how such interventions work is that: (i) people care strongly about social comparisons and self approval, and judge what "one should do" by what they believe others do, or approve of (that is, $\Delta' < 0$, generating complementarity); (ii) they often *misperceive the norm.* First, there is some dispersion in individual perceptions, which we shall relate to the convergence effect. Most importantly, there can be an aggregate misperception, termed *pluralistic ignorance* (e.g., Miller and Prentice [1994a, 1994b]). People may have incorrect beliefs about how others generally behave or assess their peers, or they may not properly account for the extent to which others are also conforming to a commonly (mis)perceived norm.[25] Dispelling excessive pessimism can then (for $\Delta' < 0$) bring about a large and beneficial shift in collective behavior, as found for

[24] For evidence, see e.g., Croson and [1998], Gerber et al. [2008], Ariely et al. [2009], Funk [2010], Della Vigna et al. [2010], Kessler [2011] or Stocking [2011]. For an analysis of the benefits and costs of such policies, see Bénabou and Tirole [2006a], Daughety and Reinganum [2009] and Ali and Bénabou [2010].

[25] This is often seen as an instance of the "fundamental attribution error" in which people systematically underestimate the power of the situation. This idea is also related to those of "social proof" (Cialdini [1984]), and "preference falsification" (Kuran [1995]). Psychologists' explanation is thus implicitly or explicitly one of limited rationality, but it does not have to be, as shown at the end of this section.

instance by Prentice and Schroeder [1998] in the context of student drinking.[26]

Where pluralistic ignorance takes the form of excessive *optimism* about others' conduct or values, however, the truth will further damage compliance. For instance, in the case of pornography –which in US law is explicitly judged according to "community standards"– there is definite evidence that actual use vastly exceeds what people think.[27] The same is likely true of drug use, and possibly of racist and sexist attitudes. In an experiment on tax evasion and welfare fraud in three European countries, Lefebvre et al. [2011] find that revealing previous instances of low average compliance increases evasion.[28] The literature's standard recommendation is to use the descriptive norm (or both) if most people behave well relative to general expectations, but if they behave badly use only the prescriptive norm. There is, however a problem of long-run *credibility* (or legitimacy; Tyler [1990], Xiao [2010]), as the principal –experimenter, NGO, policymaker– finds himself in the position of selectively disclosing and framing good behaviors, eluding or minimizing depressing truths, and relying on "soft" statements about what people declare they approve of, while doing the opposite.[29]

• **Formalizing norms-based interventions.** Descriptive interventions correspond to communicating with agents about the average $\bar{a}$, which in turn reflects some preference parameter like θ that they are imperfectly informed about. Prescriptive interventions, from public campaigns to individualized "smiley faces", can be understood as communicating about e ("people are strongly affected by this problem") or especially about μ ("people make strong judgments based on this behavior"), which boosts social pressure $\mu\Delta$ both directly and, for respectable acts, indirectly: $\partial(\mu\Delta)/\partial\mu = \Delta/(1 + \mu\Delta')$. As we show below, however, even a fully benevolent principal will always try to exaggerate or selectively disclose positive information about $\bar{a}, \theta, e$ or μ. Agents, conversely, will discount such cheap talk, a fortiori when the "norms entrepreneur" could be deriving private benefits from their compliance.

Let agents be only imperfectly informed about current "community standards", namely the overall behavior of the population against which theirs will be judged. Indeed, these shift with the

[26] Prentice and Miller [1993] showed that students overestimate the extent to which others approve of drinking, and that this perceived tolerance by peers is a strong predictor of use. Prentice and Schroeder [1998] randomly "treated" entering students with information dispelling the stereotype, and this led to significantly lower reported levels of consumption. Moreover, this effect was mediated by students' "fear of negative evaluation", as assessed by an initial psychological questionnaire. For similar findings (over-pessimistic pluralistic ignorance, effective intervention from dispelling it) in the context of tax compliance, see Wenzel [2005].

[27] Richtel [2008] relates how the defense lawyer in a recent pornography trial sought to subpoena Google for data on the frequency of sexual-term searches among internet users in the very town where the trial was held. Conversely, the prosecutor fought hard (and prevailed) to block any such information from being provided to the jury.

[28] Revealing instances of high compliance, on the other hand, has no significant effect in their sample. This can be explained by our model, given that tax compliance is arguably a respectable rather than a heroic behavior; figure 1b and equation (6) then show that the social multiplier is large when the proportion of compliers is near the middle of the respectable range, and decreases toward zero when it approaches 1.

[29] For experiments based on framing, see, e.g. Cialdini et al. [2003] or Goldstein et al. [2006]. Note also that the truth about dominant behavior is likely to leak out over time (as occurred in Prentice and Schroeder [1998]), and when the descriptive and injunctive norms visibly diverge, the former tends to trumps the latter (e.g., Tyran and Feld [2006], Bicchieri and Xiao [2010]).

underlying distribution of preferences in society, θ, which is hard for an individual to observe.[30] Agents' prior belief about θ is that it lies in some interval $[\theta_1, \theta_2] \subset \Theta$, with distribution $F(\theta)$. The legislator or principal may have information about θ (or, equivalently, $\bar{a}$), for instance from having observed the behavior of a representative sample.[31]

The principal cannot or does not vary incentives (laws with expressive content will be studied later on), so y is fixed, –possibly at zero, or more generally at a low enough level that

$$e > \mu \max\{\bar{v} - v_{\min}, v_{\max} - \bar{v}\} + (1+\lambda)y. \tag{16}$$

This condition, reducing to (14) when $y = 0$, ensures that greater participation always raises social welfare.[32]

We assume (as a simplification) that agents' social payoffs are based on *long-run* reputations, namely those that will be assigned to contributors and non-contributors after θ becomes publicly known (which ultimately happens with probability 1) –for instance, after everyone has had time to observe average compliance, $\bar{a}$. An agent's action choice is then based on his expectation of those "final" reputation payoffs conditionally on his own v, which is informative about θ since it is drawn from G_θ. Formally, $E[v|a]$ is replaced in (1) by $E_\theta[E[v|a,\theta]\,|\,v]$.

We restrict attention to equilibria in which each agent's optimal strategy is a cutoff $v_F^*(y)$, given that others follow the same rule.[33] In such an equilibrium, the expected reputational return for an agent with valuation v is $E_\theta[\Delta(v_F^*(y) - \theta)|\ F, v]$. We also focus on respectable behaviors, meaning that the support of θ lies to the right of $\theta_0 = c - e$. Strategic complementarity is indeed the relevant case for most existing norm-based interventions, which typically involve relatively low-cost behaviors (e.g., energy conservation).

• **Non-strategic revelation.** We first investigate what occurs when the principal cannot prevent the information from leaking, or it is exogenously revealed by some third party –media, academics, watchdogs, etc. Define, for all $v \in V_\theta$ and distribution F describing agents' prior beliefs about θ, the operator

$$\Phi(v, F) \equiv v - c + y + \mu E_\theta[\Delta(v - \theta)\,|\,F, v]. \tag{17}$$

Since a higher v and a lower F both represent good news about θ (in the sense of first-order stochastic dominance), and since $1 + \mu\Delta'_\theta > 0$ for all θ, Φ is increasing in v and decreasing in F. Under pluralistic ignorance the cutoff $v_F^*(y)$ is thus uniquely defined and, when interior, given as

[30] We model here descriptive interventions, but the prescriptive case could be treated very similarly.

[31] Examples include electricty consumption, recycling, tax compliance, etc. Ali and Bénabou (2010) analyze the "reverse" problem in which the principal seeks to learn about θ, and it is the population who (at least in the aggregate) has more information about it.

[32] Indeed, in any equilibrium defined by a cutoff v^*, e is what a marginal contributor brings to society, whereas $\mu E\left[\Delta_\theta(v^*)\ |\ I\right] + (1+\lambda)y$ is what he costs society (what he must be given, plus the deadweight loss from the incentive), where I is his information set ; recall, finally, that $\Delta_\theta(v^*) < \max\{\bar{v} - v_{\min}, v_{\max} - \bar{v}\}$ for all θ.

[33] For small enough μ other types of equilibria can be ruled out, but in general they cannot: the expected reputational return $E_\theta[R\ |\ v] \equiv E_\theta[E\left[\tilde{v}\ |\ a = 1,\ \theta\right] - E\left[\tilde{v}\ |a = 0,\ \theta\right]\ |\ v]$ need not be increasing in v, so the set of of contributors $\{v\ |\ v - c + y + \mu E_\theta[R\ |\ v] \geq 0\}$ need not be an interval.

the solution to $\Phi(v_F^*(y), F) = 0$. When θ becomes known, it shifts to the familiar $v_\theta^*(y)$, which is decreasing in θ.

Proposition 3 (non-strategic revelation) *Let θ have prior distribution $F(\theta)$ on $[\theta_1, \theta_2] \subset \Theta$ with $\theta_1 > \theta_0$. Compliance is a respectable act both prior to and following agents' learning θ : $\Delta'_\theta < 0$ at both $v_F^*(y)$ and $v_\theta^*(y)$. Moreover,*
(i) There exists a threshold $\hat{\theta}$ such that revelation increases compliance if $\theta > \widehat{\theta}$ and decreases it if $\theta < \widehat{\theta}$.
(ii) The higher is θ, that is, the better news it represents relative to the initial prior, the greater the gain (or the lower the loss) in social welfare resulting from revelation.[34]

Proposition 3 characterizes the aggregate response to disclosure (which is of primary interest to the policymaker), but in doing so it likely understates the extent of heterogeneity in individual responses. In Schultz et al. [2007], for instance, providing comparative data on electricity use led to a reduction for high consumers, but an increase for low consumers. The model provides a natural mechanism for such convergence: with $\Delta' < 0$ (energy conservation is not a heroic activity), individuals who find out they had overestimated the social standard θ feel decreased social pressure to contribute, and those who had underestimated it, increased pressure. At the same time, the dispelling of the average misperception, moving everyone in the same direction, implies that convergence cannot be a general result, and is thus not to be expected in every experiment.[35]

• **Strategic disclosure.** Consider now a principal who has some control over what agents get to learn. With probability q he receives (hard) information about θ; he can then choose to reveal it, or claim that he has no such data (which occurs with probability $1 - q$). Upon disclosure, the cutoff for participation is again $v_\theta^*(y)$. In the absence of disclosure it is $v_{\tilde{F}}^*(y)$, defined by

$$\Phi(v_{\tilde{F}}^*(y), \tilde{F}) = v_{\tilde{F}}^*(y) - c + y + \mu E_\theta[\Delta(v_{\tilde{F}}^*(y) - \theta) \mid \tilde{F}, v_{\tilde{F}}^*(y))] = 0, \tag{18}$$

where $\tilde{F}(\theta)$ denotes the distribution of θ conditional on non-disclosure. Given the above-noted properties of Φ, (18) uniquely defines $v_{\tilde{F}}^*(y)$ (when interior), and this cutoff is increasing in $\tilde{F}$.

Because greater participation always increase social welfare, the principal discloses if and only if $v_\theta^*(y) \leq v_{\tilde{F}}^*(y)$.[36] Since $v_\theta^*(y)$ is decreasing in θ, this occurs for θ greater than some threshold $\tilde{\theta}$. Given such a "good news only" policy by the principal, finally, agents' posterior beliefs in case of

[34] The prior F represents agents' (common) belief about θ before each one learns his own v, drawn from G_θ. Agents' beliefs at the interim stage where they know their own v are heterogenous, and also increasing (stochastically) in θ; thus in general it is uncertain whether a higher θ also represents better news relative to these interim beliefs. Agents' participation threshold $v_{\tilde{F}}^*$ is independent of θ, however, and therefore so is $\tilde{\theta}$.

[35] For instance, extending the model to three actions, $a = 0, 1, 2$, one can construct examples in which learning θ leads to convergence (adjustments occur from both $a = 0$ and $a = 2$ towards $a = 1$), or on the contrary causes divergence (adjustments from $a = 1$ to both extremes).

[36] This is formally shown as part of the proof of Proposition 4.

non-disclosure are

$$\tilde{F}(\theta) = F_{\tilde{\theta}}(\theta) \equiv \begin{cases} \frac{F(\theta)}{qF(\widetilde{\theta})+1-q} & \text{for } \theta \leq \widetilde{\theta} \\ \\ \frac{F(\widetilde{\theta})+[F(\theta)-F(\widetilde{\theta})](1-q)}{qF(\widetilde{\theta})+1-q} & \text{for } \theta \geq \widetilde{\theta}. \end{cases} . \tag{19}$$

An equilibrium therefore corresponds to a $\tilde{\theta}$ that solves the fixed-point equation

$$v^*_{\tilde{\theta}}(y) = v^*_{F_{\tilde{\theta}}}(y). \tag{20}$$

Proposition 4 (strategic revelation) *(i) The principal discloses good news and conceals bad ones: there exists a cutoff $\tilde{\theta} \in (\theta_1, \theta_2)$ such that disclosure occurs if and only if $\theta \geq \tilde{\theta}$.*
(ii) In any stable equilibrium, there is more disclosure ($\tilde{\theta}$ decreases), the higher q is.
(iii) For $\theta \geq \tilde{\theta}$, the principal would have been strictly better off under a commitment to disclose.

The second part of the proposition illustrates the credibility problem resulting from discretionary disclosure. When $\theta \geq \widetilde{\theta}$, with probability q the principal has the information to prove it, so the outcome is the same as under a commitment to disclose. With probability $1-q$ he does not have such proof available, so agents' beliefs are described by $F_{\tilde{\theta}}$, which is dominated by F and thus yields lower compliance, making the principal strictly worse off. Achieving or enhancing credibility requires, as usual, some form of costly signaling, to which we shall turn in the next section.

• **Pluralistic ignorance and "social proof".** In what precedes, the aggregate preference shock θ and average behavior $\bar{a}$ have the same informational content, so it is equivalent for the principal to disclose one or the other, and important that agents do not observe $\bar{a}$ on their own (at least, not as well as the principal). While such is indeed the case for behaviors such as electricity consumption, air pollution or tax evasion, in other instances such as drinking by student peers, shirking by co-workers or the expression of prejudice against women and minorities, people will have frequent and fairly good observations of the norm. Part of the idea of pluralistic ignorance however, is that "social proof" (equilibrium behavior $\bar{a}$) can be a misleading guide to the true underlying group preference (θ), because individuals have trouble parsing out the contribution of perceived social pressure to the observed outcome.

There are two ways to accommodate this more "resilient" form of pluralistic ignorance. First, both θ and μ may be subject to aggregate shocks, leading to a signal-extraction problem in interpreting $\bar{a}$.[37] Alternatively, pooling can also make $\bar{a}$ imperfectly informative, thereby restoring the scope for a principal's disclosures (strategic or not) to affect agents' perceptions of Δ_θ, and hence their behavior. For instance, relaxing the assumptions of continuously distributed θ and interior participation cutoff, let θ take value θ_L or θ_H, such that: (i) when agents know that $\theta = \theta_H$

[37] This is done in Ali and Bénabou [2010], with agents receiving noisy idiosyncratic signals about both aggregate shocks, from their own preferences.

(respectively, $\theta = \theta_L$) there is positive participation, $0 < \overline{a} \leq 1$ (respectively, zero participation, $\bar{a} = 0$); (ii) the prior probability that $\theta = \theta_L$ is high enough that, when agents are uninformed, no one contributing remains the (generically unique) equilibrium.[38] Thus, pluralistic ignorance prevails when agents observe $\bar{a} = 0$, and dispelling it by (credibly) disclosing that $\theta = \theta_H$ increases participation.

4 The expressive function of law

4.1 Law and societal values

• **Empirical evidence.** The idea that the law is more than a set of prices but also serves to convey a society's norms of behavior has long been discussed in the legal literature, albeit mostly in the form of thought experiments and debates over the symbolic content of legislations on sexual behavior, drinking or smoking in public, religious displays, flag-burning, etc. More recent work brings two types of empirical evidence to bear on the issue.

The first one documents the effectiveness of "symbolic" fines or sanctions –incentives too small to matter through the standard price channel, but which significantly raise compliance when combined with a simple statement or reminder of one's moral duty. Funk (2007) shows that the repeal of mandatory-voting laws in Switzerland led to statistically significant declines in turnout in cantons where the law had stipulated a trivial fine (about 1 Euro) for non-voters, whereas it had no impact where the mandate was purely declarative. Similar results are demonstrated by Galbiati and Vertova (2008) in a public-goods experiment. Stating an "obligation" to contribute above some minimum has only a weak effect, and "non-binding" incentives (fines small enough that complete free-riding remains a dominant strategy for material payoffs) have none. When the two are combined, however, contributions double, and about half of this impact operates through a shift in subjects' beliefs about what others will give.

These findings suggest a signaling role of incentives (at least as perceived by subjects), and a second set of papers specifically document such a mechanism. Tyran and Feld (2006) show that "mild law" –penalties insufficient to deter free-riding– has no effect when it is exogenously imposed in a public-goods game, but significantly raises compliance when endogenously chosen through an initial vote by the participants. Belief change is again a key element, as more votes favoring mild sanctions lead agents to expect higher compliance by others, and these expectations largely explain contributions levels (between conditions and subjects). In Galbiati et al. (2010), a pair of players engaged in a coordination (minimum-effort) game may be provided with substantial incentives. When these are exogenously imposed by the experimenter, they lead to increased effort and expectations that the partner will also respond by contributing more. When they are endogenously imposed by a benevolent third party who has observed the pair's behavior in a previous round, in contrast, subjects who had provided high effort become pessimistic about their

[38] When dealing with corner equilbria, we restrict attention to those satisfying the D1 criterion.

partner's contribution and accordingly reduce their own, making the sanctions counterproductive.[39] Bremzeny et al. [2011] also test and confirm the "bad news effects" of choosing strong incentives, this time in a setting where the principal has private information about the difficulty of the (single) agent's task.

• **Modeling expressive law.** In what follows, we formalize the expressive content of (optimal) incentives in the presence of norms, or reputational payoffs more generally. We investigate in particular a question on which neither the legal literature nor existing experiments offer consistent insights: when should expressive concerns make the law milder, or on the contrary tougher?

When a legislator or principal with privileged information about "community standards" θ or compliance $\bar{a}$ sets material incentives –law, rewards, penalties– these will inevitably convey a message about those standards, and thereby shape agents' *understanding of prevailing social norms.* Formally, the model will now involve *two-sided signaling:* agents signal their idiosyncratic types,while the principal signals the aggregate state of societal preferences.[40]

For simplicity, let θ be perfectly known by the principal.[41] Agents only know that it lies in some subinterval (θ_1, θ_2) of Θ with $\theta_1 > \theta_0$ ($a = 1$ is then a "respectable" behavior); or, alternatively, that it lies in some $(\theta_1, \theta_2) \subset \Theta$ with $\theta_2 < \theta_0$ ("admirable" behavior). Technically, this "one-sided-support" restriction is made necessary by the non-monotonic nature of the policy under symmetric information, which implies that a separating equilibrium cannot exist over all θ.

We look for a separating equilibrium where the planner's policy $y^{AI}(\theta)$ is strictly increasing on Θ if that interval lies to the left of θ_0, and strictly decreasing if it lies to the right. Agents can then invert the policy and infer the true θ as the unique solution $\hat{\theta}(y) \in \Theta$ to $y^{AI}(\hat{\theta}(y)) \equiv y$. The resulting cutoff (here again assumed interior) is then $v^*_{\hat{\theta}(y)}(y)$, so the planner maximizes[42]

$$W^{AI}_{\theta}(y) \equiv \int_{v^*_{\hat{\theta}(y)}(y)}^{+\infty} (e - \lambda y + v - c)\, g_{\theta}(v) dv + \mu(\bar{v} + \theta). \tag{21}$$

Provided W^{AI}_{θ} is quasiconcave in y for all θ and λ (which holds for λ small enough), the optimum

[39] The source of complementarity is here the nature of agents' payoffs, whereas in our model it is the reputation-based social norm (when $\Delta' < 0$). The common and key elements are the choice of incentives by an informed principal and agents' inferences from it about how others are likely to act.

[40] As in Bénabou and Tirole [2003], Ellingsen and Johannesson [2008] and Herold [2010] there is an informed-principal problem, but now the feature of the "task" which agents try to infer –the social pressure $\mu\Delta(v^*)$– embodies everyone's equilibrium actions and beliefs. In Sliwka [2008] and van der Weele [2008] incentives also convey information about the distribution of preferences but the nature of normative influences is quite different. In Sliwka [2008] social complementarities operate through "conformist" types, whose preference is to mimic whatever action the majority chooses. In van der Weele [2008] they involve "reciprocal altruists", whose taste for contributing to a public good rises with aggregate contributions (v is increasing in $\bar{a}$). Our model has no built-in complementarity; conformity ($\Delta' < 0$) or distinction ($\Delta' > 0$) effects arise endogenously, and we analyze expressive law in both cases.

[41] The main simplification is that, in a separating equilibrium, agents will not use their own realizations of $\tilde{v}$ to make inferences about θ, since it is fully revealed by y.

[42] We continue to assume that reputations (being long-term objects), are "consumed" ex-post, once agents have learned the true θ from, say, observing $\bar{a}$. This is why the last term is $\mu(\bar{v} + \theta)$, rather than $\mu(\bar{v} + \hat{\theta}(y))$.

is given, on each side of θ_0, by the first-order condition

$$\left(\frac{e-c-\lambda y+v^*_{\hat{\theta}(y)}(y))}{1+\mu\Delta'_{\hat{\theta}(y)}(v^*_{\hat{\theta}(y)}(y))}\right)\left(1-\mu\Delta'_{\hat{\theta}(y)}(v^*_{\hat{\theta}(y)}(y)))\ \hat{\theta}'(y)\right)=\frac{\lambda}{h_\theta(v^*_{\hat{\theta}(y)}(y))}. \tag{22}$$

Together with the equilibrium condition $\hat{\theta}=(y^{AI})^{-1}$, this defines a first-order differential equation in $y^{AI}(\theta)$:

$$\left(\frac{e-c-\lambda y^{AI}(\theta)+v^*_\theta(y^{AI}(\theta))}{1+\mu\Delta'_\theta(v^*_\theta(y^{AI}(\theta)))}\right)\left(1-\frac{\mu\Delta'_\theta(v^*_\theta(y^{AI}(\theta)))}{(y^{AI})'(\theta)}\right)=\frac{\lambda}{h_\theta(v^*_\theta(y^{AI}(\theta)))}. \tag{23}$$

The difference with (12) reflects the planner's taking into account that agents will draw inferences from his policy about where societal values lie (term $\hat{\theta}'=1/y^{AI\prime}$) and the social sanctions and rewards they will face as a result (term $\mu\Delta'_\theta(v^*_\theta)$). This *informational multiplier,* $1-\mu\Delta'_\theta\hat{\theta}'$, embodying the *expressive content of the law,* combines with the previously analyzed social multiplier, $1/(1+\mu\Delta'_\theta)$, to amplify or dampen agents' response to incentives, and therefore the optimal level of y. Once again, the case of no deadweight loss provides a useful benchmark.

Proposition 5 *For $\lambda=0$, the first-best symmetric-information solution $y^{FB}(\theta)=e-\mu\Delta_\theta(c-e)$ remains an equilibrium on (θ_1,θ_2), and it is the unique separating one.*

Intuitively, when the first-best can be achieved with non-distortionary incentives there is no need to resort to the norm as a substitute, and hence no need either for any "expressiveness" in the law designed to manipulate that norm. The more realistic case $\lambda>0$ requires solving the differential equation (23), with $y^{AI}=y^{SI}$ at the inner boundary. Because the Lipschitz conditions do not hold everywhere, we need to again take λ to be relatively small and impose the support restriction on θ described above.

Lemma 1 *Let (θ_1,θ_2) be any subinterval of Θ with $\theta_1>\theta_0$ (respectively, $\theta_2<\theta_0$). For all λ below some $\bar{\lambda}>0$, the differential equation (23) with boundary condition $y^{AI}(\theta_1)=y^{SI}(\theta_1)$ (respectively, $y^{AI}(\theta_2)=y^{SI}(\theta_2)$) has a unique solution on (θ_1,θ_2).*

In the process of proving existence and uniqueness we also establish the following key properties of the equilibrium policy, illustrated on Figure 2 (bottom curve).

Proposition 6 (law expressing societal standards) *(i) For all λ below some $\bar{\lambda}>0$, the equilibrium incentive $y^{AI}(\theta)$ is strictly positive, increasing to the left of θ_0, and decreasing to its right. (ii) Whether the prosocial action is respectable or admirable ($\theta_1>\theta_0$ or $\theta_2<\theta_0$), the principal always sets lower-powered incentives under asymmetric information, $y^*_{AI}(\theta)<y^*_{FI}(\theta)$ for all $\theta\in(\theta_1,\theta_2)$, and compliance is lower.*

For a respectable activity, a lower y *credibly* conveys the message: "everyone does it, except the most disreputable people, who suffer great stigma; this is why we do not need to provide strong incentives". For an admirable activity, a lower y conveys the message "the glory suffices: contributors are rare heroes, who reap such social esteem that additional incentives are unnecessary". Another interesting implication is that *expressive law is more responsive* to changes in societal values than "standard" law, at least on average, and especially for modal acts, where both are used the most: on both sides of θ_0, y^{AI} and y^{SI} start from a common value but y^{AI} is everywhere below, so its average slope is steeper. At the initial point, in particular, $y^{AI\prime}(\theta) = -\infty$, so clearly this function has (much) greater slope in a neighborhood of θ_1 and θ_2.[43]

4.2 Spillovers across spheres of behavior

What people learn or perceive concerning others' degree of prosociality or selfishness carries over between activities, leading to spillovers in behavior, both good and bad.[44] Given such "contagion", a principal setting law or other incentives for one activity needs to take into account how this will affect people's views of *general societal norms,* and thus their behavior in other realms. For instance, hard incentives conveying the sense that "society is rotten" (e.g., endemic corruption or tax evasion) can be damaging in the case of case of respectable activities where $\Delta' < 0$.

A simple case will convey the main insight, but it can be substantially generalized. Agents engage in two activities, a and b, both involving 0-1 decisions:

(i) Informal interactions. An individual's $a-$behavior is observed by other private citizens, giving rise to social sanctions and rewards, but not verifiable by the government (or other principal), who therefore cannot use incentives: cooperating with others, helping, contributing to public goods, refraining from rent-seeking, etc. Formally, $y_a = 0$ and $\mu_a = \mu > 0$.

(ii) Formal interactions. An individual's $b-$behavior, conversely, is observed and verifiable by the principal or government, but not by other private citizens. Transactions between agents and principal are of this nature, such as paying or evading taxes, an employee's productivity or a civil servant's record of corruption complaints, etc. Other agents may also be less able than the principal to sort through excuses for bad behavior (e.g., was the claimed tax deduction justified or not)? Formally, $y_b = y > 0$ and $\mu_b = 0$.

For simplicity, let the same intrinsic preferences –a general degree of prosociality– drive both activities: $v_a = v_b = v$. More generally, it suffices that the two values –or even just their distributions– be correlated. An agent chooses $b = 1$ if $v - c_b + y \geq 0$, or equivalently $v \geq c_b - y \equiv v_b^*(y)$. He

[43] Setting $\theta = \theta_1$ (say) in (23) and using the fact that $y^{AI}(\theta_1) \equiv y^{SI}(\theta_1)$, which satisfies (12), implies that $\mu\Delta'_\theta(v^*(y^{AI}_{\theta_1}(\theta_1)))/(y^{AI})'(\theta_1) = 0$. Since the numerator is strictly negative, the denominator must be infinite.

[44] For instance, Keizer et al. (2008) posted flyers (advertisements) on 77 bicycles parked along a wall and observed that the fraction of owners tossing them on the ground doubled (from one third to two thirds) after graffiti had been painted on the wall. Similarly, leaving a € 5 bill sticking out of someone's mailbox, they observed that 13% of people pocketed it when the surroundings were clean, but 23% did when there was trash lying around.

chooses $a = 1$ if $v \geq v_a^*(y)$, defined by :

$$v_a^*(y) - c_a + \mu \Delta_{\hat{\theta}(y)}(v_a^*(y)) = 0. \tag{24}$$

Note that v_a^* depends on y (which only incentivizes b behavior) solely through the inferences drawn about θ. The government or principal maximizes

$$W_\theta^{AI}(y) = \int_{v_b^*(y)}^{+\infty} (e_b + v - c_b - \lambda y)\, g_\theta(v) dv + \int_{v_a^*(y)}^{+\infty} (e_a + v - c_a)\, g_\theta(v) dv + \mu(\bar{v} + \theta), \tag{25}$$

leading to the first-order-condition:

$$\begin{aligned} \frac{\partial W_\theta^{AI}(y)}{\partial y} &= (e_b + v_b^*(y) - c_b - \lambda y)\, g_\theta(v_b^*(y)) - \lambda \left[1 - G_\theta(v_b^*(y)\right] \\ &\quad - (e_a - c_a + v_a^*(y))\, g_\theta(v_a^*(y)) \left(\frac{\partial v_a^*(y)}{\partial y} \right). \end{aligned} \tag{26}$$

Under symmetric information the last term vanishes, so the first-best ($\lambda = 0$) policy is given by $v_b^*(y) = c_b - e_b$, hence $y^{FB} \equiv e_b$ for all θ. When $\lambda > 0$, the Ramsey condition takes the form

$$y^{SI}(\theta) = \frac{e_b}{1+\lambda} - \frac{\lambda}{(1+\lambda)\, h(v_b^*(y^{SI}(\theta)) - \theta)}. \tag{27}$$

which has a unique solution $0 < y^{SI}(\theta) < y^{FB}(\theta)/(1+\lambda)$, decreasing in θ and λ, as long as the hazard rate $h(v)$ is increasing and $0 < \lambda < h\left(v_b^*(0) - \theta\right) e_b$. We shall assume both of these conditions.

Under asymmetric information the social cost of a marginal rise in y now includes, on top of the usual rents to inframarginal agents choosing $b = 1$, a reduction in $\bar{a}$ that arises from agents' inferring that they face a "worse" society and therefore weaker social enforcement in their a decisions. The optimal policy, taking account of these expressive spillovers, is given by $y^{AI} = \hat{\theta}^{-1}$, where $\hat{\theta}$ solves the differential equation

$$e_b - (1+\lambda)y = \frac{\lambda}{h_{\hat{\theta}(y)}(v_b^*(y))} + (e_a - c_a + v_a^*(y)) \left(\frac{g_{\hat{\theta}(y)}(v_a^*(y))}{g_{\hat{\theta}(y)}(v_b^*(y)} \right) \left(\frac{\hat{\theta}'(y) \cdot \mu \Delta'_{\hat{\theta}(y)}(v_a^*(y))}{1 + \mu \Delta'_{\hat{\theta}(y)}(v_a^*(y))} \right) \tag{28}$$

with boundary condition $\hat{\theta}\left(y^{SI}(\theta_1)\right) = \theta_1$. For simplicity we focus on the case where, under symmetric-information, the a activity is in the respectable range, $\Delta'_\theta(v_a^*(y^{SI}(\theta))) < 0$, and reputational pressure is insufficient to ensure the first-best, meaning that $v_a^*(y^{SI}(\theta)) > c_a - e_a$. The latter condition is ensured as long as e_a satisfies (14), and the former provided that θ has support in some $(\theta_1, \theta_2) \subset \Theta$ with $\theta_1 > e_a - e_b - \mu\Delta(0)$, and λ is not too large.[45]

[45] Indeed, $v_a^*(y^{SI}(\theta)) < \theta$ if $\theta - c_a + y^{SI}(\theta) + \mu\Delta_\theta(\theta) > 0$; recall that $\Delta_\theta(\theta) = \Delta(0)$, while for λ small enough $y^{SI}(\theta)$ is close to $y^{FB}(\theta) = e_b$.

Proposition 7 (expressive spillovers) *When the socially-enforced behavior a is respectable or admirable ($\theta_1 > \theta_0$), the principal sets lower-powered incentives for the incentivized action b under asymmetric information: $y^{AI}(\theta) < y^{SI}(\theta)$ for all θ, with $y^{AI}(\theta)$ decreasing everywhere. Participation in b is lower than under full information, participation in a is unchanged (since θ is revealed).*

These results are reminiscent of those in the multi-tasking literature, but operate through a different mechanism. The literature has emphasized the hazards associated with giving incentives in one task when they cannot also be adjusted on another, competing one. Relatedly, when some aspects of performance are unverifiable, it may be desirable to leave other, verifiable ones, unspecified (see, e.g., Baker et al. (1994), Bernheim and Whinston (1998)). Rather than effort substitution, crowding out occurs in our case through what incentives reveal about the standards of behavior to be expected in other activities.

• **Society's resistance to economists' prescriptions.** In nearly all countries, economists' typical message about the effectiveness and desirable normative properties of incentives meets with considerable resistance. Examples includes tradeable pollution permits, financial incentives for students, teachers and civil servants, unemployment benefits that decrease over time to encourage job search, layoff taxes rather than regulation, markets for blood and organs, taxes rather than prohibition for drugs and prostitution, etc. While misinformation and special-interest considerations may be relevant in some cases, they do not come close to explaining the nearly universal reluctance toward what many in the lay public perceive as a nefarious "commodification" of human activity.

Our framework can be used to shed some light on this phenomenon. Economists typically bring a message, both positive (based on empirical studies) and normative (policy recommendations) that is *bad news about human nature* and behavior. In terms of the model, policies implementing the standard advice to "put a price on everything" constitute strong public signals (and daily reminders) that altruism is generally low (a low θ), and greed generally high.[46] Society may then resist such a message and the policies embodying it, for two reasons.

First, individuals and societies alike often just *do not like* to hear bad news, preferring to maintain pleasant (but ultimately, costly) illusions about themselves. Such is the case, for instance, with political and economic ideologies, national founding "myths", etc.[47] A related form of purely affect-driven preference for collective self-image will be analyzed in the next section.

Second, societies could be justifiably concerned about spillovers from policies that express too dim or mercantile a view of human nature. For instance, economists' lessons may be drawn predominantly from b -type behaviors, where incentives are easily available and social norms weak. Less attention might be paid –perhaps simply for lack of data– to a -type behaviors, in which incentives are unavailable and reliance on social norms important.[48] As shown above, in such cases

[46] See footnote 8 for a more general version of the model in which agents differ in their marginal utility of money v_y. The distribution of v_y can then also be subject to aggregate uncertainty, and thus revealed by policy.

[47] For models of the persistence of collective ideologies through (equilibrium) cognitive dissonance, see Bénabou and Tirole [2006b] and Bénabou [2008].

[48] Note the striking contrast between economists' typical findings and message to society of "θ and μ are *lower* than

bringing bad news about θ, by stating and especially by concretely signaling that strong incentives are effective or needed in b, has the collateral effect of undermining the social norms in a. This creates a need for incentives to replace them, but the nature of the activity can make this a much less cost-effective way of achieving compliance, resulting in a welfare loss.

4.3 When expressiveness makes law tougher

We have so far seen how expressive concerns about the nature of society (others' goodness or mediocrity) always makes, perhaps surprisingly, the law more gentle. So when do signaling considerations lead instead to stronger incentives (along the lines of "lock them up and throw away the key: we need to send a message")?

Intuition suggests that this should occur when uncertainty bears on *how damageable* to society selfish behavior is –or, conversely, on how important the social spillovers (e) from good behavior are. Also required, however, are preferences linking agents' intrinsic motivations, v, to the social value of their contribution, e.[49] Since such a link does seem intuitive, we now consider a variant of the model in which intrinsic utilities are $v_a = ve$, with $v \sim G(v)$ and $G(\cdot)$ having the same properties as before.[50] Reputation (or self-image) still bears on v, which represents an agent's general degree of social concern. Agents know their own v, while the principal knows e –e.g., how damaging to the environment CO_2 emissions are, how much good \$1 can do in poor countries, the negative externalities created by drunk driving, drugs, etc.

Under symmetric information, the cutoff is now given by $ev_e^*(y) - c + y + \mu\Delta(v_e^*(y)) \equiv 0$, leading again to a modified version of Pigovian taxation. Thus, for $\lambda = 0$,

$$y^{FB}(e) = e - \mu\Delta\left(\left(c - e\right)/e\right). \tag{29}$$

In general, it could be the case that $y^{FB} < 0$ (people demonstrate great social concern by paying significant costs for trivially small social benefits), or that $dy^{FB}/de < 0$ (reputation increases more than 1 for 1 with e, hence also the reputation tax). We shall abstract from such cases, as we are interested in relatively large e's.[51]

Under asymmetric information we look for a separating equilibrium on the support $[e_1, e_2]$ of e such that $y^{AI}(e)$ is increasing everywhere, as are $y^{FB}(e)$ and $y^{SI}(e)$. Agents then infer e as the solution to $y^{AI}(\hat{e}(y)) \equiv y$, so the participation cutoff is $v^*_{\hat{e}(y)}(y)$, with

most people think", leading to a prescription for incentives; and those of social psychologists working on norms-based interventions, which are that "θ and μ are *higher* than most people think" (pessimistic pluralistic ignorance).

[49] Otherwise, learning from y that e is high (say) has no impact on the reputational pressure Δ that agents face, once the direct impact of y on v^* is accounted for; see (5).

[50] In small-group interactions, such a link is consistent with (and follows from) "pure", consequentialist altruism: an individual values the difference he makes to $e\bar{a}$. With large numbers, it is not, since each individual has a negligible effect. Intrinsic motivation must then arise from a pure preference for ("joy of") giving. Nonetheless, it remains sensible that one should derive more intrinsic utility from giving to more useful causes, rather than unimportant ones. In particular, this is what will arise from a Kantian-type or similar rule-based reasoning.

[51] Thus (14) suffices to ensure $y^{FB} > 0$ and , since Δ' is bounded, a low enough μ ensures that $dy^{FB}/de > 0$. For λ small enough, these properties remain true for the full information second-best policy, $y^{FI}(e)$.

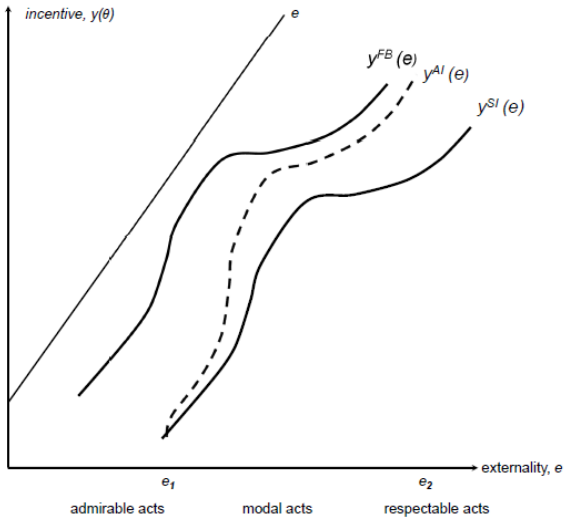

Figure 3: law and social externalities

$$\frac{dv^*_{\hat{e}(y)}(y)}{dy} = -\frac{1 + v^*_{\hat{e}(y)}(y)\hat{e}'(y)}{\hat{e}(y) + \mu\Delta'(v^*_{\hat{e}(y)}(y))}. \tag{30}$$

Knowing this, the principal maximizes

$$W^{AI}_e(y) = \int_{v^*_{\hat{e}(y)}(y)}^{+\infty} (e + ev - c - \lambda y)\, g(v)dv + \mu\bar{v},$$

leading to the differential equation in $\hat{e}(y)$

$$\left(\frac{e\left[1 + v^*_{\hat{e}(y)}(y)\right] - c - \lambda y}{\hat{e}(y) + \mu\Delta'(v^*_{\hat{e}(y)}(y))}\right)\left[1 + v^*_{\hat{e}(y)}(y)\hat{e}'(y)\right] = \frac{\lambda}{h(v^*_{\hat{e}(y)}(y))} \tag{31}$$

or, conversely, in $y^{AI}(e)$:

$$\left(\frac{e\left[1 + v_e(y^{AI}(e))\right] - c - \lambda y^{AI}(e)}{e + \mu\Delta'(v_e(y^{AI}(e)))}\right)\left(1 + \frac{v^*_e(y)}{y^{AI\prime}(e)}\right) = \frac{\lambda}{h(v_e(y^{AI}(e)))}, \tag{32}$$

with boundary condition $y^{AI}(e_1) = y^{SI}(e_1)$. Since $y^{SI}(e)$ is increasing we expect that for λ small, so is $y^{AI}(e)$, implying that the "expressive" term $v^*_{\hat{e}(y)}(y)\ \hat{e}'(y)$ has the sign opposite to that obtained in (22) for signaling over θ. These intuitions lead to the following results, illustrated in Figure 3.

Proposition 8 (law expressing externalities) *Whether the prosocial action is respectable or admirable, for all $\lambda > 0$ below some $\bar{\lambda} > 0$, the principal sets higher-powered incentives under asymmetric information: $y^{AI}(e) > y^{SI}(e)$ for all e, and participation is correspondingly higher. The function $y^{AI}(e)$ is everywhere increasing on $[e_1, e_2]$.*

4.4 Cruel and unusual punishments

As legal sanctions for antisocial behavior (harm to others, negative externalities), standard social-welfare considerations generally argue for using fines, compensation of victims, community service and other "efficient" punishments. Such alternatives are, however, politically unpopular: large fractions of the electorate demand not only long and harsh prison sentences but also various forms of public humiliation.[52] In many countries, public executions and corporal punishments are still the law of the land and, when public, heavily attended.[53] At the same time, a growing number of nations are renouncing what they deem "cruel and unusual" punishments or means of coercion.[54] Such decisions, moreover, are not based on any real considerations of optimal deterrence, but on "what it makes us", what "civilized" peoples do or don't do –in other words, on *expressive reasons.*

What exactly is it, however, that makes caning, whipping, flogging, public shaming and the like qualitatively different –and "expressively" worse– than very long prison sentences or drastic financial penalties, especially when the condemned himself would rather take the pain or shame?

The key variable in the answer we develop here could be called "the banality of evil": a fraction κ (for "cruel") of agents in society actually enjoy the suffering of others (either all others, or only the guilty, in which case this is a taste for vengeance). Seeing criminals, cheaters and other law-breakers punished harshly –a high level of physical or psychological pain, p– and publicly (being a spectator enhances this form of "consumption") is an opportunity, and possibly an excuse, to obtain such enjoyment. The total utility flow thus derived is then $\kappa pG\left(v^*(p)\right)$, where $v^*(p)$ is the threshold below which people break the law and are subject (with some probability) to the punishment p.

The second important assumption is that many people do not like to think –find scary, disturbing to acknowledge– that their society or community comprises a lot of cruel or vengeful types. These could, for instance, hurt them in certain circumstances (traffic dispute, breakdown of law and order, etc.). Bad news about human nature is also, inevitably, bad news about oneself. We shall assume here a simple (linear) affective dislike for ugly truths about the banality of evil: agent's utility functions are now $U - \beta E\left[\kappa \mid p\right]$, with U still given by (1) and β measuring the intensity of preferences over beliefs about $\kappa \in [\kappa_1, \kappa_2]$.[55]

The government or legislator knows more about κ –having access to observations from the judicial system, prison life, how people behave in blackouts, wars, etc. With such knowledge it sets the level $p = p(\kappa)$ of "painful" penalties levied on those who choose $a = 0$ in serious offences such

[52] See, e.g., Kahan [1996, 1998] who argues that alternative sentences (e.g., community service) are seen by the public as not carrying appropriate symbolism – conferring insufficient stigma on the condemned and devaluing victims –whereas shaming sanctions, such as practiced in several U.S. states (internet postings, compulsory lawn signs, license plates, etc.) better satisfy this demand.

[53] To take only a rich and educated country as an example, Singaporean law allows caning for over 30 offences and makes it mandatory punishment for several, such as rape or drug trafficking. In 2007 there were 6,404 such sentences. Caning is also used in prisons, the military, and schools.

[54] For instance, the European Community makes renouncing the death penalty a precondition for membership, and the United States declares (with some debate over exceptions) torture contrary to "American values".

[55] Alternatively, one could endogenize the presence of such a term in the principal's objective function from instrumental concerns, arising for instance from coordination externalities in investment.

as murder, theft, fraud, drunk driving, child abuse, etc. We make p here the only policy tool, but one could also allow for "non-cruel" incentives y such as fines, jail time, community service, etc. The key points should be unchanged, as these do not generate as much enjoyment for cruel types, and are also more costly. In setting policy, let $0 < \gamma \leq 1$ be the weight placed on the utility of the cruel types –or equivalently, their political influence.[56] Finally, the infliction of harsh punishments involves direct enforcement costs, represented by a unit shadow cost $\lambda \geq 0$.

Given a level of harshness p, agents infer κ as the solution $\hat{\kappa}(p)$ to $p(\hat{\kappa}(p)) = p$. The cutoff between law-abiding and law-breaking is then determined by $v^*(p) - c + p + \mu\Delta(v^*(p)) = 0$, and the planner maximizes

$$W_{\kappa}^{AI}(p) = \int_{v^*(p)}^{+\infty} (e + v - c)\, g(v)dv + \mu(\bar{v} + \theta) - p\,(\lambda - \kappa\gamma)\, G\,(v^*(p)) - \beta\hat{\kappa}(p). \tag{33}$$

The combined optimality and equilibrium conditions now yield

$$\left(\frac{e - c + v^*(p) + p\,(\lambda - \kappa\gamma)}{1 + \mu\Delta'(v^*(p))}\right) g(v^*(p)) = (\lambda - \kappa\gamma)\, G(v^*(p)) + \beta\hat{\kappa}'(p). \tag{34}$$

Under symmetric information, or in the absence of expressive considerations ($\beta = 0$), we get back the case of Section 4.1, with $\lambda' \equiv \lambda - \kappa\gamma$. Under asymmetric information, $p^{AI}(\kappa)$ is given by (34) with the boundary condition $p^{AI}(\kappa_1) = p^{FI}(\kappa_1)$.

Proposition 9 (civilized punishments) *Let $eg(c-e)/G(c-e) < 1$. For all μ below some $\bar{\mu} > 0$ and all κ such that $|\lambda - \gamma\kappa|$ is below some $\bar{\lambda} > 0$,*

(i) The symmetric-information policy $p^{FI}(\kappa)$ and its asymmetric-information counterpart $p^{AI}(\kappa)$ are both increasing in κ.

(ii) Punishments are less harsh under asymmetric information than under full information: $p^{AI}(\kappa) < p^{FI}(\kappa)$ for all $\kappa \in (\kappa_1, \kappa_2]$, and compliance is correspondingly lower.

Implications. The presence of κ-types reduces the effective deadweight loss from punishment (to the extent that society internalizes their utility), from λ to $\lambda - \kappa\gamma$. This implies harsher sanctions, closer to or in excess of first-best deterrence level ($\lambda = 0$). On the other hand, people's desire to believe, or signify to the world, that they are part of a non-barbaric society leads to restrictions on cruel punishments, whether or not efficient at the margin (above or below the first-best). This is captured by the last term in (34), which distinguishes such punishments from standard incentives. To the extent that cruel types' enjoyment is enhanced by witnessing harsh treatments being administered, expressive concerns will also –and first of all– lead to eliminating *public displays* of judicially sanctioned pain and executions. Indeed, much of what goes on in prisons is more cruel than some forms of corporal punishments or public shaming, but it remains out of sight.[57]

[56] Equivalently, uncertainty could be over the political weight γ of cruel types, rather than their number κ. Only the product $\kappa\gamma$ matters.

[57] Note also that –here as in previous cases– the equilibrium is separating, so ultimately no one is fooled and

5 Robustness and extensions

• **Action space.** The zero-one assumption yields very sharp results concerning the shape or equilibrium reputations and optimal incentives. The underlying insights are much more general, however. With several discrete actions, for instance, there will be multiple reputation levels corresponding to successive intervals of v, but the pursuit of reputation will remain a zero-sum gam and be reflected in a nonlinear tax embodied in the optimal incentive. Reputation levels will again change with shifts in the distribution of preferences, leading to a scope for norm manipulation and expressive law. In particular, whenever the principal has private information about any parameter directly affecting equilibrium reputations he will weaken incentives, so as to economize on their cost, by signaling that social payoffs are high. This remains true even with continuous actions, as long as preference heterogeneity is multidimensional (e.g., prosocial motivation and utility for money, or prosocial motivation and image concerns), so that behavior remains an imperfect signal of type.[58]

• **Reputational payoffs.** Most of the paper's results were derived under the joint assumption that types are distributed according to a single-peaked (possibly monotonic) density and that reputational payoffs are linear in the posterior belief about v. This is a natural benchmark, but the results can be substantially generalized. Thus, if reputational payoffs are of the form $E[\varphi(v)|a]$, where φ is increasing and differentiable, this operates much as a change in the distribution of v – more specifically, in its skewness. Let $w \equiv \varphi(v)$, which has c.d.f. $F \equiv G \circ \varphi^{-1}$ and associated density f, and denote $\Delta_F(w^*) = \mathcal{M}_F^+(w^*) - \mathcal{M}_F^-(w^*)$ the associated linear reputational payoff. The equilibrium cutoff is now given by $v^* - c + y + \mu\Delta_F(\varphi(v^*)) = 0$, and the social multiplier equals $[1 + \Delta_F'(w^*)\varphi'(w^*)]^{-1}$. Actions are therefore strategic complements when F is concave (f is decreasing), which by definition means that G is more concave than φ (e.g., g is decreasing and φ is convex). Conversely, they are strategic substitutes when G is more convex than φ. Similarly, f is single-peaked if G is more convex than φ up to a point, then more concave.[59] The analysis of optimal incentives under symmetric or asymmetric information then proceeds along lines qualitatively similar to those of the paper, albeit with more complicated expressions.

These observations allow the global monotonicity or single-peakedness of equilibrium reputation and optimal incentives to be preserved for certain more general payoff specifications. Even when they are not, however, it is important to note that the paper's key results do not strictly

average welfare would be higher (more generally, the principal would be better off ex-ante) if he could commit to acting according to the truth, whether pleasant or not, about human nature. This is a standard type of inefficient-signaling result, but it also reflects specific assumptions that one may want to relax. Thus, with discrete types there could be pooling equilibria; ignorance could then generate (ex-ante) welfare gains if agents' utility is concave in $\hat{\kappa}$, or if increasing returns in cooperative investments make social payoffs nonlinear in trust. Alternatively, a fraction of agents may be naifs rather than perfect Bayesians.

[58] For agents' equilibrium behavior and reputations in such a model (with an exogenously given level of incentives), see Bénabou and Tirole [2006a]. In this continuous specification the mean of the preference distribution (the v's) no longer affects reputations, but its variances and covariances do (as does the mean of agents' μ's).

[59] If reputational payoffs are of the form $\psi(E[v|a])$, similarly, the reputational return $\Delta_\psi(v)$ is such that $\Delta_\psi' = \left(\mathcal{M}_F^+\right)'\psi' - \left(\mathcal{M}_F^-\right)'\psi'$, so a concave ψ pushes toward complementarity, and a convex one toward substitutability.

require such global properties. Proposition 2 applies whatever is the shape of g and thus Δ, and Propositions 2 and 6 apply on any interval of θ or e such that $\Delta'(v^*)$ remains bounded away from zero. Furthermore, whereas the size of the multiplier and the direction in which a principal wants to strategically bias his disclosures hinge on the sign of Δ', our results about when expressive concerns lead to weaker or tougher laws (Propositions 6 and 8) are entirely independent of it. They would thus also carry over to most other forms of reputational payoffs, including for instance non-monotonic "preferences for conformity" as in Bernheim [1994].

• **The scope of norms.** While we do not explicitly model the enforcement of social sanctions and rewards that typically underlie reputational payoffs, one can already identify several intuitive factors that will also contribute to the emergence of norms ($\Delta' < 0$) or a quest for distinction ($\Delta' > 0$). A first one is asymmetry in the feasibility of social rewards and punishments. In most public-goods experiments, for instance, agents can punish free-riders but not reward model citizens. This perhaps reflects the fact that, in many decentralized interactions, it is cheaper to hurt than to reward someone (at the same monetary or utility-equivalent level). In such cases the socially enforced payoff reduces to $-\mu\mathcal{M}^-(v)$, leading to unmitigated complementarity and strong norms. The same effect obtains when good actions may be unrelated to type (false positives) with some probability, while conversely the existence of plausible but unverifiable excuses for not contributing weakens the inferences that can be drawn from it, thus dampening stigma relative to honor.[60]

Next, when agents are more concerned about a given behavior (say, affecting the environment), they tend to pay more attention to how others behave along this dimension, and to be more willing to enforce social sanctions on them. By making μ increasing in θ, this will cause optimal incentives to decline faster (or, where $\Delta' > 0$, increase slower) with average societal concerns, and thus further amplify the "bad news" about average preferences expressed by strong incentives. Finally, participating more in some behavior, such as volunteering or fighting for one's group, can make it easier to observe who else does so. In this case μ depends positively on $\bar{a}$, resulting in a larger social multiplier and a greater likelihood that powerful social norms will emerge, but again with a qualitatively similar impact on how optimal incentives vary with societal preferences.

• **Individual and social preferences.** We have assumed that agents try to signal their commitment (v) to a specific cause –environment, firm, country– to an audience that cares about it. In other contexts, they could instead signal a broader concern for welfare. For example, an individual with type v, instead of internalizing ve where e is the externality, could also have concern for taxpayers's costs and thus internalize $v[e - (1 + \lambda)y]$. Stocking [2011] provides experimental evidence that agents take into account the "eviction effect" of receiving an incentive. Relatedly, Deffains and Fluet [2010] consider strict liability law, which forces harm-doers to fully compensate their victims; this is similar to $y = e$ (and implicitly assumes $\lambda = 0$). The individual experiences no moral disutility since the victim is made whole (zero net externality); as a result, there is no

[60] For the exact form of $\Delta(v^*)$ when participation or non-participation may be subject to unobservable shocks, see Bénabou and Tirole [2006a].

reputational effect either, since behavior becomes uncorrelated with concern for others.[61] With realistic transaction or enforcement costs, however, strict liability will be suboptimal, thus bringing back reputation and the interactions of laws and norms.

It could also be the principal who has different preferences. First, the case where he puts a weight less than one on agents' total welfare and derives private benefits from their contributions can be renormalized into a planner's problem (see Section 1.1). Second, a social planner could value differently the material welfare of agents with different types, or those of contributors and abstainers. This corresponds to an objective function of the form

$$W(v^*, y, \theta) = \int_{v^*}^{+\infty} h^+(v; e, c)\, g_\theta(v) dv + \int_{-\infty}^{v^*} h^-(v; e, c)\, g_\theta(v) dv - \lambda y[1 - G_\theta(v^*)] \qquad (35)$$

where v^* is the equilibrium cutoff, h^+ and h^- are any two functions of v such that $h^+ \geq h^-$, and the last term represents the deadweight loss from transfers. More generally, consider: (i) any cutoff rule $v^* = v^*(y, \hat{\theta})$ describing agent's behavior as a function of the incentive and their (point) belief about θ, with $\partial v^*/\partial y < 0$; (ii) any objective function of the form $W = \Phi(v^*; \theta) - \lambda y[1 - G_\theta(v^*)]$ such that $\partial^2\Phi/\partial v^* \partial\theta = 0$ where $\partial\Phi/\partial v^* = 0$ –satisfied, in particular, by (35). We show in the Appendix that (under appropriate regularity conditions) the key result on expressive content over θ leading to weaker incentives no matter whether $\partial v^*/\partial\hat{\theta} < 0$ (corresponding to complementarity) or $\partial v^*/\partial\hat{\theta} > 0$ (substitutability) will again hold in any separating equilibrium.

6 Conclusion

The paper's main results can be summarized by two multipliers: a social multiplier, measuring how reputational payoffs depend on the frequency of different behaviors in the population, and an informational multiplier, reflecting how perceptions of societal preferences and prevailing norms are affected by the policies of an informed principal. Optimal incentives take both into account, resulting in two departures from standard Pigou-Ramsey taxation. First, because incentives are shown to generate crowding out for rare, heroic behaviors but crowding-in for common, merely respectable ones, their optimal level depends (nonmonotonically) on the private cost of the behavior and the distribution of intrinsic motivations in society, neither of which plays a role in the standard Pigovian rule. Second, expressive concerns always lead to weaker incentives when the principal's information involves the general "goodness" of society (more generally, the strength of social norms),

[61] Fuster and Meier [2010] find that when incentives are provided (by the experimenter) a public-goods game with punishments, players sanction free-riders less, and the latter show a lesser response (increase in contribution) in subsequent rounds. Due to this form of crowding out of norm enforcement, incentives are found to be much less effective than when decentralized punishments are not feasible. The finding is again consistent with the idea that when shirkers already pay compensation there is less need for them to feel guilty, and for others to stigmatize them, than when their actions remain unpriced. Of course, agents' views on how to price an externality (or on the social welfare criterion) will often differ considerably. For instance, a polluting firm that pays a tax equal to the average (or median) agent's assessment of the externality it imposes may still be subject to boycotts by those who care more about its environmental impact.

and to stronger ones when it concerns the spillovers created by agents' behavior.

There are several directions in which our analysis could be interestingly expanded. First, we have taken the distribution of preferences as exogenous. This is a good approximation when the population is fixed, such as for a country. By contrast, a firm may choose to segregate workers with heterogeneous values into sub-units where different norms will prevail, and likewise for a school with its students. There can also be self-sorting through cooptation and exit in organizations, or through migration across neighborhoods and regions. Extending the model to deal with segregation –both equilibrium and optimal– could thus shed light on local variations in norms and institutions. Second, the coevolution of norms, law, and the social meaning of private and public actions, offers a vast and promising topic for future research.[62]

[62] On socially-minded behavior and sorting, see Besley and Ghatak [2005] and Fisher and Huddart [2008]. On the evolutionary dynamics of norms, see e.g., Guiso et al., [2008] Tabellini [2008], Greif [2009] and Acemoglu and Jackson [2011].

Appendix

Properties of the Δ function. Recall that Δ is minimized at $v = 0$ (by normalization of the v's) and that we strengthened the quasiconcavity implied by the second part of Lemma 1 to assume that it is strictly decreasing on $[v_{\min}, 0]$ and strictly increasing on $[0, v_{\max}]$. This implies that for any small $\tilde{\varepsilon} > 0$,there exists $\eta^*(\tilde{\varepsilon}) > 0$ such that

$$\Delta'(v) < -\eta^*(\tilde{\varepsilon}) \quad \text{on} \quad (v_{\min} + \tilde{\varepsilon}, -\tilde{\varepsilon}) \qquad \text{and} \qquad \Delta'(v) > \eta^*(\tilde{\varepsilon}) \quad \text{on} \quad (\tilde{\varepsilon}, v_{\max} - \tilde{\varepsilon}). \tag{A.1}$$

Note also that

$$\Delta'(v) = \frac{g(v)}{1 - G(v)}[\mathcal{M}^+(v) - v] - \frac{g(v)}{G(v)}[v - \mathcal{M}^-(v)],$$

so $|\Delta'|$ is clearly bounded on $(v_{\min}, v_{\max})$. At the boundaries, l'Hopital's rule yields $\Delta'(v_{\min}) = g(v_{\min})(\bar{v} - v_{\min}) - 1/2$ and $\Delta'(v_{\max}) = 1/2 - g(v_{\max})(v_{\max} - \bar{v})$, hence Δ' is bounded on $[v_{\min}, v_{\max}]$.

Proof of Proposition 2 (i) Let us express (12) as $F(y^{SI}(\theta), \theta) = 0$, where

$$F(y, \theta) \equiv e + v_\theta^*(y) - c - \lambda \left[y + \frac{1 + \mu \Delta'_\theta(v_\theta^*(y))}{h_\theta(v_\theta^*(y))} \right], \tag{A.2}$$

with all functions in the bracketed term evaluated at $v_\theta^*(y) = \theta + v_0^*(y + \theta)$. Since h_θ is strictly positive and continuously differentiable (C^1) everywhere, so is F, with

$$\begin{aligned} F_y(y, \theta) &= \frac{-1}{1 + \mu \Delta'_\theta} - \lambda \left[1 - \frac{\mu \Delta''_\theta h_\theta - h'_\theta (1 + \mu \Delta'_\theta)}{h_\theta^2} \left(\frac{1}{1 + \mu \Delta'_\theta} \right) \right] \\ &\equiv \frac{-1}{1 + \mu \Delta'(v^*(y, \theta) - \theta)} \left[1 - \lambda \chi(v_\theta^*(y) - \theta) \right]. \end{aligned} \tag{A.3}$$

Since h and Δ have continuous derivatives, the function $\chi(v)$ is bounded on V. Let $\lambda_1 \equiv 1/\sup_{v \in V} \{\chi(v)\}$ when this number is positive and $\lambda_1 = +\infty$ otherwise. Thus, $F(y, \theta)$ is strictly decreasing in y whenever $\lambda < \lambda_1$. Next, observe that for $y = y^{FB}(\theta)$ the non-bracketed terms in (A.2) sum to zero, so $F(y^{FB}(\theta), \theta) < 0$ for all θ. We also have $F(0, \theta) > 0$ if

$$e + v_\theta^*(0) - c > \lambda \left(\frac{1 + \mu \Delta'(v_\theta^*(0) - \theta)}{h(v_\theta^*(0)) - \theta)} \right),$$

or equivalently by (5) and the identity $v_\theta^*(0) - \theta = v_0^*(\theta)$:

$$e - \mu \Delta(v_0^*(\theta)) > \lambda \left(\frac{1 + \mu \Delta'(v_0^*(\theta))}{h(v_0^*(\theta))} \right). \tag{A.4}$$

From (14), $e - \mu \Delta(v) > 0$ for all $v \in V$, so this expression is bounded on $V = [v_{\min}, v_{\max}]$. Therefore

$$\lambda_2 \equiv \inf_{v \in V} \left[\frac{[e - \mu \Delta(v)] h(v)}{1 + \mu \Delta'(v)} \right] > 0, \tag{A.5}$$

and for $\lambda < \min\{\lambda_1, \lambda_2\}$ the function $F(\cdot, \theta)$ has a (unique) zero $y^{SI}(\theta) \in (0, y^{FB}(\theta))$. ||

(ii) We focus here on the case $\theta_1 > \theta_0$, and denote $\varepsilon_1 \equiv \theta_1 - \theta_0$; the case with $\theta_2 < \theta_0$ can be treated symmetrically. By the implicit function theorem,

$$\frac{dy^{SI}(\theta)}{d\theta} = \frac{-F_\theta(y,\theta)}{F_y(y,\theta)} = \frac{\frac{\mu\Delta'_\theta}{1+\mu\Delta'_\theta} + \lambda\left[\frac{\mu\Delta''_\theta h_\theta - h'_\theta(1+\mu\Delta'_\theta)}{h_\theta^2}\left(\frac{1}{1+\mu\Delta'_\theta}\right)\right]}{\frac{1}{1+\mu\Delta'_\theta} + \lambda\left[1 - \frac{\mu\Delta''_\theta h_\theta - h'_\theta(1+\mu\Delta'_\theta)}{h_\theta^2}\left(\frac{1}{1+\mu\Delta'_\theta}\right)\right]}, \tag{A.6}$$

evaluated at $v^*_\theta(y^{SI}(\theta))$. We next show that that $\mu\Delta'_\theta(v^*_\theta(y^{SI}(\theta)))$ is negative and bounded away from zero on (θ_1, θ_2). First, note that

$$v^*_\theta(y^{SI}(\theta)) - \theta = \theta_0 - \theta + \lambda\left[y^{SI}(\theta) + \frac{1 + \mu\Delta'_\theta(v^*_\theta(y^{SI}(\theta)))}{h_\theta(v^*_\theta(y^{SI}(\theta)))}\right]. \tag{A.7}$$

Fix ε'' with $\varepsilon'' < \theta_1 - \theta_0$ and define

$$\lambda_3 \equiv \frac{\theta_1 - \theta_0 - \varepsilon''}{\sup_{v\in V}\left[e + (1+\mu\Delta'(v))/h(v)\right]} > 0. \tag{A.8}$$

Since $y^{SI}(\theta) < y^{FB}(\theta) < e$, when $\lambda < \min\{\lambda_1, \lambda_2, \lambda_3\}$, $v^*(y^{SI}, \theta) - \theta < -\varepsilon'$ for all θ in (θ_1, θ_2). Next, since $(\theta_1, \theta_2) \subset \Theta$, we have by (8) $\theta_0 - \theta \equiv c - e - \theta > v_{\min} + \varepsilon$, so there exists $\lambda_4 \in (0, \lambda_3)$ such that for all $\lambda < \lambda_4$, $v^*(y^{SI}, \theta) - \theta > v_{\min} + \varepsilon/2$. Denoting $\varepsilon' \equiv \min\{\varepsilon'', \varepsilon/2\}$, and $\eta' \equiv \eta^*(\varepsilon')$, property (A.1) therefore implies that

$$\Delta'(v^*_\theta(y^{SI}(\theta)) - \theta) \ < -\eta' \text{ for all } \theta \in (\theta_1, \theta_2). \tag{A.9}$$

Finally, let us define

$$\lambda_5 \equiv \eta' \Big/ \sup_{v\in V}\left[\frac{\mu\Delta''h - h'(1+\mu\Delta')}{h^2}\right] > 0. \tag{A.10}$$

Thus, for $\lambda < \bar{\lambda} \equiv \min\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$, $(\partial F/\partial y)(y^{SI}(\theta), \theta) < 0$ and (A.6) implies that $y^{SI}(\theta)$ is strictly decreasing on (θ_1, θ_2). We shall denote $\theta^{FI}(y)$ its inverse function. ■

Proof of Proposition 4 We first verify that the principal discloses if and only if $v^*_\theta(y) \le v^*_{\tilde{F}}(y)$. Indeed, in that case disclosure leads to a welfare gain of

$$\begin{aligned}\int_{v^*_\theta(y)}^{v^*_{\tilde{F}}(y)} (e + v - c - \lambda y) g_\theta(v) dv \ &> \ \int_{v^*_\theta(y)}^{v^*_{\tilde{F}}(y)} (e + v^*_\theta(y) - c - \lambda y) g_\theta(v) dv \\ &= \ \int_{v^*_\theta(y)}^{v^*_{\tilde{F}}(y)} (e - \mu\Delta_\theta(v^*_\theta(y)) - (1+\lambda)) g_\theta(v) dv > 0.\end{aligned}$$

by (16), whereas when $v^*_\theta(y) > v^*_{nd}(y)$ it generates a loss of

$$\begin{aligned}\int_{v^*_\theta(y)}^{v^*_{\tilde{F}}(y)} (e+v-c-\lambda y)g_\theta(v)dv &> \int_{v^*_\theta(y)}^{v^*_{\tilde{F}}(y)} (e+v^*_\theta(y)-c-\lambda y)g_\theta(v)dv \\ &= \int_{v^*_\theta(y)}^{v^*_{\tilde{F}}(y)} (e-\mu\Delta_\theta(v^*_\theta(y))-(1+\lambda))g_\theta(v)dv > 0.\end{aligned}$$

due to the same condition.

We next analyze the fixed-point problem (20). Since $F_{\tilde{\theta}}$ is decreasing in $1-q$ and in $\tilde{\theta}$, so is $v^*_{F_{\tilde{\theta}}}(y)$, by (18): a lack of disclosure is better news (and therefore leads to more participation) the less likely it is that the principal has information, and the more selective his disclosure policy. Consequently, the right-hand side of (20) is decreasing in $\tilde{\theta}$ and $1-q$; since the left-hand-side is decreasing in $\tilde{\theta}$ there could in general be multiple intersections. We now show that there always exists one where $\tilde{\theta} \mapsto v_{\tilde{\theta}}(y)$ cuts $\tilde{\theta} \mapsto v_{F_{\tilde{\theta}}}(y)$ from above, which corresponds (since both functions are decreasing) to a stable equilibrium (i.e., a point where $d[\left(v_{\tilde{\theta}}\right)^{-1}(v_{F_{\tilde{\theta}}}(y))]/d\tilde{\theta} < 1$). Indeed, as $\tilde{\theta}$ tends to θ_2, $F_{\tilde{\theta}}$ tends to the prior F and $v^*_{\theta_2}(y) < v^*_F(y)$ since F puts positive mass on $[\theta_1, \theta_2)$; as $\tilde{\theta}$ tends to θ_1, $F_{\tilde{\theta}}$ also tends to F (see (19)) and $v^*_{\theta_1}(y) > v^*_F(y)$ since F puts positive mass on $(\theta_1, \theta_2]$. Therefore, there is at least one $\tilde{\theta}_q \in (\theta_1, \theta_2)$ where the two curves intersect and where $d[v_{\tilde{\theta}}(y) - v_{F_{\tilde{\theta}}}(y)]/d\tilde{\theta} < 0$; since this difference is decreasing in q, $\tilde{\theta}_q$ is locally decreasing in q. ■

Proof or Proposition 5 For $\lambda = 0$, $\partial W^{AI}_\theta/\partial y$ is proportional to $e - c + v^*_{\hat{\theta}(y)}(y)$. If the planner sets $y^{AI}(\theta) \equiv e - \mu\Delta_\theta(c-e)$, this policy is invertible on any interval that does not contain θ_0. Agents thus correctly infer θ, and the participation cutoff is

$$v^*_\theta(y^{AI}(\theta)) \equiv v^*_\theta(e - \Delta_\theta(c-e)) = c - e + \mu\Delta_\theta(c-e) - \mu\Delta_\theta(c-e) = c - e.$$

Therefore, for all θ, $\partial W^{AI}_\theta/\partial y = 0$ for $y = e - \mu\Delta_\theta(c-e)$. Strict quasiconcavity then implies that this is the optimal policy under asymmetric information when the planner observes that value θ.

To show uniqueness, let $y(\theta)$ be some other function that equates the left-hand side of (23) to zero for all $\theta \in [\theta_1, \theta_2]$. If the first term in brackets is zero, then $v^*(y(\theta)) = c - e = v^*(y^{FB}(\theta))$, therefore $y(\theta) = y^{FB}(\theta)$. If the second term is zero, $y(\theta)$ is a solution to the differential equation $y'(\theta) = \mu\Delta'_\theta(v^*_\theta(y(\theta)) = \mu\Delta'(v^*_\theta(y(\theta)) - \theta)$, and as any separating equilibrium it must also satisfy the boundary condition $y(\theta_1) = y^{SI}(\theta_1) = y^{FB}(\theta_1)$. Since Δ' has bounded derivatives, standard standard theorems ensure that the solutions to this initial-value problem is unique. Note, however, that $y^{FB}(\theta) = e - \mu\Delta_\theta(c-e)$ satisfies that same differential equation and coincides with $y(\theta)$ at the boundary. Therefore, the two must be equal. ■

Proof of Lemma 1 and Proposition 6 Fix $(\theta_1, \theta_2) \subset \Theta$, with $\theta_0 < \theta_1$, and again denote $\varepsilon_1 \equiv \theta_1 - \theta_0 > 0$. The case $\theta_2 < \theta_0$ can be treated symmetrically**.** By Proposition 2**,** for $\lambda < \bar{\lambda}$ there exists a decreasing function (which depends on λ) $y^{SI} : [\theta_1, \theta_2] \to [y^{SI}(\theta_2), y^{SI}(\theta_1)] \subset (0, y^{FB}(\theta))$ that solves the full-information problem, $F(y^{SI}(\theta), \theta) = 0$.

Let $y_1 \equiv y^{SI}(\theta_1)$ and consider now the *initial-value problem* defined by $\hat{\theta}(y_1) \equiv \theta_1$ and the differential equation (22), which we rewrite as

$$IVP(\lambda): \quad \hat{\theta}'(y) = \Psi(y, \hat{\theta}(y)), \quad \text{with } \hat{\theta}(y_1) \equiv \theta_1, \tag{A.11}$$

where

$$\Psi(y, \theta) = \frac{F(y, \theta)}{\mu \Delta'_\theta(v^*_\theta(y))[e - c - \lambda y + v^*_\theta(y)]}, \tag{A.12}$$

and $F(y, \theta)$ is still given by (A.2).

The proof will proceed in three steps. First, we show existence of a unique local solution $\hat{\theta}(y)$ on some left-neighborhood of y_1. We then establish key properties of this function, including monotonicity; this is the most difficult step. Finally, we use these properties to show that the function can be (uniquely) extended to a global solution, mapping some interval $[y_2, y_1]$ with $y_2 > 0$ into $[\theta_1, \theta_2]$; its inverse, $y^{AI}(\theta)$, is therefore defined on all of (θ_1, θ_2). To lighten notation, we shall abbreviate the function $v^*_{\hat{\theta}(y)}(y)$ as simply $\hat{v}(y)$.

• *Step 1: local existence and uniqueness.* The function $\Psi(y, \theta)$ and its partial derivatives are well-defined and continuous at every point where the denominator terms, $e - c - \lambda y + v^*_\theta(y)$ and $\Delta'(v^*_\theta(y) - \theta)$, are non-zero. In particular, at $(y_1, \theta_1) = (y^{SI}(\theta_1), \theta_1)$ we have $e - c - \lambda y_1 + v^*_{\theta_1}(y_1) > 0$ due to (15). Moreover, (A.9) implies that, at $\theta = \theta_1$, $\Delta'_{\theta_1}(v^*_\theta(y_1)) - \theta_1) \ < -\eta' < 0$. Therefore, $\Psi(y, \theta)$ has bounded derivatives in a neighborhood of the form $[y_1 - z, y_1] \times [\theta_1 - z', \theta_1 + z']$, implying by standard theorems that the initial-value problem $IVP(\lambda)$ has a unique solution defined on some local left-neighborhood of y_1. Let $(\tilde{y}_\lambda, y_1]$ denote the maximal (left-)interval on which such a unique solution satisfying $\theta(y) \in [\theta_1, \theta_2]$ exists, and let $\hat{\theta}_\lambda$, or for short $\hat{\theta} : (\tilde{y}_\lambda, y_1] \to (\theta_1, \tilde{\theta}_\lambda]$, denote that solution. ‖

Step 2: properties of the solution. Fix any ε' such that $h(-\varepsilon') > 0$ and

$$0 < \varepsilon' < \min\left\{\varepsilon,\ \theta_1 - v^*_{\theta_1}(y_1),\ \theta_1 - \theta_0\right\} \tag{A.13}$$

and define

$$\lambda^* \equiv \min\left\{\frac{h(-\varepsilon')\,(\theta_1 - \theta_0 - \varepsilon')}{1 + y^{FB}(\theta_1) h(-\varepsilon')}, \bar{\lambda}\right\} > 0. \tag{A.14}$$

Lemma 2 *For all $\lambda < \lambda^*$, the function $\hat{\theta}$ has the following properties on its support:*
(i) $b(y) \equiv e - c - \lambda y + \hat{v}(y)$ is strictly decreasing, and therefore bounded below by $b(y_1) > 0$.
(ii) $\hat{v}(y) - \hat{\theta}(y)$ is bounded above by $-\varepsilon'$, implying in particular $\Delta'(\hat{v}(y) - \hat{\theta}(y)) < 0$.

Proof. (i) We have

$$b'(y) = -\lambda + \frac{-1 + \mu \Delta'_{\hat{\theta}(y)}(\hat{v}(y))\ \widehat{\theta}'(y)}{1 + \mu \Delta'_{\hat{\theta}(y)}(\hat{v}(y))} = -\lambda - \frac{\lambda}{h_{\hat{\theta}(y)}(\hat{v}(y))[e - c - \lambda y + \hat{v}(y)]}.$$

Therefore, $b'(y) < 0$ wherever $b(y) > 0$. Since $b(y_1) > 0$ by (15), this implies that b is decreasing on all of $[\theta_1, \tilde{\theta}_\lambda]$, and thus bounded below by $b(y_1) > 0$.

(ii) Note first that:

$$\begin{aligned}\frac{d[\hat{v}(y) - \hat{\theta}(y)]}{dy} &= \frac{d(v_0^*(y + \hat{\theta}(y)))}{dy} = -\frac{1 + \hat{\theta}'(y)}{1 + \mu\Delta'(\hat{v}(y) - \hat{\theta}(y))} \\ &= \frac{-1}{\mu\Delta'(\hat{v}(y) - \hat{\theta}(y))}\left[1 - \frac{\lambda}{h(\hat{v}(y)) - \hat{\theta}(y))[e - c - \lambda y + \hat{v}(y)]}\right]. \quad \text{(A.15)}\end{aligned}$$

Suppose now that (ii) does not hold, and let y' be the largest $y \in [\tilde{y}_\lambda, y_1]$ such that $\hat{v}(y) - \hat{\theta}(y) = -\varepsilon'$. Then,

$$\begin{aligned}& h(v^*_{\hat{\theta}(y')}(y')) - \hat{\theta}(y'))[e - c - \lambda y' + v^*_{\hat{\theta}(y')}(y')] \\ = \;& h(-\varepsilon')(-\theta_0 - \lambda y' + \hat{\theta}(y') - \varepsilon') > h(-\varepsilon')\left(\theta_1 - \theta_0 - \varepsilon - \lambda y'\right) \\ > \;& h(-\varepsilon')\left(\theta_1 - \theta_0 - \varepsilon' - \lambda y_1\right) > h(-\varepsilon')\left(\theta_1 - \theta_0 - \varepsilon' - \lambda y^{FB}(\theta_1)\right) > \lambda. \quad \text{(A.16)}\end{aligned}$$

The bracketed term in (A.15)is therefore positive, and since $\Delta'(v^*_{\hat{\theta}(y')}(y') - \hat{\theta}(y')) = \Delta'(-\varepsilon') < 0$ this implies that the function $\hat{v}(y) - \hat{\theta}(y)$ is increasing at y'. Since at y_1 it is strictly below $-\varepsilon'$ by (A.13), there must exists some $y'' \in (y', y_1)$ where it equals $-\varepsilon$ 'again, a contradiction. ‖

Lemma 3 *For all $\lambda < \lambda^*$:*
(i) Wherever $\hat{\theta}(y)$ lies below $\theta^{FI}(y)$ (respectively, above it) on $[\tilde{y}_\lambda, y_1]$, $\hat{\theta}$ must be decreasing (respectively, increasing).
(ii) Consequently, the two curves intersect only at y_1, $\hat{\theta}$ lies everywhere below θ^{FI}, and the function $\hat{\theta}(y)$ is strictly decreasing.
(iii) Compliance is strictly lower under asymmetric information.
(iv) $\hat{v}(y) - \hat{\theta}(y) \in (v_{\min} + \varepsilon', -\varepsilon')$ therefore $\Delta'(\hat{v}(y) - \hat{\theta}(y))$ is bounded above by $-\eta^*(\varepsilon')$.

Proof : (i) We have shown that

$$e + \hat{v}(y) - c - \lambda y > 0 > \mu\Delta'_{\hat{\theta}(y)}(\hat{v}(y)). \quad \text{(A.17)}$$

Equation (22) therefore implies that $\hat{\theta}'(y) \leq 0$ if and only if

$$\frac{e + \hat{v}(y) - c - \lambda}{1 + \mu\Delta'_{\hat{\theta}(y)}(\hat{v}(y))} \geq \frac{\lambda}{h_{\hat{\theta}(y)}(\hat{v}(y))}, \quad \text{(A.18)}$$

which by (12) means that $\partial W^{FI}_{\hat{\theta}(y)}/\partial y \geq 0$ at y. By strict quasiconcavity of $W^{FI}_{\hat{\theta}(y)}(y)$, this is equivalent to $y \leq y^{SI}(\hat{\theta}(y))$, or $\theta^{FI}(y) \leq \hat{\theta}(y)$.

(ii) Where the two curves intersect, the above inequalities must all be equalities, and in particular it must be that $\hat{\theta}'(y) = 0$. Since θ^{FI} is a decreasing function, $\hat{\theta}'(y_1) = 0 > (\theta^{FI})'(y_1)$, so just to

the left of y_1, $\hat{\theta}(y)$ lies below the decreasing curve $\theta^{FI}(y)$. It cannot cut it elsewhere, since at any such intersection $\hat{\theta}$ would have to be steeper than θ^{FI}, while at the same time having a horizontal derivative, a contradiction. The last part of the claim follows from (i).

(iii) and (iv) From (9), we have $\hat{v}(y)-\hat{\theta}(y) = v^*(y+\hat{\theta}(y)) > v^*(y+\theta^{AI}(y)) > v^*(y+\theta^{FB}(y)) = c-e > v_{\min}+\varepsilon$, where the first inequality (establishing (iii)) follows from (ii) above, the second from the fact that $y^{SI}(\theta) < y^{FB}(\theta)$ for all θ, and the last one from (8) together with $\varepsilon' < \varepsilon$. In Lemma 2 we showed that $\hat{v}(y)-\hat{\theta}(y)$ is bounded above by $-\varepsilon'$, so we now have both parts of (A.1), implying the last claim in (iv). ‖

Step 3: existence and uniqueness of a global solution for y^{AI} on (θ_1,θ_2). Recall that $\hat{\theta}(y)$ is strictly decreasing on $[\tilde{y}_\lambda, y_1]$ and that $\hat{\theta}(y) \in [\theta_1,\theta_2]$ as this is part of the joint definition of $[\tilde{y}_\lambda, y_1]$ and $\hat{\theta}$. Therefore, as $y \to \tilde{y}_\lambda$ from above, $\hat{\theta}(\tilde{y}_\lambda)$ tends to a limit $\hat{\theta}(\tilde{y}_\lambda) \leq \theta_2$. Note now that Lemmas 2 and 3 imply that Ψ has bounded derivatives (hence satisfies the Lipschitz conditions) on $[\tilde{y}_\lambda, y_1]\times[\theta_1, \hat{\theta}(\tilde{y}_\lambda)]$. It therefore cannot be that $\hat{\theta}(\tilde{y}_\lambda) < \theta_2$, otherwise we can (uniquely) extend $\hat{\theta}$ to some left-neighborhood of $\tilde{y}_\lambda$ by solving the differential equation (A.12) with initial condition $(\tilde{y}_\lambda, \hat{\theta}(\tilde{y}_\lambda))$, and still have $\hat{\theta}(y)$ remain in (θ_1,θ_2), contradicting the earlier definition of the maximal interval $(\tilde{y}_\lambda, y_1]$. Therefore $\hat{\theta}(\tilde{y}_\lambda) = \theta_2$, proving that $\hat{\theta}$ is a (unique) global solution to $IPV(\lambda)$, mapping $[\tilde{y}_\lambda, y_1]$ onto $[\theta_1,\theta_2]$, with $\hat{\theta}' < 0$ and (by Lemma 3(i), $\hat{\theta}(y) < \theta^{AI}(y)$ for all $y < y_1$. Defining $y_2 \equiv \tilde{y}_\lambda$ the inverse function $y^{AI} \equiv \hat{\theta}^{-1}$ concludes the proof. ■

Proof of Proposition 7 Given the assumptions made following (27), the function $\phi(y,\theta) \equiv y + (1+\lambda)^{-1}\left[\lambda/h\left(v^*(y)-\theta\right) - e_b\right]$ is then increasing in y, θ and λ, with $\phi(0,\theta) < 0$, or $< \phi(e_b/(1+\lambda), \theta)$. Hence $y^{SI}(\theta)$ is uniquely defined and decreasing in θ and λ, with $0 < y^{SI}(\theta) < y^{FB}(\theta)/(1+\lambda)$. Turning now to the differential equation (28), it can be rewritten as

$$\hat{\theta}'(y) = \left(\frac{1+\mu\Delta'_{\hat{\theta}(y)}\left(v_a(y)\right)}{\mu\Delta'_{\hat{\theta}(y)}\left(v_a(y)\right)}\right)\left(\frac{e_b + v_b^*(y) - c_b - \lambda y - \lambda/h_{\hat{\theta}(y)}\left(v_b^*(y)\right)}{e_a - c_a + v_a^*(y)}\right)\left(\frac{g_\theta\left(v_b^*(y)\right)}{g_\theta\left(v_a^*(y)\right)}\right). \quad \text{(A.19)}$$

Consider the initial-value problem defined by (A.19) and the initial condition $\hat{\theta}(y_1) = \theta^{FI}(y_1)$, where $\theta^{FI} \equiv \left(y^{SI}\right)^{-1}$ is a decreasing function, and $y_1 \equiv y^{SI}(\theta_1)$. Since we imposed $e_a - c_a + v_a^*\left(y^{SI}(\theta)\right) > 0$, the differential equation is well-behaved, so the initial-value problem has a unique solution $\hat{\theta}(y)$ in a some left-neighborhood $[y_2, y_1]$ of the initial condition. Moreover since $\Delta'_\theta\left(v_a^*\left(y^{SI}(\theta)\right)\right) \leq 0$, $\hat{\theta}(y)$ is strictly decreasing as long as

$$e_b + v_b^*(y) - c_b - \lambda y > \frac{\lambda}{h_{\hat{\theta}(y)}\left(v_b^*(y)\right)}. \quad \text{(A.20)}$$

Where this last condition holds, we also know that $\partial W_\theta^{FI}(y)/\partial y > 0$ and, by quasiconcavity of W^{FI}, we conclude that $y < y^{SI}(\hat{\theta}(y))$, or equivalently $\theta^{FI}(y) < \hat{\theta}(y)$. Therefore $\hat{\theta}(.)$ is decreasing if and only if $\theta^{FI}(y) \leq \hat{\theta}(y)$.

Now, observe that wherever $\hat{\theta}(y)$ and $\theta^{FI}(y)$ intersect, it must be that $e_b + v_b^*(y) - c_b - \lambda y =$

$\lambda/h_\theta\left(v_b^*\left(y\right)\right)$, and consequently $\hat{\theta}' = 0$ at the intersection point. Such is the case at the initial point $y_1 \equiv \theta^{FI}(\theta_1)$, therefore $\hat{\theta}'\left(y\right) > \left(\theta^{FI}\right)'\left(y\right)$ and $\hat{\theta}\left(y\right) > \theta^{FI}\left(y\right)$ on some left-neighborhood of y_1. Suppose the two curves intersect at more than one point, and let $z < y_*$ be the largest such intersection. At that point $\hat{\theta}$ must cut θ^{FI} from above, meaning that $\hat{\theta}'(y_*) < \left(\theta^{FI}\right)'(y_*) < 0$, which contradicts the fact that $\hat{\theta}'\left(y_*\right) = 0$. Therefore the two curves intersect only at y_1, implying that $\hat{\theta}\left(y\right) > \theta^{FI}\left(y\right)$ and $\hat{\theta}' < 0$ on $(y_2, y_1]$. Equivalently, $y^{AI} \equiv \hat{\theta}^{-1}$ is decreasing on $[\theta_1, \theta_2]$, and $y^{AI}\left(\theta\right) \leq y^{SI}\left(\theta\right)$ with strict inequality except at θ_1. ■

Proof of Proposition 8. The proof follows steps very similar to those used for Proposition **6**; it is omitted to avoid repetition and economize on space, but is available upon request. ■

Proof of Proposition 9. We will show the claimed properties of the symmetric-information solution $p^{FI}(\kappa)$ for $\mu = 0$, in which case $v^*(p) = c - p$. By continuity, they extend to μ small enough. Denoting $\lambda' \equiv \lambda - \kappa\gamma$, under full information the first-order condition (equation (34) with $\beta = 0$) becomes

$$\left(W_{\lambda'}^{FI}\right)'\left(p\right) \equiv \left(e - p\right) g\left(c - p\right) + \lambda' \left[pg\left(c - p\right) - G\left(c - p\right)\right] = 0,$$

For $\lambda = 0$, the objective function $W_0^{FI}(p)$ is strictly quasiconcave, since it is clear that if $\left(W_0^{FI}\right)'\left(p\right) = 0$ then $\left(W_0^{FI}\right)''\left(p\right) < 0$. By continuity (since all functions involved are continuously differentiable), this remains true for $|\lambda|$ small enough. The optimal policy $p^{FI}(\lambda')$ is then uniquely defined by the first order condition, which can be rewritten as

$$\psi(p, \lambda') \equiv e - p + \lambda' \left[p - k\left(c - p\right)\right] = 0, \tag{A.21}$$

where $k(v) \equiv G(v)/g(v)$. The function ψ is such that

$$\frac{\partial \psi}{\partial p}(p, \lambda') \equiv -1 + \lambda' \left[1 + k'\left(p - c\right)\right] = 0, \quad \frac{\partial \psi}{\partial \lambda'}(p, \lambda') = p - k\left(c - p\right). \tag{A.22}$$

The first derivative is negative as long as $\left|\lambda'\right| \left|1 + k'\left(p - c\right)\right| < 1$, which holds for λ not too large since k' is bounded from below (as $g > 0$ on $[v_{\min}, v_{\max}]$. The second derivative is also negative provided $p < k\left(c - p\right)$; for $\lambda' > 0$ this holds at $p = p^{FI}(\lambda')$ if and only if $e > p$, meaning that $\psi(e, \lambda') < 0$, or $e < k(c - e)$; for $\lambda' < 0$ it holds if and only if $e < p$, meaning that $\psi(e, \lambda') > 0$, or again $e < k(c - e)$. Since we assumed that $eg(c-e)/G(c-e) < 1$ (which can be ensured, for instance, as long as $eg(v_{\max}) < 1$), this condition holds as well. Consequently, for $\left|\lambda'\right| \left|1 + k'\left(p - c\right)\right| < 1$, the implicit function theorem ensures that $p^{FI}(\lambda')$ is strictly increasing in λ'.

Consider now the case of asymmetric information. The differential equation (34) can be rewritten as

$$\begin{aligned} \beta\hat{\kappa}'\left(p\right) &= \left(\frac{e - c + v^*\left(p\right) + p\left(\lambda - \hat{\kappa}\left(p\right)\gamma\right)}{1 + \mu\Delta'\left(v^*\left(p\right)\right)}\right) g\left(v^*\left(p\right)\right) - \left(\lambda - \hat{\kappa}\left(p\right)\gamma\right) G\left(v^*\left(p\right)\right) \\ &\equiv \Gamma\left(p, \kappa\left(p\right)\right), \end{aligned} \tag{A.23}$$

Since $\Gamma(p,\kappa)$ has bounded derivatives, there is a unique solution to (A.23) with boundary condition $\hat{\kappa}(p_1) = \kappa^{FI}(p_1) = \kappa_1$, where $p_1 \equiv p^{FI}(\kappa_1)$. Furthermore, at any point where $\hat{\kappa}(p_1) = \kappa^{FI}(p_1)$, it must be that $\Gamma(p, \hat{\kappa}(p)) = 0$, hence $\hat{\kappa}'(p) = 0 < \left(\kappa^{FI}\right)'(p)$, where the inequality was established earlier (for $|\lambda'| = |\lambda - \gamma\kappa|$ and μ small enough). Therefore, $\hat{\kappa}'(p)$ is everywhere below $\kappa^{FI}(p)$ on its support $[p_1, p_2]$, implying in turn $\Gamma(p, \hat{\kappa}(p)) > 0$, which by (34) yields $\hat{\kappa}'(p) > 0$. ■

Proof for Section 5. We provide here the key arguments in the proof relative to expressive content over θ under the more general objective function and threshold rules introduced at the end of Section 5. Since our objective is only to show the robustness of the key insight, we simply take as given the existence and differentiability of the relevant solutions (but these could be established as they were for previous propositions).

Denoting $\hat{\theta}(y)$ the value of θ which agents will infer (in a separating equilibrium) from his choice of y, the principal maximizes over y

$$\Psi(y) = \Phi(v^*(y, \hat{\theta}(y)); \theta) - \lambda y[1 - G_\theta(v^*(y, \hat{\theta}(y)))],$$

leading to the first-order condition:

$$\left[\frac{\partial \Phi}{\partial v^*} + \lambda y g(v^*(y, \hat{\theta}(y)) - \theta)\right]\left[\frac{\partial v^*}{\partial y} + \frac{\partial v^*}{\partial \hat{\theta}}\hat{\theta}'(y)\right] = \lambda[1 - G_\theta(v^*(y, \hat{\theta}))]. \tag{A.24}$$

In the Pigovian case where $\lambda = 0$ this reduces (together with the equilibrium condition $\hat{\theta} = \theta$) to $\partial\Phi(v^*(y^{FB}(\theta), \theta); \theta)/\partial v^* = 0$, or $v^*(y^{FB}(\theta), \theta); \theta) = \arg\max_y \Phi$. When $\lambda > 0$, the (Ramsey) solution under symmetric information, $y^{SI}(\theta)$, is given by (A.24) where $\hat{\theta}'$ is set to zero; therefore $\partial\Phi(v^*(y^{SI}(\theta), \theta); \theta)/\partial v^* < 0$, implying $v^*(y^{SI}(\theta), \theta) > v^*(y^{FB}(\theta), \theta)$ and hence $y^{SI}(\theta) < y^{FB}(\theta)$ provided Φ is strictly quasi-concave in v^*, which we will assume. If W has that same property (which will be the case for λ small enough), the solution under asymmetric information, $y^{AI}(\theta)$ is then such that $y^{AI}(\theta) < y^{FB}(\theta)$ if and only if $\partial W(v^*(y^{AI}(\theta), \theta); \theta)/\partial y > 0$, if and only if $\hat{\theta}'(y^{AI})(\partial v^*/\partial y) < 0$, or equivalently $(y^{AI})'(\theta)(\partial v^*/\partial y) < 0$. For λ small enough, $(y^{AI})'$ will have the same sign as $(y^{FB})'$ (the formal proof would follow the same steps as in that of Lemma 1 and Proposition 6); totally differentiating $\partial\Phi(v^*(y^{FB}(\theta), \theta); \theta)/\partial v^* = 0$ yields

$$\frac{\partial^2 \Phi}{\partial^2 v^*}\left[\frac{\partial v^*}{\partial y} dy^{FB} + \frac{\partial v^*}{\partial \hat{\theta}} d\theta\right] + \frac{\partial^2 \Phi}{\partial v^* \partial \theta} = 0, \tag{A.25}$$

Given the assumed property of the objective function that $\partial^2\Phi/\partial v^*\partial\theta = 0$ where $\partial\Phi/\partial v^* = 0$, it follows that $(dy^{AI}/d\theta)(\partial v^*/\partial y) < 0$, hence the result. ■

REFERENCES

Acemoglu, D and M. Jackson (2011) "History, Expectations, and Leadership in the Evolution of Cooperation," NBER W. P. 17066.

Ali, N. and Bénabou, R. (2010) "Privacy and Evolving Societal Values: Image vs. Information," UCSD mimeo.

Andreoni, J. (1989) "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy,* 97(6), 1447-58.

Ariely, D. (2008) *Predictably Irrational: The Hidden Forces That Shape Our Decisions.* New York, Harper-Collins.

Ariely, D. Bracha, A. and S. Meier (2009) "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially", *American Economic Review,* 99(1), 544-555.

Ayres, I., Raseman, S. and A. Shih (2010) "Evidence From Two Large Field Experiments that Peer Comparison Feedback Can Reduce Energy Usage," NBER W.P. 15386.

Baker, G., Gibbons, R. and K. Murphy (1994) "Subjective Performance Measures in Optimal Incentive Contracts," *Quarterly Journal of Economics,* 109: 1125-56.

Bar-Gill, O. and Fershtman, C. (2004) "Law and Preferences," *Journal of Law Economics and Organization,* 20, 331-352.

Bar-Isaac, H. (2009) "Transparency, Career Concerns, and Incentives for Acquiring Expertise." NYU mimeo, January.

Bem, D. (1972) "Self-Perception Theory." in L. Berkowitz , ed., *Advances in Experimental Social Psychology*, Vol. 6, New York: Academic Press, 1-62.

Bénabou, R. (2008) "Ideology" *Journal of the European Economic Association,* (2-3), 321-352.

Bénabou, R. and J. Tirole (2003) "Intrinsic and Extrinsic Motivation," *Review of Economic Studies,* 70(3), 489-520.

—— (2004) "Willpower and Personal Rules," *Journal of Political Economy*, 112(4): 848–886.

——(2006a) "Incentives and Prosocial Behavior," *American Economic Review*, 96(5), 1652-1678.

—— (2006b) "Belief in a Just World and Redistributive Politics," *Quarterly Journal of Economics,* 121(2) , 699-746.

—— (2011) "Identity, Morals and Taboos: Beliefs as Assets," *Quarterly Journal of Economics,* 126, 805–855.

Bernheim, D. (1994) "A Theory of Conformity." *Journal of Political Economy*, 102(5), 842–877.
—— and M. Whinston (1998) "Incomplete Contracts and Strategic Ambiguity," *American Economic Review,* 88: 902-932.

Besley, T. and Ghatak, M. (2005) "Competition and Incentives with Motivated Agents," *American Economic Review*, 95(3), 616-636.

Bicchieri, C. and E. Xiao (2010) "Do the Right Thing: But Only if Others Do So," University of Pennsylvania mimeo, Politics and Economics Program.

Bohnet, I., Frey, B., Huck, S. (2001) "More Order with Less Law: on Contract Enforcement, Trust and Crowding," *American Political Sciences Review,* 9, 131-144.

Bowles, S. (2008) "Policies Designed for Self-Interested Citizens May Undermine "The Moral Sentiments": Evidence from Economic Experiments," *Science,* 320, 1605-1609.

—— and S. Reyes (2009) "Economic Incentives and Social Preferences: A Preference-Based Lucas Critique of Public Policy," Santa Fe Institute mimeo, May.

Brekke, K., Snorre, K. and K. Nyborg (2003) "An Economic Model of Moral Motivation." *Journal of Public Economics,* 87 (9-10), 1967-1983**.**

Bremzeny, A. Khokhlovaz, E., Suvorov, A. and J. van de Ven (1991) "Bad News: An Experimental Study On The Informational Effects Of Rewards," mimeo, New Economic School.

Brennan, G. and M. Brooks (2007) "Esteem, Norms of Participation and Public Goods Supply," *Public Economics and Public Choice,* 63-80.

Chicone, C. (1999) *Ordinary Differential Equations, with Applications.* Secaucus, NJ, USA: Springer-Verlag New York, Incorporated.

Cialdini, R. (1984) *Influence.* New York: William Morrow and Company,

—— , Demaine, L., Sagarin, B., Barrett, D., Rhoads, K. and P. Winter (2006) "Managing Social Norms for Persuasive Impact," *Social Influence,* 1(1), 3-15.

Cooter, R.(1998) "Expressive Law and Economics," *Journal of Legal Studies,* 27(2), 585-608.

Corneo, G. (1997) "The Theory of the Open Shop Trade Union Reconsidered," *Labour Economics,* 4(1), 71-84.

Croson, R. and M. Marks (1998) "Identifiability of Individual Contributions in a Threshold Public Goods Experiment," *Journal of Mathematical Psychology* 42, 167 190.

Daughety, A. and J. Reinganum (2009) "Public Goods, Social Pressure, and the Choice Between Privacy and Publicity," *American Economic Journal: Microeconomics,* 2(2), 191-221.

Deffains, B. and C. Fluet (2010) "Legal Liability when Individuals Have Moral Concerns," mimeo, University Paris 2 and Université du Québec à Montréal.

Della Vigna, S., List, J. and U. Malmendier (2011) "Testing for Altruism and Social Pressure in Charitable Giving," *Quarterly Journal of Economics,* forthcoming.

Ellickson, R. (1998) "Law and Economics Discovers Social Norms," *Journal of Legal Studies,* 27(2), 537-552.

Ellingsen, T. and M. Johannesson (2008): Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review,* 98(3), 990-1008.

Falk, A. and M. Kosfeld, (2006), "The Hidden Costs of Control," *American Economic Review*, 96(5), 1611-1630.

Fehr, E. and Falk, A. (2002) "Psychological Foundations Of Incentives," *European Economic Review,* 46, 687-724.

—— and S. Gächter (2002) "Do Incentive Contracts Undermine Voluntary Cooperation?" Institute for Empirical Research in Economics, Zurich University, Working Paper No. 34.

—— and Rockenbach (2003) "Detrimental Effects of Sanctions on Human Altruism," *Nature,* 422, 137-140.

Fischer, P. and S. Huddart (2008) "Optimal Contracting with Endogenous Social Norms", *American Economic Review,* 98, 1459-1475.

Frey, Bruno S. (1997) *Not Just for the Money: An Economic Theory of Personal Motivation.* Cheltenham: Edward Elgar.

Fryer, R. (2010) "Financial Incentives and Student Achievement: Evidence from Randomized Trials," *Quarterly Journal of Economics,* forthcoming.

Funk, P. (2007) "Is There An Expressive Function of Law? An Empirical Analysis of Voting Laws with Symbolic Fines", *American Law and Economics Review,* 9(1), 135-159.

Fuster, A. and S. Meier (2010 "Another Hidden Cost of Incentives: The Detrimental Effect on Norm Enforcement," *Management Science,* 56(1), 57-70.

—— (2010) "Social Incentives and Voter Turnout: Evidence from the Swiss Mail Ballot System", *Journal of the European Economic Association,* 8(5), 1077–1103.

Galbiati, R., Schlag K. and J. van der Weele (2010), "Sanctions that Signal: An Experiment," University of Sienna Working Paper 24/2009.

—— and P. Vertova (2008) "Obligations and Cooperative Behaviour in Public Good Games," *Games and Economic Behavior*, 64(1), 146-170.

Gibbons, R. (1997) "Incentives and Careers in Organizations," in: David Kreps and Ken Wallis, eds., *Advances in Economic Theory and Econometrics*, vol. 2. Cambridge University Press.

Gneezy, U., and A. Rustichini (2000a) "Pay Enough or Don't Pay At All," *Quarterly Journal of Economics,* 791-810.

—— (2000b) "A Fine is a Price," *Journal of Legal Studie*s, 29(1), 1-18.

Greif, A. (2009) "Morality and Institutions: Moral Choices Under Moral Network Externalities," Stanford University mimeo, November.

Guiso, L., Sapienza, P. and L. Zingales (2008) "Social Capital as Good Culture," *Journal of the European Economic Association,* 6(2-3):295–320.

Herold, F. (2010) "Contractual Incompleteness as a Signal of Trust," *Games and Economic Behavior,* 68(1), 180-191.

Jewitt, I. (2004) "Notes on the Shape of Distributions," Mimeo, Oxford University.

Kahan, D. (1996) "What Do Alternative Sanctions Mean?" *University of Chicago Law Review,* 63, 591-653.

—— (1997) "Between Economics and Sociology: The New Path of Deterrence," *Michigan Law Review,* 95(8), 2477-2497.

Kaplow, L. and S. Shavell (2007) "Moral Rules, the Moral Sentiments and Behavior: Toward a Theory of a Moral System that Optimally Channels Behavior," *Journal of Political Economy,* 115, 494-514.

Karlan, D. and J. List (2007) "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," *American Economic Review,* 97(5), 1774-1793.

Keizer. K. Lindenberg, S. and L. Steg. (2008) "The Spreading of Disorder", *Science,* 322 (5908), 1681-1685.

Knez, M. and D. Simester (1981) "Firm-Wide Incentives and Mutual Monitoring at Continental Airlines," *Journal of Labor Economics,* 19(4), 743-772.

Kessler, J. (2011) "Signals of Support and Public Good Provision," Harvard University mimeo.

Krupka, E. and Weber, R. (2009) "The Focusing and Informational Effects of Norms on Pro-Social Behavior," *Journal of Economic Psychology,* 30, 307-320.

Lefebvre, M., Pestieau, P., Riedl, A., and M.C. Villeval (2011) IZA Discussion Paper 5609, "Tax Evasion, Welfare Fraud, and "The Broken Windows" Effect: An Experiment in Belgium, France and the Netherlands,"March.

Lessig, L. (1998) "The New Chicago School," *Journal of Legal Studies,* 27(2), 661-691.

Licht, A.N., (2008), "Social Norms and the Law: Why Peoples Obey the Law?", *Review of Law and Economics*, 4(3), 715-750.

McAdams, R. (2000) "An Attitudinal Theory of Expressive Law," *Oregon Law Review,* Summer 79(2), 339-390.

McAdams, R. and Eric Rasmusen (2007) "Norms and the Law," in *Handbook of Law and Economics,* M. Polinsky and S. Shavell eds. Elsevier B.V., Chapter 20, 1573-1618.

Miller, D. and D. Prentice (1994) "Collective Errors and Errors About the Collective," *Personality and Social Psychology Bulletin,* 20, 541-550.

Miller, D. and D. Prentice (1994) "The Self and the Collective," *Personality and Social Psychology Bulletin,* 20, 451-453.

Panagopoulos C. (2009) "Turning Out, Cashing In: Extrinsic Rewards, Intrinsic Motivation and Voting," Fordham University mimeo, March.

Pesendorfer, W. (1995) "Design Innovation and Fashion Cycles," *American Economic Revie*w, 85(4), 771-792.

Posner, E. (1998) "Symbols, Signals, and Social Norms in Politics and the Law," *The Journal of Legal Studies,* 27(2) -798.

—— (2000a) *Law and Social Norms.* Cambridge, MA: Harvard University Press.

Prat, A. (2005) "The Wrong Kind of Transparency," *American Economic Review,* 95(3), 862-877.

Prendergast, C. (1999) "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37(1), 7-63.

Prentice, D. and Miller, D. (1993) "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm," *Journal of Personality and Social Psychology,* 64, 243-256.

Rasmusen, E. (1996) "Stigma and Self-Fulfilling Expectations of Criminality," *Journal of Law and Economics,* 39, 519-544.

Gerber, A., Green, D. and Larimer, C. (2008) "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment," *American Political Science Review,* 102, 33-48.

Reiss, P. and White, W. (2008) "What Changes Energy Consumption? Prices and Public Pressures," *RAND Journal of Economics,* 39(3), 636–663.

Richtel, M. (2008) "What's Obscene? Google Could Have an Answer", *New York Times,* June 24.

Rosenthal, E. (2008) "Motivated by a Tax, Irish Spurn Plastic Bags," *New York Times,* February 2.

Rotemberg, J. (2008) "Minimally Acceptable Altruism and the Ultimatum Game," *Journal of Economic Behavior and Organization,* 66(3-4), 457-476.

Schroeder, C. and Prentice, D. (1998) "Exposing Pluralistic Ignorance to Reduce Alcohol Use Among College Students", *Journal of Applied Social Psychology,* 28(23), 2150-2180.

Schultz, W., Nolan, J., Cialdini, R. Goldstein, N. and V. Griskevicius (2007) "The Constructive, Destructive, and Reconstructive Power of Social Norms", *Psychological Science,* 18(5), 429-433.

Shavell, S. (2002) "Law versus Morality as Regulators of Conduct," *American Law and Economics Review,* 4(2), 227-257.

Sliwka, D. (2008) "Trust as a Signal of a Social Norm and the Hidden Costs of Incentives Schemes," *American Economic Review*, 97 (3), 999-1012.

Smith, A. (1759) *The Theory of Moral Sentiments.* Reedited (1997), Washington, D.C.: Regnery Publishing, Conservative Leadership Series.

Stocking, A. (2011) "The Incentives Underlying Prosocial Behavior From Six Field Experiments," C.B.O. mimeo, May.

Sunstein, C. (1996) "On the Expressive Function of Law," *University of Pennsylvania Law Review,* 144(5), 2021-2053.

Tyler, T. (1990). *Why People Obey the Law.* New Haven: Yale.

Tyran, J. and L. Feld (2006) "Achieving Compliance When Legal Sanctions are Non-Deterrent", *Scandinavian Journal of Economics,* 108(1), 135-156.

Xiao, E. (2010) "Profit-Seeking Punishment Corrupts Norm Obedience," Carnegie Mellon University mimeo, January.

Weele, van der, J. (2009) "The Signalling Power of Sanctions in Social Dilemmas," *Journal of Law, Economics and Organization,* 10.1093/jleo/ewp039.

Wenzel, M. (2005) "Misperceptions of Social Norms about Tax Compliance: From Theory to Intervention," *Journal of Economic Psychology,* 26, 862–883.

Weibull, J. and E. Villa (2005) "Crime, Punishment and Social Norms," Working Paper Series in Economics and Finance 610, Stockholm School of Economics.