Laws and Norms

Roland Bénabou, Jean Tirole

Forthcoming, *Journal of Political Economy*

Laws and Norms*

Roland Bénabou† and Jean Tirole‡

Tuesday 21st January, 2025

Abstract

We analyze how private decisions and optimal public policies are shaped by personal and societal preferences, material incentives, and social norms. We show how honor and stigma interact with incentives and derive optimal taxation. We then analyze the expressive role of law as embodying society's values and identify when it calls for a weakening or a strengthening of incentives. The law should be softened when it signals agents' general willingness to contribute to the public good and toughened when it signals social externalities. We also shed light on norms-based interventions, societies' resistance to economists' messages, and the avoidance of cruel and unusual punishments.

Keywords: motivation, incentives, social norms, expressive law.

JEL Classification: D64, D82, H41, K1, K42, Z13.

*We are grateful to Daron Acemoglu, Tim Besley, Betsy Paluck, Torsten Persson, Aleh Tsyvinski, Glen Weyl, Yao Zeng and participants at many seminars, named 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 acknowledges support from the European Research Council (Grant No. FP7/2007-2013 - 249429, Grant No. 669217 - ERC MARKLIM and Grant No. 101098319 - HWS) and from TSE-P's "Chaire Association Finance Durable et Investissement Responsable (AFG)."

†Princeton University, NBER, CEPR, and IZA. rbenabou@princeton.edu

‡Toulouse School of Economics and IAST. jean.tirole@tse-fr.eu

I Introduction

To foster desired behaviors, economists generally emphasize (with a number of caveats) material incentives provided through contracts, markets or policy. While these often work very effectively, there also many 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 choose what seem like inefficient forms of incentives (e.g., prison rather than fines or reparations) or renounce others that might be cheap or effective (paying for organ donations, corporal punishments, public shaming).

Rather than incentives, psychologists emphasize persuasion and social influence, in particular through informational manipulations aimed at changing the "social meaning" of actions and shifting the norms that prevail in a population. A growing literature in experimental economics also explores such norms-based interventions, but theoretical work remains relatively scarce.

Legal scholars 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. Exactly how laws do or should convey societal values remains elusive, however. 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 (such as shaming or the death penalty).

These apparently disjoint approaches are in fact highly complementary and can be brought together to shed new light on the determinants of compliance and the effects of incentives. To this effect, we develop a unifying framework to analyze how private decisions are shaped by personal and societal preferences ("values"), material or other explicit incentives ("laws") and social sanctions or rewards ("norms"), and how optimal policy should be set in such environments. The core model, presented in Section II, involves a continuum of agents who interact socially and a principal who sets incentives for them. The agents differ in their prosocial orientation and their behavior is shaped by a mix of intrinsic, extrinsic and reputational motivations (or other social payoffs). The principal optimally takes into account how her policies will interact with the social equilibrium, both through endogenous complementarities or substituabilities in agents' actions (social multiplier), and by conveying private information she may have about the environment in which they operate (informational multiplier), such as the distribution of preferences in society or the magnitude of externalities.

Focusing first on the case of symmetric information about the environment, we show in Section III how honor, stigma and social norms endogenously arise from individuals' behaviors and inferences, how they generate a social multiplier, and when they are strengthened or undermined by the presence of material incentives. Moving from positive to normative analysis, 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

[1] Examples of such puzzles include e.g., Gneezy and Rustichini (2000), Fehr and Gächter (2001), Knez and Simester (2001), Fehr and Rockenbach (2003), Karlan and List (2007), Ariely, Bracha and Meier (2009), Funk (2010), Fryer (2010), Ashraf and Bandiera (2014), and Alfitian, Sliwka and Vogelsang (2024). See, e.g., Bowles (2008) and Bowles and Reyes (2009) for surveys, and Gibbons (1997) and Prendergast (1999) for the more "classical" literature on incentives in organizations.

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 should be lowest for behaviors with very high or very low participation rates (as these respectively induce maximal stigma and maximal honor), and highest for behaviors in the "grey zone" where compliance and noncompliance are both common behaviors (and social pressure is thus at its weakest). More generally, we derive a rich set of comparative-statics results on how key parameters (distribution of preferences, intensity of social monitoring, cost and externality of individual actions) affect private behaviors and optimal policy, characterizing for each one: (i) when it encourages or discourages agents to engage in pro-social behavior; (ii) when it should induce the principal to raise or reduce extrinsic incentives. Besides providing testable predictions, this fourfold typology of results will also prove key to the analysis with asymmetric information.

In Section IV we extend the normative analysis to the expressive role of law or incentives –we will use the two terms interchangeably. Policymakers will often have information that is not available to economic agents about societal values or compliance, (e.g., the tax evasion rate), or about the consequences of anti-social behavior (e.g., the social cost of pollution). Whether intended to foster the common good or more narrow objectives, the chosen laws and other policies will then reflect such knowledge, which in turn will affect intrinsic motivation and social norms. Thus, imposing a heavy sentence for some offense or a zero price on certain transactions such as organ donations means both setting material incentives and sending a message about society's values, and hence about the norms according to which different behaviors will 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.

We first demonstrate that law should be expressive only when incentives generate a deadweight loss (because they are non-monetary or because funds are costly). When incentives are costless, the policymaker can achieve her preferred level of compliance by setting them at the proper level, and therefore has no reason to distort the message sent by the policy. By contrast, costly incentives must be employed parsimoniously, so the policymaker will attempt to optimally use expressive law to send a compliance-enhancing message and harness agents' other sources of motivation, both intrinsic and reputational.

To determine when these expressive concerns should result in softer or tougher law, we examine the existence and properties of a separating equilibrium of the game between the principal and the (mutually interacting) agents. The answer, provided by optimal-tax formulae that incorporate an informational multiplier, turns out to hinge on what specific variable the law signals, such as agents' general willingness to contribute to the public good, or the value to society of these contributions. When better informed about prevailing standards of behavior or preferences, the principal optimally tries to signal that the social norm is strong by lowering extrinsic incentives, at some cost in compliance (soft law). In contrast, when the asymmetric information concerns the magnitude of the externalities that agents impose on each other, she 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 (tough law).

More broadly, we characterize the circumstances under which equilibrium law can convey

information and thereby be expressive, and those under which it cannot. For instance, it cannot signal the intensity of social monitoring and sanctions: when this is the parameter subject to asymmetric information, the equilibrium involves full pooling by the principal.

As an alternative to incentives, we allow in Section V the policy maker to engage in direct communication similar to the norms-based interventions advocated by social psychologists: she can disclose, or withhold, information that alters agents' perception of the social norm (dispelling pluralistic ignorance) or the consequences of their action (externality-awareness raising). We make clear how such messages operate, but also how their effectiveness is limited by the credibility problem of a principal who communicates good news about prosocial behavior or community values and withholds bad ones, while doing the reverse for negative externalities.

In Section VI we extend the model in several important directions. Investigating spillovers between domains of behavior, we first shed light on why societies are often resistant to economists' advocacy of incentives, which are perceived as bringing about a nefarious "commodification" of human activity. We consider two activities driven by the same prosocial proclivity of agents. For observability or enforcement reasons, one is subject to social sanctions and rewards, but not regulated by incentives set by the government; a contrario, the other is controlled by standard extrinsic incentives. Setting or arguing for strong incentives in this latter activity communicates a negative message about general prosociality, which erodes the social norm on the other one. The consequence of this expressive spillover is a lower use, in optimal policy-making, of economists' recommendations about the importance of incentives. In another extension, we analyze why societies forego "cruel and unusual" punishments, irrespective of effectiveness considerations, in order to express their being "civilized". Finally, in the Online Appendix we extend the model to social interactions and norms that operate through channels other than reputation, such as reciprocity or a pure taste for conformity, or for exclusive status.

All proofs are gathered in the Online Appendix

Related literature. The need for an integrated analysis of law and social norms is stressed by Ellickson (1998), Lessig (1998) and McAdams and Rasmusen (2007).

The interaction of incentives with other forms of motivation under symmetric information about the social environment is studied by, among others, Frey (1997), Brekke and Nyborg (2003), Besley and Ghatak (2005) and Bénabou and Tirole (2006a).[2] We provide here a comprehensive framework that generates both new testable predictions and optimal-tax formulae. Besley, Jensen and Persson (2023), Jia and Persson (2021) and Chen (2016) build on it to study empirically the interaction of norms and incentives in the context of, respectively, tax evasion, ethnic-identity choice, and military desertions. Lane, Nozenso and Sonderegger (2023) extend it to document experimentally the discontinuity in stigma that occurs when someone breaks a law defined by a threshold (e.g., maximum driving speed, minimum age of consent).

The expressive role of law is emphasized by Sunstein (1996), Kahan (1997), Cooter (1998), Posner (1998, 2000a,b) and McAdams (1999). Our signaling approach is most closely related

[2]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.

to the one informally advocated by the last two authors.[3] The informed-principal problem that formalizes expressive law bears a relationship to those in Bénabou and Tirole (2003), Ellingsen and Johannesson (2008) and Herold (2010), but with important differences. In particular, agents must now try to infer the prevailing social standard, which 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 (2012), but the nature of normative influences is 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. In our model, conformity or distinction effects arise endogenously, and we analyze the potential for expressive law in both cases, as well as in settings where the asymmetric information bears on the shape of the preference distribution, the magnitude of externalities, or the intensity of social monitoring. We provide general results on when expressive concerns will lead to weaker or stronger incentives than under symmetric information, deriving optimal tax formulae here as well. We also identify cases in which no separating equilibrium exists, preventing the law from conveying information. Because our model is not covered by Mailath's (1987) classical results on signaling games, as part of our analysis we derive a more general incentive-compatibility condition.

A number of papers provide evidence of the signaling effect of incentives. 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 a key element, as more votes favoring mild sanctions lead agents to expect higher compliance by others, and these expectations largely explain contributions levels. Galbiati et al. (2021) show that the UK government's introduction of lockdown measures during the COVID-19 health crisis substantially changed the public perception of the norms regarding social distancing: the fraction of survey respondents believing that most other people approved of such measures rose substantially, and this shift, rather than the weakly enforced policies, was associated with significantly reduced mobility. Turning to the effects of more high-powered incentives, in Galbiati, Schlag and van der Weele (2013) a pair of players engaged in a minimum-effort game may be subject to substantial sanctions for shirking. 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 players' behavior in a previous round, by contrast, subjects who had provided high effort become pessimistic about their partner's contribution and accordingly reduce their own, making the sanctions counterproductive. Bremzeny et al. (2015) and Danilov and Sliwka (2017) also document the bad-news effect of choosing strong incentives in settings where the principal has private information about, respectively, the difficulty of a task and the previous effort norm among a set of agents.

The analysis of direct disclosure, finally, connects the paper to the literature on norms-based interventions and pluralistic ignorance (e.g., Cialdini 1984, Miller and McFarland 1987,

[3] An alternative route for laws to affect social norms is an evolutionary process of preference adaptation; e.g., Huck (1997), Bohnet, Frey and Huck (2001), Bar-Gill and Fershtman (2004), Tabellini (2008), Guiso, Sapienza and Zingales (2008), Greif (2009) and Acemoglu and Jackson (2015).

Prentice and Miller 1993). Campaigns and experiments targeting the *descriptive norm* (the norm of "is") consist in informing agents of the average (or distribution of) behavior among comparable peers, bringing into play social comparisons and self-image concerns. Schultz et al. (2007), Ayres, Raseman and Shih (2010) and Allcott (2011) demonstrate these effects for electricity conservation, and Lefebvre et al. (2011) for tax evasion. Bursztyn, Egorov and Fiorin (2020) show that raising subjects' perceptions about the fraction of Trump voters in their local area during the 2016 presidental election made them more likely to donate to an anti-immigrant organization. Interventions targeting the *prescriptive* or *injunctive* norm (the norm of "ought") consist in communicating to agents what most of their peers (say they) approve of. The idea is to dispel *pluralistic ignorance,* which occurs when people underestimate the extent to which observed behavior is driven by adherence to a commonly misperceived norm rather than by true values. Prentice and Miller (1993) found that students overestimate the extent to which their peers approve of drinking, and that this perceived tolerance is a strong predictor of use. Prentice and Schroeder (1998) used anonymously elicited students' attitudes to dispel the stereotype, resulting in lower reported levels of consumption. Bursztyn, González and Yanagizawa-Drott (2020) show that Saudi men substantially underestimate the percentage of other men in their social network who approve of a wife working outside the home, and that correcting this misperception leads more of them to allow their own wife to do so.

II The image-concern model

In our core framework, the norms shaping agents' behavior operate through their social image, while a principal sets incentives to correct the various externalities created by their actions.

A Basic framework

We index all relevant aspects of the economic and social environment by a parameter θ, with support Θ. Letting θ affect any key component of agents' or the principal's payoffs will allow us to derive unified results on how equilibrium behavior and optimal policies vary with each of them. In Sections II-III, θ is common knowledge; in Sections IV-VI it will be private information of the principal.

A continuum of agents with mass 1 each choose some discrete action $a \in \{0, 1\}$, where $a = 1$ entails a personal cost (time, effort) $c_\theta > 0$ and creates an externality $\epsilon_\theta > 0$ onto others, while also earning the individual an incentive of y, provided by some principal. In a public-goods context, $a = 1$ 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 $a = 0$). In a firm or organization, $a = 1$ represents working rather than shirking, abstaining from opportunism, helping co-workers, etc., and y a wage rate, performance-contingent bonus, or prospect of a promotion.

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

$$U = (ve_\theta - c_\theta + y)a + \epsilon_\theta \bar{a}_\theta + \mu_\theta \left(E_\theta[\tilde{v} \mid a, y] - \bar{v}_\theta \right), \tag{1}$$

The term ve_θ is the agent's *intrinsic motivation*, in which v measures the general intensity of his social preferences and $e_\theta \equiv \gamma\epsilon_\theta + 1 - \gamma$, with $\gamma \in [0,1]$, reflects the extent to which these are of a "consequentialist" (caring about externalities from one's own action) or a "warm glow" nature.[4] In a public-goods context, ve_θ represents the agent's degree of altruism or prosocial orientation, whether general or domain-specific (e.g., concern for the environment). In a firm or organization it corresponds to work ethic, liking and motivation for the task (sales, research) or mission, concern for colleagues, etc. Since the true externality is ϵ_θ, each agent derives a benefit $\epsilon_\theta \bar{a}_\theta$ from the aggregate supply $\bar{a}_\theta$.

To analyze most transparently the interplay of individual and aggregate uncertainty we focus on a single source of heterogeneity, namely intrinsic motivation. Let $F_\theta(v)$ denote the distribution of these preferences, which are private information, with finite support $V_\theta \equiv [v_\theta^{\min}, v_\theta^{\max}]$, continuously differentiable density $f_\theta(v) > 0$, a strictly increasing hazard rate $h_\theta(v) \equiv f_\theta(v)/[1 - F_\theta(v)]$ and mean $\bar{v}_\theta$. By contrast, all agents share the same marginal valuation, normalized to 1, for money or other (net) extrinsic incentives $y - c_\theta$; they also care equally about social (or self) esteem, to which we now turn.[5]

The last term in (1) captures image concerns. The observation of a leads the agent's audience to update their beliefs about his type, resulting in payoffs that reflect the posterior mean $E_\theta[\tilde{v} \mid a]$ with (common) intensity μ_θ. This value of image can be purely hedonic (enjoying social esteem *per se*), or instrumental. 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, being a team player, 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 (dis) utility experienced from *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 for external visibility.[6]

Given y, an agent chooses $a(v,y) = 1$ if $ve_\theta \geq c_\theta - y - \mu_\theta \left(E_\theta\left[\tilde{v} \mid a = 1\right] - E_\theta\left[\tilde{v} \mid a = 0\right]\right)$, implying a cutoff rule. From the preference distribution $F_\theta(v)$, we therefore define two important conditional moments:

$$E_\theta^+(v) \equiv E_\theta\left[\tilde{v} \mid \tilde{v} \geq v\right], \quad E_\theta^-(v) = E_\theta\left[\tilde{v} \mid \tilde{v} < v\right], \quad \text{for all } v \in V. \tag{2}$$

Thus $E_\theta^+(v^*)$ governs the "honor" conferred by participation, and $E_\theta^-(v^*)$ the "stigma" from abstention, when types above v^* contribute and those below do not. In the self-image interpretation of the model, they correspond respectively to feelings of pride and shame. The net reputational incentive to contribute is

$$\Delta_\theta(v^*) \equiv \mu_\theta[E_\theta^+(v^*) - E_\theta^-(v^*)]. \tag{3}$$

[4] Consequentialism is taken here in the sense of a motivation that reflects the social value of the activity in question, e.g. is higher for saving lives in an epidemic than for recycling. In a large population each individual has a negligible impact on $\bar{a}_\theta$, so this desire to nonetheless "do one's part" could also be thought of as Kantian, namely reflecting what the agent could "will" that everyone would do (e.g., Brekke, Snorre and Nyborg 2003, Alger and Weibull 2013). On intrinsic motivation, see also Besley and Ghatak (2005), Prendergast (2007) and Bénabou and Tirole (2016). The term $v(1-\gamma)$, conversely, represents pure warm glow (Andreoni 1989)

[5] Bénabou and Tirole (2006a) allow for heterogenous marginal utilities of money and reputational concerns. We abstract here from the "overjustification" and full-crowding-out effects that can arise with multidimensional types, focusing instead on new questions, such as the setting of optimal incentives and the expressive role of law.

[6] See Smith (1759), Bem (1972), Bodner and Prelec (2003), Bénabou and Tirole (2004, 2011a).

We shall focus for simplicity on the case where the equilibrium cutoff $v_\theta^*(y)$ —sometimes abbreviated as v_θ^* —is interior, and thus given by the fixed-point equation[7]

$$v_\theta^*(y)e_\theta - c_\theta + y + \Delta_\theta(v_\theta^*) = 0. \tag{4}$$

Note that reputation is here a positional good: $E_\theta[E_\theta(\tilde{v}\,|a,y)] = \bar{v}_\theta$.[8] Agents' average utility is thus

$$\bar{U}_\theta = \int_{v_\theta^*(y)}^{+\infty}(ve_\theta - c_\theta + y)dF_\theta(v) + \epsilon_\theta\bar{a} = \int_{v_\theta^*(y)}^{+\infty}(ve_\theta + \epsilon_\theta - c_\theta + y)dF_\theta(v). \tag{5}$$

B The calculus of esteem and 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_θ^* decreases (see Figure 1a), honor declines but stigma worsens, since both E_θ^+ and E_θ^- are increasing functions. Depending on which effect dominates, the net social or moral pressure Δ_θ can increase or decrease. In the first case, $\Delta'_\theta(v_\theta^*) < 0$, decisions are (locally) strategic complements, which corresponds to the usual definition of a *norm.* In the latter, $\Delta'_\theta(v_\theta^*) > 0$, they are (locally) strategic substitutes, giving rise to an *anti-norm* effect.

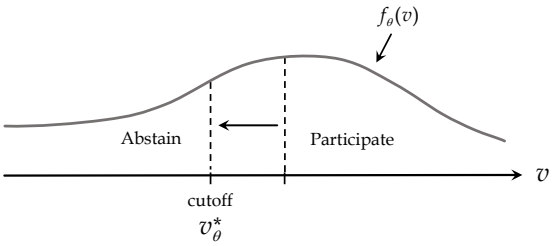

Figure 1a (preference distribution)

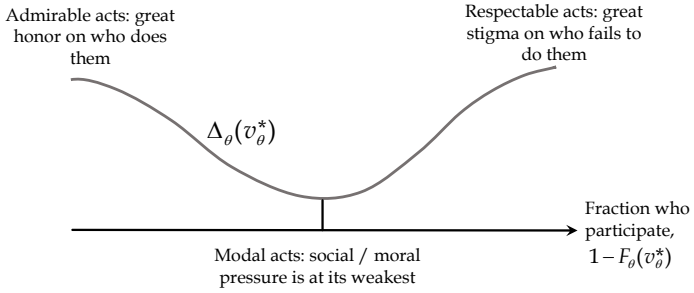

Figure 1b (reputational returns)

If strategic complementarity is strong enough and μ_θ high enough, there can be multiple equilibria –that is, self-sustaining norms. From here on, we ensure uniqueness by imposing $e_\theta + \Delta'_\theta(v) > 0$ for all $v \in V_\theta$, which holds for μ_θ not too large.[9] The slope of aggregate supply $\bar{a}_\theta(y) = 1 - F_\theta(v_\theta^*(y))$ is then $f_\theta(v_\theta^*(y))$ times the *social multiplier,*

$$s_\theta(y) \equiv -\frac{\partial v_\theta^*}{\partial y} = \frac{1}{e_\theta + \Delta'_\theta(v_\theta^*(y))}. \tag{6}$$

[7] An interior equilibrium will be ensured by assuming (or, later on, ensuring that the optimal y satisfies) $v_\theta^{\min}e_\theta + \mu_\theta(\bar{v}_\theta - v_\theta^{\min}) < c_\theta - y < v_\theta^{\max}e_\theta + \mu_\theta(v_\theta^{\max} - \bar{v}_\theta)$, together with the condition stated below for monotonicity of $ve_\theta + \Delta_\theta(v)$.

[8] Reputational value functions derived from an explicit second-stage game may not be linear (e.g, Rotemberg 2008), or involve type-dependent weights, in which cases signaling can be a positive or negative-sum game. The linear case serves as a natural and important benchmark. For a field experiment in which image payoffs are estimated to be concave, making reputation seeking a negative-sum game, see Butera et al. (2022).

[9] The fact that $|\Delta'_\theta|$ is bounded is shown in the Online Appendix. Bénabou and Tirole (2006a) provide sufficient conditions and explicit examples for the case of multiplicity, $e_\theta + \Delta'_\theta < 0$. 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).

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 ($a = 0$). This is indeed true when the distribution of agents' preferences is single-peaked, but otherwise need not be:

Lemma 1 (Jewitt; Harbaugh and Rasmusen; Adriani and Sonderegger)

(i) If f_θ is everywhere decreasing (increasing), then Δ_θ is everywhere increasing (decreasing).

(ii) If f_θ is unimodal (strictly quasi-concave), then Δ_θ is strictly quasi-convex. Its minimum is interior, provided that $f_\theta(v_\theta^{\min})$ and $f_\theta(v_\theta^{\max})$ are sufficiently small.

(iii) If f_θ is U-shaped, then Δ_θ is strictly quasi-concave. Its maximum is interior, provided that $f_\theta(v_\theta^{\min})$ and $f_\theta(v_\theta^{\max})$ are sufficiently large.

(iv) In general, the extrema of f_θ and Δ_θ do not *coincide. If, in addition to (ii) or (iii), f_θ is symmetric around its extremum, then so is Δ_θ and the two extrema coincide, implying that f_θ and Δ_θ are countermonotonic: $f'_\theta(v^*)\Delta'_\theta(v^*) < 0$ for all v^*.*[10]

We shall focus on the unimodal case (ii), which is empirically the most natural and allows for both strategic substitutability and complementarity, with (i) being a subcase; see Figure 1b. For concreteness, we shall refer to the socially desirable behavior $a = 1$ as being (in equilibrium):

- *"Respectable" or "normal"*, if v_θ^* is in the lower tail where $\Delta'_\theta < 0$, for instance because the cost c_θ is low. These are things that "everyone but the worst people do", such as not committing serious offenses or mistreating one's spouse and children, and which are consequently normative in the usual sense that the pressure to conform rises with the behavior's prevalence.

- *"Admirable" or "distinguished"*, if v_θ^* is in the upper tail where $\Delta'_\theta > 0$, for instance because the cost c_θ 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, or actions that confer a rare "status" more generally.

- *"Modal"* if v_θ^* is in the middle range around the minimum of Δ_θ. Both $a = 1$ and $a = 0$ are then common behaviors, leading to weak inferences about agents' types.[11]

It is worth noting that the model generates endogenously the two types of signaling motives which, in previous literature, were taken as alternative assumptions: a desire to signal *conformity* (e.g., Bernheim 1994), and a desire to signal status or *distinction* (e.g., Pesendorfer 1995).[12]

[10] Jewitt (2004) established (i) and the first parts of (ii)-(iii). Harbaugh and Rasmusen (2018) refined the results to include the second parts, and Adriani and Sonderegger (2019) to include (iii).

[11] Other factors affecting the relative strength of honor and stigma include nonlinearities in reputational payoffs, $E[\varphi(v) \mid a]$ (e.g., Corneo and Jeanne 1997), which are equivalent to transformations of the density $f_\theta(v)$, and differential visibility of good and bad deeds (Bénabou and Tirole 2006a).

[12] Brennan and Brooks (2007) 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 unimodal.

C Comparative statics and empirics

C.1 Comparative statics

Let $m_\theta(v, v^*)$ denote agent v's non-extrinsic motivation when the cutoff is some v^*:

$$m_\theta(v, v^*) \equiv ve_\theta - c_\theta + \Delta_\theta(v^*). \tag{7}$$

Recalling that $e_\theta + \Delta'_\theta > 0$, or equivalently $\partial m/\partial v + \partial m/\partial v^* > 0$, the equilibrium cutoff $v^*_\theta(y)$ is thus uniquely given by $m_\theta(v^*_\theta(y), v^*_\theta(y)) + y = 0$.

Definition 1 (motivation-enhancing or -reducing parameter) *A parameter θ is (in equilibrium)*

- *motivation-enhancing* (M^+) *if* $\frac{\partial m_\theta}{\partial \theta}(v^*_\theta(y), v^*_\theta(y)) > 0$*;*
- *motivation-reducing* (M^-) *if* $\frac{\partial m_\theta}{\partial \theta}(v^*_\theta(y), v^*_\theta(y)) < 0$.

This condition is equivalent to the equilibrium cutoff $v^*_\theta(y)$ being decreasing (resp. increasing) with θ.

A sufficient condition for M^+ (resp. M^-) is that $\frac{\partial m_\theta}{\partial \theta}(v, v^*) > 0$ (resp., < 0) for all (v, v^*). Differentiating condition (4) and recalling that $e_\theta \equiv \gamma e_\theta + 1 - \gamma$ yields

$$\frac{\partial v^*_\theta(y)}{\partial \theta} = -\frac{v^*_\theta(y)\gamma\frac{\partial \epsilon_\theta}{\partial \theta} - \frac{\partial c_\theta}{\partial \theta} + \frac{\partial \Delta_\theta}{\partial \theta}(v^*_\theta(y))}{e_\theta + \Delta'_\theta(v^*_\theta(y))}. \tag{8}$$

As can be seen from (7) –or, in equilibrium, (8)– increases in cost (c_θ) are motivation reducing, whereas more intense social monitoring or a greater importance attached by peers to the activity in question (μ_θ, scaling Δ_θ) is motivation increasing. So are increases in the externality (ϵ_θ), but only to the extent that agents are as least partly consequentialist, in the sense that their intrinsic motivation e_θ is tied to the perceived of impact of their actions, as measured by γ. An interesting hybrid case is when learning of a more important externality (e.g., from carbon emissions) also causes agents to monitor more closely and respond more strongly to others' behavior. In a sense, it is now *social vigilance* and enforcement that obey a consequentialist logic. Formally, μ_θ is an increasing function $\psi(\epsilon_\theta)$, so a more significant externality boosts both the intrinsic and the social-esteem motives. All these effects, finally, are amplified by the social multiplier, $s_\theta = 1/[e_\theta + \Delta'_\theta(v^*_\theta)]$.

The results encapsulated in (8) will also be key to understanding expressive law and other forms of persuasion by the principal: if she can alter agents' beliefs about θ, she can influence their motivation, and thus their behavior, without need for material incentives. A particularly illuminating case in that respect is the first type of distributional shift considered below.

D Shifts in societal values

The distribution of preferences in a society or group is a fundamental determinant of what norms and standards will emerge among its members. To show precisely how, we consider here variations in F_θ, while c, ϵ and μ remain fixed.

1. Uniform shift. Let $F_\theta(v) \equiv F(v-\theta)$, with density $f_\theta(v) = f(v-\theta)$, support $V_\theta \equiv [v^{\min}+\theta, v^{\max}+\theta]$ and hazard rate h_θ. Conditionally on θ, the reputational return to choosing $a=1$ is easily seen to be $\Delta_\theta(v) \equiv \Delta(v-\theta)$, where Δ is the reputational concern for $\theta = 0$. Without loss of generality, we can normalize the v's (adding a constant) so that the minimum of Δ occurs at $v=0$, and that of Δ_θ therefore at $v=\theta$.[13] Assuming as before that the equilibrium cutoff $v_\theta^*(y)$ is always interior and thus given by (4), it is easily seen that

$$v_\theta^*(y) - \theta = v_0^*(y+\theta e) \quad \text{for all } \{y, \theta\}, \tag{9}$$

where v_0^* is the cutoff for $\theta = 0$. A known or perceived shift in societal values θ therefore has *the same effect* on equilibrium social norms $\Delta_\theta(v_\theta^*(y))$ and aggregate behavior $\bar{a}_\theta(y)$ as an increase in material *incentives* y of magnitude θe. This equivalence already suggests that, for a principal, communicating about community standards or a firm's culture (θ, v_θ^*, or $\bar{a}_\theta$) can be an attractive substitute to costly rewards and punishments, provided she can achieve credibility.

2. Shifts affecting the tails. Adriani and Sonderegger (2019) extend the analysis in our working paper (Bénabou and Tirole 2011b) to other types of shifts, emphasizing how fatter tails magnify Δ_θ in two important cases.

(a) Truncations. Cutting off either the right or left tail of some initial distribution $F(v)$ at some point in $(v^{\min}, v^{\max})$ reduces signaling. A right truncation ($F_\theta(v) \equiv F(v)/F(v^{\max}-\theta)$, truncated at $v^{\max}-\theta$ with $\theta \geq 0$) reduces the honor $E_\theta^+(v^*)$ from providing the costly signal, without affecting the stigma of not doing so. In contrast, a left truncation ($F_\theta(v) \equiv [F(v) - F(v^{\min}+\theta)]/[1-F(v^{\min}+\theta)]$, truncated at $v^{\min}+\theta$, with $\theta \geq 0$) reduces the stigma $E_\theta^-(v^*)$ associated with the absence of contribution, without affecting the honor. We will say that θ is a "truncation parameter" if θ increases, reducing Δ_θ.

(b) Mean-preserving spreads (MPS). Similarly, signaling incentives intensify when the population becomes more diverse, in the sense of second-order stochastic dominance. If $\int_{v^{\min}}^{v} \frac{\partial F_\theta(v)}{\partial \theta} dv > 0$ for all $v \in (v^{\min}, v^{\max})$ while $\int_{v^{\min}}^{v^{\max}} \frac{\partial F_\theta(v)}{\partial \theta} dv = 0$, then E_θ^+ rises and E_θ^- declines, so both now contribute to raising Δ_θ.

E Empirical Applications

Several recent papers build on the framework of this section to study empirically the determinants of norm compliance and its response to incentives, providing tests of the model in the process. Besley, Jensen and Persson (2023) use a dynamic version of equation (4) to study tax evasion in local British and Welch councils between 1980 and 2009. They first show that when $\mu > 0$, the social multiplier makes temporary shocks to general intrinsic motivation (θ in $F(v-\theta)$) have long-lasting effects, with equilibrium behavior returning only slowly to its original value –monotonically when $\Delta' < 0$, or with oscillations when $\Delta' > 0$. Compliance was initially very high (97%), making it *a priori* a respectable behavior for which a norm (first case) should prevail. The authors then analyze the effect of the temporary (1990 to 1993) switch in the local-tax regime from the traditional property-based one to a poll tax that was highly unpopular (decrease in general motivation θ). In line with the model's predictions about how a temporary negative shock to θ lastingly weakens the social norm, they find that it led to an

[13] Of course, in practice $v \geq 0$ for most agents; the normalization is only meant to simplify the notation.

initial spike in evasion (to about 15%), followed by a long period (until at least 2009, by which time the regime had long reverted to the original one) during which it stayed above its previous level, decreasing back only slowly and monotonically. Also in line with the model, local councils where the initial backlash to the poll tax (increased evasion) was higher remained less compliant (on property-based taxes) than other ones throughout the convergence process, controlling for many economic and political factors.

Jia and Persson (2020) exploit another comparative-statics property, namely how the social multiplier varies with the initial compliance level. As seen in (6), if the reputation function Δ is convex, $s_\theta(y) = 1/[e + \Delta'(v^* - \theta)]$ increases with the level of participation (lower v^*, or higher θ), which tends to make the equilibrium more responsive to incentives when compliance is initially high (and stigma considerations prevail) than when it is low (honor considerations dominate).[14] The application pertains to the choice, by mixed Chinese couples where the man is of the majority Han and the woman of a minority group, of the official ethnic identity they select for their child. The paper first documents a strong society-wide norm to pass on the father's ethnicity, then exploits the gradual introduction by the Chinese government of affirmative-action benefits for minority children. In line with the model, they find that the share of couples passing on the mother's minority identity (thus breaking the patriarchal norm) increased by more, when the incentives went into effect, in places where initial compliance with the norm was higher. Chen (2016) uses a similar comparative static to predict and then verify that executions of deserters in the British Army during World War I had much weaker (even negative) effects on dissuading absences for Irish soldiers (who had weaker identification with the army) than on British ones.

III Optimal laws and incentives in the presence of norms

A Principal's objective function

We now turn from the positive analysis of agents' equilibrium behaviors to the normative analysis of how to set policy in the presence of such social interactions. Consider therefore a principal ("she") who sets the incentive y (subsidy or tax, wage etc.) under symmetric information about θ. This is "the law", whether that of the company or that of the land.[15] The principal's objective function is

$$W_\theta^{SI}(y) \equiv \int_{v_\theta^*(y)}^{+\infty} [ve_\theta + \epsilon_\theta - c_\theta - \lambda y]\, f_\theta(v) dv. \tag{10}$$

[14] Recalling that Δ is quasiconvex, convexity is a relatively weak assumption, at least if the cutoff is not too far away from the mode. The impact of incentives also involves the local density at the cutoff, as $\partial \bar{a}_\theta / \partial y = f(v^* - \theta) s_\theta(y)$. If f is not too decreasing (or sufficiently right skewed as, Persson and Jia assume), however, the sign of $\partial^2 \bar{a}_\theta / \partial y \partial \theta$ is primarily governed by that of Δ''.

[15] We assume costless observation of behaviors by the principal. 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 Huddart (2008) for a model with norms and an informationally constrained principal. Another policy tool can be for the principal to affect the public visibility or memorability of agents' actions, thus scaling the reputational weight at some cost. On the benefits and costs of visibility-based policies, see Prat (2005), Bénabou and Tirole (2006a), Daughety and Reinganum (2009), Bar-Isaac (2012) and Ali and Bénabou (2020).

A first interpretation of (10) is that of a *social planner* facing a cost $\lambda \geq 0$ of public funds.[16] That is, she internalizes agents' welfare, including the fiscal cost that they will collectively bear to provide individual incentives:

$$W_\theta^{SI}(y) = \bar{U}_\theta - (1+\lambda) y \bar{a}_\theta,$$

where $\bar{U}_\theta$, defined in (5), is the sum of all agents' equilbrium utilities, $\bar{a}_\theta$ their total contribution, and $\lambda \bar{a}_\theta y$ the deadweight loss from the required taxation.[17]

In the second interpretation of (10), the principal is a *for-profit company* that trades off inducing greater effort by workers ($\bar{a}_\theta$) against some shadow cost of providing performance-based incentives. We provide in the Online Appendix a simple example of a firm's compensation structure and worker-participation constraint that leads to a profit function reducing to (10). More generally, one could consider principals with other objective functions, such as politicians with private benefits from office holding and career or reelection concerns.

B Pigou and Ramsey with image concerns

In all that follows, we will assume that the principal's objective function (10) is strictly quasi-concave in y (such is clearly the case provided λ is small enough[18]), and the equilibrium cutoff interior. To ensure the latter, we restrict the model's parameters to satisfy, for all $\theta \in \Theta$,

$$v_\theta^{\min} e_\theta + \varepsilon < c_\theta - \epsilon_\theta < v_\theta^{\max} e_\theta - \varepsilon \tag{11}$$

for some fixed, arbitrarily small $\varepsilon > 0$.[19] The optimal incentive is then given as the solution to

$$\frac{\epsilon_\theta + v_\theta^*(y) e_\theta - c_\theta - \lambda y}{e_\theta + \Delta'_{\hat{\theta}(y)}(v^*_{\hat{\theta}(y)}(y))} = \frac{\lambda}{h_\theta(v^*_{\hat{\theta}(y)}(y))}. \tag{12}$$

The interpretation is familiar from Ramsey taxation: the net social marginal benefit of a unit increase in y (inducing $d\bar{a}_\theta = (-\partial v_\theta^*/\partial y)\, f_\theta(v_\theta^*) dy$ new agents to participate) is equated to the deadweight loss from paying the extra reward to all inframarginal agents, $\lambda \left[1 - F_\theta(v_\theta^*(y))\right]$.

In the first-best (FB) case ($\lambda = 0$), (12) reduces to $\epsilon_\theta + v_\theta^*(y) e_\theta = c_\theta$, which is the standard *Samuelson condition* equating the total social benefit and cost of a marginal contribution. Substituting the definition of the cutoff yields the explicit solution $y_\theta^{FB} = \epsilon_\theta - \Delta_\theta\left(\frac{c_\theta - \epsilon_\theta}{e_\theta}\right)$, which we discuss below. In general, y_θ^{FB} could be positive or negative (taxing image-seeking behaviors with low or negative social value). When the externality is sufficiently high,

$$\epsilon_\theta > \max\left\{\Delta_\theta(v_\theta^{\min}), \Delta_\theta(v_\theta^{\max})\right\} = \mu_\theta \max\left\{\bar{v}_\theta - v_\theta^{\min}, v_\theta^{\max} - \bar{v}_\theta\right\}, \tag{13}$$

[16] Equation (10) incorporates agents' utility from contributing (ve_θ) into the principal's welfare function. There are pros and cons of doing so, as discussed by, e.g., Diamond (2006). Our results do not hinge on this specific formulation of W_θ, which only affects the definition of regions in which there is an over- or under-provision of prosocial behavior. That is, we could alternatively assume that $W_\theta^{SI} = \int_{v_\theta^*(y)}^{+\infty} (\epsilon_\theta - c_\theta - \lambda y) f_\theta(v) dv$.

[17] While $y > 0$ is more intuitive, and we will later on impose conditions ensuring that it holds in equilibrium, the linearity of (10) also allows for $y < 0$: action $a = 1$ is then taxed, generating valuable revenue. When levying fines or inflicting other sanctions is costly, the principals' incentive cost is somewhat different: letting $y > 0$ be the fine on $a = 0$, for instance, the last term is replaced by $(1+\lambda) y F_\theta(v^*)$.

[18] At the first-order condition, $\partial W_\theta^{SI}/\partial y = (-\partial v_\theta^*/\partial y)[\epsilon_\theta + v_\theta^*(y) e_\theta - c_\theta - \lambda y] f_\theta(v^*(y)) = 0$, and, for small λ, $\partial^2 W_\theta^{SI}/\partial y^2 \approx (-\partial v^*/\partial y)^2 e_\theta f_\theta(v^*(y)) > 0$.

[19] Condition (11) means that it is socially inefficient (respectively, efficient) for the least (most) motivated agents, with types close to $v_\theta^{\min}$ ($v_\theta^{\max}$) to contribute. It will imply that for y close to the first-best optimum (which delivers $v_\theta^* e_\theta = c_\theta - \epsilon_\theta$), the cutoff remains interior (i.e., the condition given in footnote 7 is satisfied).

it will be the case that $y_\theta^{FB} > 0$, since the function Δ_θ is strictly quasiconvex.

With costly incentives, substituting (8) into (12) leads to an expression closely related to that for y_θ^{FB}, parametrized by λ:

Proposition 1 (modified Pigou and Ramsey) *Under symmetric information,*

(i) The first-best ($\lambda = 0$) incentive is equal to the net externality:

$$y_\theta^{FB} = \epsilon_\theta - \Delta_\theta\Big(\frac{c_\theta - \epsilon_\theta}{e_\theta}\Big). \tag{14}$$

(ii) Let (11) and (13) hold. With costly incentives ($\lambda > 0$), the second-best subsidy solves:

$$y_\theta^{SI} = \frac{\epsilon_\theta - \Delta_\theta(v^*(y_\theta^{SI}))}{1+\lambda} - \frac{\lambda}{(1+\lambda)h_\theta(v_\theta^*(y_\theta^{SI}))s_\theta(v_\theta^*(y_\theta^{SI}))}, \tag{15}$$

where $v_\theta^*(y)$ is given by (4). It is always below the first-best, $y_\theta^{SI} < y_\theta^{FB}$, and decreases with λ, implying the same properties for aggregate compliance, $\bar{a}_\theta^{SI}$.

The first-best case will prove to be an important benchmark under both symmetric and asymmetric information. One must subtract from the standard Pigovian subsidy, ϵ_θ, the *reputational rent* Δ_θ 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 penalty. Our modified-Pigou formula then yields a rich set of comparative-statics results, which we detail below. In particular, given the properties shown in Section II.B for the reputation function Δ_θ, the first-best incentive y_θ^{FB} is *bell-shaped* in the general prosociality or "goodness" of society (uniform shift $F(v-\theta)$), and in the contribution cost c_θ: see Figure 2.

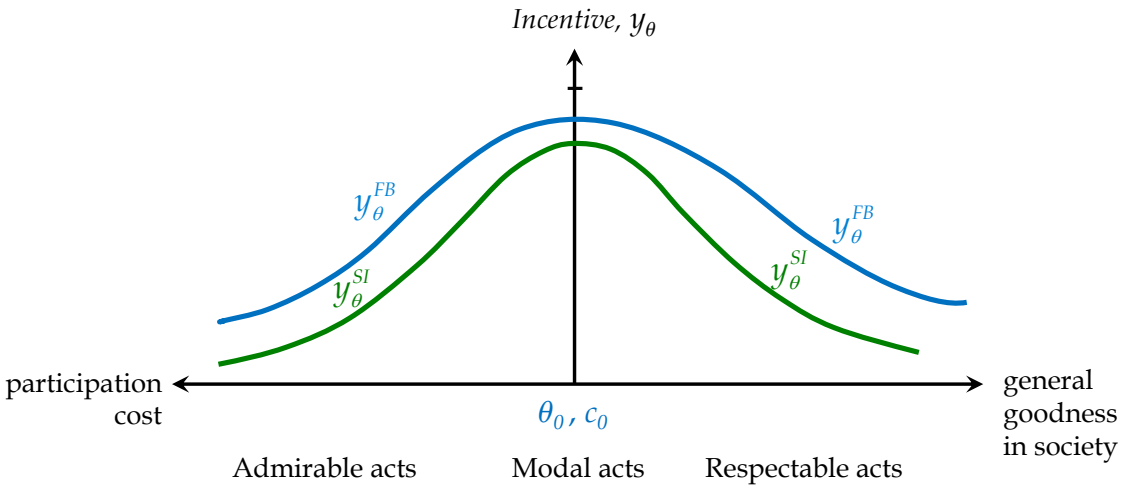

Figure 2: Formal incentives and societal values

The latter part of Proposition 1 demonstrates the robustness of these insights: the second-best y_θ^{SI} also involves a tax on reputation-seeking, and for λ not too large it will share the shape and comparative-statics properties of y_θ^{FB}, while shifting down relative to it as λ increases; see again Figure 2. Furthermore, with a positive shadow cost of providing material incentives, there

is always *under-provision* of prosocial behavior: by (12),

$$\epsilon_\theta + v^*_\theta(y^{SI}_\theta)e_\theta - c_\theta - \lambda y^{SI}_\theta > 0, \tag{16}$$

meaning that the social benefit from the marginal contribution exceeds its social cost. Consequently, any instrument other than y that raises participation is welcomed by the principal.

The intuition for the bell shape is as follows. When prosociality (referring to the shift in $F(v-\theta)$) is generally low or the contribution cost c_θ high, most people do not contribute, so the few who do reap significant honor. Conversely, when prosociality is high or the cost is low, most do contribute, so "bad apples" who fail to participate incur strong stigma. At both ends there is thus a high reputational incentive, making a relatively low y optimal.[20] 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. Formally, we show:

Proposition 2 (uniform shifts in societal preferences) *Let θ index the "goodness of society", shifting the values' distribution uniformly: $F_\theta(v) = F(v-\theta)$, so $y^{FB}_\theta = e - \Delta\left(\frac{c-\epsilon}{e} - \theta\right)$, and let the corresponding versions of (11) and (13) hold.*

(i) When f is strictly unimodal in v (with the mode of Δ normalized to be 0), y^{FB}_θ is single-peaked with respect to θ and c, and maximized at $\theta_0 \equiv \frac{c-\epsilon}{e}$ and $c_{0,\theta} = \epsilon + \theta e$ respectively.

(ii) For any $\varepsilon > 0$, there exists $\bar{\lambda} > 0$ such that for all $\lambda < \bar{\lambda}$, the symmetric-information policy y^{SI}_θ is uniquely defined by (15), strictly increasing for $\theta < \theta_0 - \varepsilon$, and strictly decreasing for $\theta > \theta_0 + \varepsilon$.

Implications

1. The tax deduction rate for 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 reasonably well established "market prices" for naming rights to a university or hospital building, an endowed chair, etc. Similarly, agencies computing corporate social responsibility (CSR) indices could aim to incorporate a "publicity discount" in their scores.

2. Similar distortions driven by visibility (high μ_θ) 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 –typically poorer agents, moreover. As a result, too many dollars flow toward hybrid cars and solar panels relative to housing insulation and efficient furnaces (Ariely, Brach and Meier 2009), and toward fair-trade coffee compared to food kitchens.

3. Consider a new environment-friendly technology, such as electric vehicles, that diffuses more widely as its cost c_θ falls due to technological progress. The optimal subsidy rate should

[20] This result has 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 (E^-_θ and E^+_θ) arise in equilibrium from everyone's actions and inferences. This makes them interdependent, and vary with ("control for") the level of material incentives.

first rise, then fall over time, as owning such a good gradually progresses from being an enviable signal of virtue to a relatively nondescript choice and, finally, a strong social norm.

4. Because they partially crowd out social esteem, material incentives, laws, fines and subsidies 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 social multiplier $s_\theta(y)$ is less than $1/\epsilon_\theta$.[21] Incentives are much more effective (the multiplier exceeds $1/\epsilon_\theta$) for respectable behaviors, such as not stealing or evading taxes, as they are amplified by the dynamics of stigma (crowding in). Where net costs are not too high (a relatively low v^*_θ) and actions easily observable (a high μ_θ), small variations in incentives such as "symbolic" fines can induce large changes in aggregate behavior (e.g., Funk 2007).

C Comparative statics of optimal incentives

We now study more generally how the optimal policy depends on each aspect of the environment encapsulated in the "synthetic" parameter θ. Differentiating (14),

$$\frac{dy^{FB}_\theta}{d\theta} = \left(1 + (\gamma c_\theta + 1 - \gamma)\frac{\Delta'_\theta}{e^2_\theta}\right)\frac{\partial \epsilon_\theta}{\partial \theta} - \frac{\Delta'_\theta}{e_\theta}\frac{\partial c_\theta}{\partial \theta} - \frac{\partial \Delta_\theta}{\partial \theta}. \tag{17}$$

We will assume that the Pigovian subsidy is increasing in the externality, meaning that the term in parentheses is positive. This is always true for admirable, distinction-driven acts, $\Delta'_\theta \geq 0$. For respectable, norms-driven ones, $\Delta'_\theta \leq 0$, variations in image motivation should not be too large (e.g., μ_θ is not too large). Condition (17) then implies that the optimal incentive y^{FB}_θ grows with the size of the externality, decreases with image concerns, and increases (respectively, decreases) with the private cost of prosocial behavior in the case of a norm (respectively, an antinorm). By continuity, the same properties will hold for the second-best level of incentives y^{SI}_θ provided that λ is not too large, which we will assume.

To summarize these results and later on derive their implications under asymmetric information, it will be convenient to introduce the following definition.

Definition 2 (policy monotonicity) *On any interval $[\theta_1, \theta_2]$ where the second-best policy y^{SI}_θ is differentiable, we shall say that condition P^+ (respectively, P^-) holds if there exists $\varepsilon > 0$ such that $dy^{SI}_\theta/d\theta > \varepsilon$ (respectively, $dy^{SI}_\theta/d\theta < -\varepsilon$) for all θ.*

Table 1 below summarizes the effects of parameter changes on individual motivation and optimal incentives, from which we draw the following conclusions:

- Whenever θ affects image incentives (Δ_θ) alone, the only possible configurations are (M^+, P^-) and (M^-, P^+). This holds more generally and does not rely on θ being a goodness, MPS, distribution-truncating or social-monitoring intensity parameter. It reflects instead the fact that (keeping other agents' behaviors fixed), material and image incentives are substitutes in inducing compliance.

[21] Full crowding out (a negative supply response to incentives) requires multidimensional heterogeneity, as described in footnote 5. This phenomenon was investigated elsewhere and is therefore not our focus here.

- By contrast, when θ measures the externality ϵ_θ (and $\gamma > 0$, so that this affects e_θ), (M^+, P^+) obtains: a higher externality intrinsically motivates agents to comply (as long as they are somewhat consequentialist) and simultaneously raises Pigovian taxation, given that their response remains insufficient, by (16).
- The cost of compliance c_θ is fully internalized by the agent (unlike the two externalities ϵ_θ and $-\Delta_\theta$ that directly enter Pigovian taxation) and so does not require any correction of its own. However, it indirectly (i.e., in equilibrium) affects image concerns. A higher cost renders the act less common, making compliance even more admirable when $\Delta' > 0$ (anti-norm), and non-compliance more respectable when $\Delta' < 0$ (norm), affecting optimal taxation accordingly.

	θ is motivation-enhancing (M^+)	θ is motivation-reducing (M^-)
Increase in θ leads to higher incentives (P^+)	• externality $(\theta = e)$	• distribution F_θ, when θ is a truncation parameter • distribution F_θ, when θ is a parameter of goodness and $\Delta'_\theta > 0$ • agent's cost $(\theta = c)$, when $\Delta' < 0$
Increase in θ leads to lower incentives (P^-)	• intensity of social monitoring $(\theta = \mu)$ • distribution F_θ, where θ is a MPS • distribution F_θ, when θ is a parameter of goodness and $\Delta'_\theta < 0$	• agent's cost $(\theta = c)$ when $\Delta' > 0$

Table 1: Comparative statics of individual motivation and optimal incentives

IV The expressive function of law (and other incentives)

A Intuitions: soft or tough law?

When a legislator or other principal with private information about agents' environment sets material incentives –law, rewards, penalties– these will inevitably convey a message about what she knows, and thereby shape their understanding of the prevailing social norms, externality,

or cost of behaving prosocially.[22] Keeping with the normative focus of the previous section, we therefore investigate here a question on which the previous legal and economic literatures do not seem to offer general insights: when should expressive concerns make the law (or other formal incentive) milder, or on the contrary tougher? The intuition for the analysis is as follows:

(a) We saw that prosocial contributions are always insufficient when the provision of incentives is costly. The principal would therefore like to boost motivation through the signal sent to the agents by her choice of y, denoted y_θ^{AI}.

(b) Motivation can be enhanced by signaling a high (low) θ when this parameter is motivation enhancing, M^+ (reducing, M^-). Which case obtains depends on what aspect of agents' environment the informational asymmetry bears on, as shown in Table 1.

(c) Credible signaling hinges, as usual, on a global second-order condition: a principal of type θ must not want to induce in agents a belief $\hat{\theta}$ (say, $\hat{\theta} > \theta$ under M^+, when this would be motivation-enhancing) by setting incentive $y(\hat{\theta})$ instead of the equilibrium $y(\theta)$. Whether this condition holds or fails (the equilibrium must then involve some pooling) depends again, as we will show, on what facet of agents' problem θ corresponds to.

(d) In cases where the first-order condition indeed defines a global optimum, the answer to the expressive-law question can be directly read off from the four combinations of M^+/M^- and P^+/P^- in Table 1, with tough law ($y_\theta^{AI} > y_\theta^{SI}$) on the diagonal and soft law (($y_\theta^{AI} < y_\theta^{SI}$) on the off-diagonal.

We now formalize these intuitions. For simplicity, let θ be perfectly known by the principal, whereas agents only know that it lies in some interval $[\theta_1, \theta_2]$. The legislator or principal's information about θ (or, equivalently, $\bar{a}_\theta$) may for instance derive from having observed the previous behavior and experience of a representative sample. Assume (as a simplification) that social payoffs are based on *long-run* reputations, namely those that will be assigned to contributors and non-contributors after θ becomes publicly known, for instance after everyone has had time to observe average compliance $\bar{a}_\theta$: 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 v is drawn from F_θ. Formally, $E[v \mid a, y]$ is replaced by $E[E_\theta[\tilde{v} \mid a, y] \mid v]]$. When the equilibrium is separating there is no such conditioning on v, as the policy perfectly reveals θ. As to the principal, we will assume that she seeks to maximize social welfare evaluating each of its components (externality, cost, agents' ultimate satisfaction from their contributions) according the objective (or ex-post) value of θ, rather than agent's interim beliefs about it. Finally, in all that follows we impose conditions (11) and (13).

B The informational multiplier

We look for a separating equilibrium in which the planner's policy y_θ^{AI} is strictly increasing (or, decreasing) on $[\theta_1, \theta_2]$. Agents can then invert the policy and infer the true θ as the

[22] When θ indexes c_θ or μ_θ, one can think of each agent's long-run participation cost or magnitude of reputational payoffs being independently drawn from an unknown distribution, with the principal having better information about its mean from previous periods or related population samples.

unique solution $\hat{\theta}(y) \in [\theta_1, \theta_2]$ to $y^{AI}_{\hat{\theta}(y)} \equiv y$. The resulting cutoff (here again assumed interior) is then $v^*_{\hat{\theta}(y)}(y)$, which depends on y through both the standard and the signaling channels. The principal's objective function is now

$$W^{AI}_{\theta}(y) \equiv \int_{v^*_{\hat{\theta}(y)}(y)}^{+\infty} [ve_\theta + \epsilon_\theta - c_\theta - \lambda y] \, f_\theta(v) dv. \tag{18}$$

The first-order condition (FOC) for maximizing $W^{AI}_{\theta}(y)$ is:

$$\left(\frac{\epsilon_\theta + v^*_\theta(y) e_\theta - c_\theta - \lambda y}{e_\theta + \Delta'_{\hat{\theta}(y)}(v^*_{\hat{\theta}(y)}(y))} \right) \left(1 + \left(v^*_{\hat{\theta}(y)} \gamma \frac{\partial \epsilon_\theta}{\partial \theta} - \frac{\partial c_\theta}{\partial \theta} + \frac{\partial \Delta_\theta}{\partial \theta}(v^*_{\hat{\theta}(y)}(y)) \right) \hat{\theta}'(y) \right) = \frac{\lambda}{h_\theta(v^*_{\hat{\theta}(y)}(y))}, \tag{19}$$

recalling that $\partial v^*_\theta / \partial y$ is given by (8). This equation embodies the key idea of "the law as a signal." The difference with (12), i.e. the second bracket on the left-hand side of (19), thus reflects the way the principal takes into account that: (i) agents will draw inferences from her policy choice, as captured by the term $\hat{\theta}'(y) = 1/(y^{AI}_\theta)'$, the sign of which is governed by condition P^+/P^-, provided the incentive varies with θ the same way for symmetric and asymmetric information, i.e. the policy schedules are "comonotonic"; (ii) their resulting beliefs over θ will affect their behavior, through either intrinsic or social-image motivation; this corresponds to the term multiplying $\hat{\theta}'(y)$, previously encountered in equation (18) and the sign of which corresponds to the M^+/M^- property.[23] This entire *informational multiplier*, embodying the *expressive content of the law*, then combines with the previously analyzed social multiplier, $1/(e_\theta + \Delta'_\theta)$, to amplify or dampen agents' response to incentives, and therefore the optimal policy.

C Optimal incentives with norms: asymmetric information

C.1 Properties of separating equilbria

Here again, the case of no deadweight loss provides a useful benchmark.

Proposition 3 (costless incentives) *Let $\lambda = 0$. If the first-best solution $y^{FB}_\theta = \epsilon_\theta - \Delta_\theta(\frac{c_\theta - \epsilon_\theta}{e_\theta})$ satisfies P^+ or P^- over $[\theta_1, \theta_2]$, it remains on this interval an asymmetric-information equilibrium of the game in which the principal just selects an incentive. When neither property holds, the first-best outcome can still be implemented in equilibrium through an announcement of the state of nature θ and the choice of incentive $y = y^{FB}_\theta$.*

Intuitively, when the principal can avail herself of a costless instrument to set the cutoff v^*_θ to its optimal level, she has no need to manipulate agents' beliefs about their environment (θ). By contrast, when incentives are costly, she will try (although ultimately not succeed, in a separating equilibrium) to distort beliefs in the direction that raises compliance. To show precisely how, we establish key properties of the solution to the first-order condition for the optimal policy.

[23] Both P^+/P^- and M^+/M^- pertain here to the policy y^{AI}_θ. The former was initially defined for y^{SI}_θ, but will carry over to y^{AI}_θ for λ small enough. The latter was defined for any incentive level y.

Proposition 4 (solution to FOC) *Let $\lambda > 0$ be small enough. Under policy monotonicity, P^+ or P^-, of the symmetric-information incentive y_θ^{SI}, the differential equation (19) has a unique solution y_θ^{AI} on $[\theta_1, \theta_2]$ satisfying the no distortion at the boundary condition (NDB)*

$$y_{\theta_1}^{AI} = y_{\theta_1}^{SI} \text{ if } M^+ \text{ holds or } y_{\theta_2}^{AI} = y_{\theta_2}^{SI} \text{ if } M^- \text{ holds,}$$

and it is comonotonic (COM) with the symmetric information policy: $(y_\theta^{AI})'(y_\theta^{SI})' > 0$. [24]

The next proposition, illustrated in Figure 3, focuses on this solution (thus neglecting the second-order condition, which we examine later on). It states when expressive concerns will make the principal want to give agents weaker or stronger incentives than she would under symmetric information. Note that the four quadrants map exactly to those of Table 1.

Proposition 5 (determinants of soft or tough law) *When the global second-order condition is satisfied, so that the necessary condition (B.2) indeed defines a separating equilbrium, (i) For all λ below some $\bar{\lambda} > 0$, the equilibrium incentive y_θ^{AI} is, like y_θ^{SI}, strictly positive and increasing in θ under P^+, and strictly positive and decreasing under P^-.*

(ii) The principal's private information about θ makes her set, for all $\theta \in (\theta_1, \theta_2)$:

- *Lower-powered incentives, $y_\theta^{AI} < y_\theta^{SI}$, under either (M^+, P^-) or (M^-, P^+).*
- *Higher-powered incentives, $y_\theta^{AI} > y_\theta^{SI}$ under either (M^+, P^+) or (M^-, P^-).*

(iii) There is always underprovision of prosocial behavior: $b_\theta^{AI} \equiv v_\theta^*(y_\theta^{AI})e_\theta + \epsilon_\theta - c_\theta - \lambda y_\theta^{AI} > 0$.

The intuition for how expressive concerns make the law softer or tougher depending on whether the "signs" of properties M and P are opposite or the same can be read off the second term in (B.2): in the first case the informational multiplier is smaller than 1, as the "message" conveyed by higher incentives crowds out some other (intrinsic or reputational) source of reputation, and as a result the principal uses them less. In the second case the multiplier is greater than 1, and this greater effectiveness of incentives (relative to their cost) makes the principal use them more. We provide below examples of both cases.

The intuition for why the net social product b_θ^{AI} of the marginal contribution is strictly positive, finally, reflects the fact that the informational multiplier is (in equilibrium) always positive; see again (B.2). This is obvious when M and P are of the same "sign", and indeed if b_θ^{AI} was negative, the principal could then simultaneously economize on incentives and decrease the excessive participation by lowering y marginally. When M and P are of opposite "signs" this is more subtle, as the direct and informational effects of incentives on participation go in opposite directions; we show in the Online Appendix (Lemma 3) that the former always dominates (there is never net crowding out), so that, here again, if b_θ^{AI} were negative the principal could simultaneously raise it toward zero and save money by reducing y.

Softer law. An important illustration of soft law, $y_\theta^{AI} < y_\theta^{SI}$ is provided by signaling social standards in *norm-driven* activities, $\Delta'(v - \theta) < 0$. In essence, a lower y credibly conveys the

[24] The Online Appendix shows that an allocation that satisfies (COM) as well as no-distortion-at-one-of-the boundaries necessarily satisfies NDB at θ_1 under M^+, and at θ_2 under M^-.

message: "everyone does it, except disreputable people, who suffer substantial stigma; that is why we do not need to provide very strong extrinsic incentives"

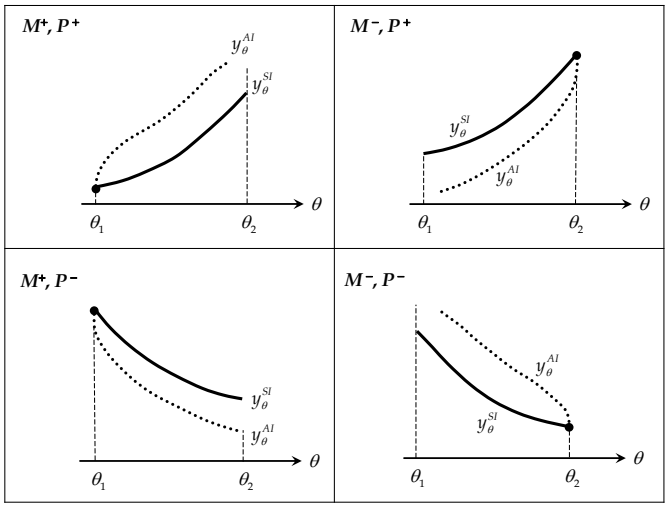

Figure 3: Incentives under symmetric and asymmetric information.

Tougher law. When the principal signals the *magnitude of the externality,* expressive concerns now call for tougher law. Intuitively, the perception of a high e_θ enlists each individual's prosociality, (as long as $\gamma > 0$). Because a high e_θ also "naturally" (under symmetric information) leads to a high Pigovian subsidy, the principal optimally credibly it by setting $y_\theta^{AI} > y_\theta^{SI}$.

Further applications, corresponding to the other quadrants of Figure 3, are presented below.

D Sufficient conditions for a separating equilibrium

Mailath (1987)'s classic analysis of signaling games, in which the sender of the signal is always better off when thought of as a higher type (more productive, more generous, etc), does not apply to our game, as our payoff is not monotonic in the beliefs induced by the policy variable y. Our analysis therefore involves a non-trivial extension of Mailath's pioneering work. Even under his regularity conditions, for instance, our second-order conditions need not be satisfied by the solution to (FOC), which then pushes toward pooling in some environments.

Let $\mathcal{W}(\theta, \hat{\theta}, y)$ denote the payoff of a type-θ principal when offering incentive y and being perceived as type $\hat{\theta}$:

$$\mathcal{W}(\theta, \hat{\theta}, y) \equiv \int_{v_{\hat{\theta}}^*(y)}^{+\infty} [ve_\theta + \epsilon_\theta - c_\theta - \lambda y] f_\theta(v) dv, \tag{20}$$

and define the benefit from a marginal contribution as:

$$b(\theta, \hat{\theta}, y) = v_{\hat{\theta}}^*(y) e_\theta + \epsilon_\theta - c_\theta - \lambda y. \tag{21}$$

The next proposition provides two sufficient conditions, referred to jointly as (SOC), which guarantee that the solution y_θ^{AI} to (FOC), (NDB) and (COM) from Proposition 4 also satisfies global incentive compatibility. This means that the principal has no profitable deviation, whether to on-path incentives (belonging to the graph $\mathcal{G} \equiv \{y_\theta^{AI}\}_{\theta \in ([\theta_1, \theta_2])}$, nor to off-path ones

(beliefs following $y \notin \mathcal{G}$ must then be chosen appropriately), and therefore $\{y_\theta^{AI}\}_{\theta\in[\theta_1,\theta_2]}$ defines a separating equilibrium.

Proposition 6 (second-order conditions/SOC) *Let $\lambda > 0$ be small enough and consider $y(\theta)$, a differentiable and strictly monotonic function satisfying (FOC), (COM), and (NDB). The condition:*

$$\mathcal{A}(\theta,\hat{\theta}) \equiv y'(\hat{\theta})b(\theta,\hat{\theta},y(\hat{\theta}))\frac{\partial(b(\theta,\hat{\theta},y(\hat{\theta}))h_\theta(v^*_{\hat{\theta}}(y(\hat{\theta}))))}{\partial\theta} \geq 0, \qquad (\text{SOC}_1)$$

(i) taken at $\hat{\theta} = \theta$, is a necessary condition for the function $y(\cdot)$ to be a separating equilibrium; (ii) satisfied at all $\hat{\theta}$ and θ, is a sufficient condition for the function $y(\cdot)$ to be a separating equilibrium, provided that the single-crossing condition

$$\mathcal{B}(\theta,\hat{\theta},y) \equiv y'(\hat{\theta})\frac{\partial^2\mathcal{W}(\theta,\hat{\theta},y)}{\partial\theta\partial y} \geq 0 \qquad (\text{SOC}_2)$$

holds at $\hat{\theta} \in \{\theta_1,\theta_2\}$, for all y.[25]

The key condition is $\mathcal{A} \geq 0$, which we show ensures that no type θ wants to deviate to some other equilibrium $y(\hat{\theta})$; the single-crossing condition $\mathcal{B} \geq 0$ is used only to rule out off-path deviations.

E Expressive law and its limits

Recall that the parameter θ on which the principal has private information may affect either agents' image concerns (through various changes in the distribution $F_\theta(v)$ of societal values, or social vigilance μ_θ), or their intrinsic sources of motivation (externality e_θ or participation cost c_θ). In the Online Appendix we examine in each case the signs of $\mathcal{A}$ and $\mathcal{B}$, leading to three propositions. The first one establishes the existence of a separating equilibrium of each type illustrated in a quadrant of Figure 3, in settings where both (SOC_1) and (SOC_2) hold with strict inequality.[26] The second pertains to knife-edge cases in which $\mathcal{A} = 0$, and the third to settings where $\mathcal{A} < 0$, so that some pooling must occur.

Proposition 7 (expressive law) *Let $\lambda > 0$ be small enough. The solution to (FOC) satisfying (COM) and (NDB) also satisfies (SOC_1)-(SOC_2) with strict inequalities, and therefore defines a separating equilibrium when θ :*

(a) shifts the distribution $F_\theta(v) = F(v-\theta)$, which has an increasing density f_θ, and a norm prevails, $\Delta'_\theta < 0$, [27] *or*

(b) operates a right truncation of the distribution ($F_\theta(v) = F(v)/F(v^{\max}-\theta)$), or

(c) affects the externality ϵ_θ (with $\partial\epsilon_\theta/\partial\theta > 0$), or

[25] In some applications, we will use the weaker condition that (SOC_2) hold over the set of y's such that $b(\theta,\hat{\theta},y) > 0$, to which we show that the search for the optimal incentive can be restricted.

[26] The four cases correspond respectively to the Southwest, Northeast, Northwest and Southeast quadrants.

[27] The condition $f'_\theta \geq 0$ is sufficient (but not necessary) to ensure $\mathcal{B} > 0$ over the relevant range of y. When f_θ is symmetric around its mode, Lemma 1(iv) implies that it is equivalent to the activity being a norm, so this is not an additional assumption. For a unimodal distribution and a widespread activity ($F_\theta(v^*_\theta(y_\theta^{AI}))$ small enough), $f'_\theta(v^*_\theta(y_\theta^{AI})) > 0$ as well.

(d) affects the contribution cost c_θ (with $\partial c_\theta/\partial\theta > 0$) and an anti-norm prevails, $\Delta' > 0$.

The equilibrium involves soft law ($y_\theta^{AI} < y_\theta^{SI}$) when θ shifts societal values (cases (a,b)) and tough law ($y_\theta^{AI} > y_\theta^{SI}$) when it affects the externality (case (c)), or the contribution cost in the presence of an anti-norm (case (d)).

Proposition 8 (full pooling) *For both the social-vigilance and the left-truncation applications, $\mathcal{A} = \mathcal{B} = 0$. There exists a full-pooling equilibrium, and it is strictly preferred to any other equilibrium by all types of principals in the first case, and in the second one when $F(v)$ is uniform.*

Proposition 9 (absence of separating equilibrium) *For a distributional shift $F(v-\theta)$ in the case of an anti-norm ($\Delta' > 0$), or a cost uncertainty in the case of a norm ($\Delta' < 0$), $\mathcal{A} < 0$, so there exists no separating equilibrium.*

To get some intuition about how these results reflect the incentive compatibility of the message which each principal wants to send through her choice of y, and the role of the hazard rate h_θ therein (see SOC_1), consider the case of a shift $F(v-\theta)$, for which $\mathcal{A} = b^2 y'(\hat{\theta})(\partial h_\theta/\partial\theta)$. Under a norm ($\Delta'_\theta < 0$), the principal wants to claim a high θ to increase the shame attached to not contributing. By (COM) this must be expressed through a low-powered incentive, as under symmetric information. Furthermore, a high-θ principal gains more than a low-θ one from any decrease in y, as this payment is pocketed by a higher fraction of agents scaled by the marginal impact: $(1 - F_\theta(v))/f_\theta(v)$ rises with θ, by the monotone-hazard-rate property. The message sent to the agents through a lower y is thus concordant with the principal's incentive ($\mathcal{A} > 0$). With an anti-norm, ($\Delta'_\theta > 0$), in contrast, the principal wants to claim a low θ in order to leverage honor seeking. This would again have to be signaled by low-powered incentives, since (COM) now implies that $y'(\hat{\theta}) > 0$. As just seen, however, a low-θ principal gains less than a high-θ one from reducing y. Lower incentives therefore cannot be a credible signal of a lower θ ($\mathcal{A} < 0$) and so there is no separating equilibrium.

Suppose, finally that θ indexes social vigilance, μ_θ. All types of principals would like to signal that vigilance is high by setting a low y: because μ_θ does not enter their payoff function (as reputation is positional) there is no sorting condition ($\mathcal{A} = 0$).

V Persuasion and norms-based interventions

When material incentives are unavailable or too costly, a principal may try to affect collective behavior through direct communication. Social scientists distinguish between two types of interventions aimed at altering norms. *Descriptive norm interventions* correspond to communicating with agents about the average $\bar{a}_\theta$, which in turn reflects some preference parameter like θ that they are imperfectly informed about. *Prescriptive norm interventions*, from public campaigns to individualized "smiley faces", can be understood as communicating about ϵ ("people are strongly affected by this problem") or about μ ("people make strong judgments based on this behavior"), which boosts social pressure Δ both directly, through an increase in the perception of social vigilance, and, for respectable acts, indirectly by making good behavior

more of the norm. As we show below, however, even a fully benevolent principal will try to selectively disclose positive information about $\bar{a}_\theta, \epsilon$, or μ.[28] Agents, conversely, will interpret negatively the absence of evidence disclosure.

We assume here that the principal cannot or does not vary incentives, so y is fixed, say at $y = 0$ for notational simplicity. More generally, the material incentive is low enough that there is always too little prosocial behavior. We posit condition (13), so that for $y = 0$ greater participation always raises social welfare.

Let agents be imperfectly informed about current "community standards", namely the overall behavior of the population against which theirs will be judged. Indeed, these standards shift with the underlying distribution of preferences in society, $F(v - \theta)$, which is hard for an individual to observe. In contrast, we take e, c and μ as fixed.[29] Agents' prior belief about θ is that it lies in some interval $[\theta_1, \theta_2] \subset \Theta$, with distribution $G(\theta)$. The principal, on the other hand, may learn the value of θ, for instance from having observed previous aggregate behavior $\bar{a}_\theta$.[30] Specifically, suppose that she receives hard information about θ with probability q; she can then reveal it, or claim that she has no such data (probability $1 - q$). Upon disclosure, the cutoff is the symmetric-information one v^*_θ, given by (4). In the absence of disclosure, we can show that (provided μ is not too large) agents' equilibrium choices are again defined by a cutoff, denoted $v^*_\emptyset$. Since greater participation increases social welfare, the principal discloses if and only if $v^*_\theta \leq v_\emptyset$. Recalling that v^*_θ is decreasing (resp. increasing) in θ under M^+ (resp., M^-), this implies that, in any equilibrium, the disclosure rule is defined by a cutoff for θ.

Proposition 10 (norms-based interventions)

(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 $\theta \geq \tilde{\theta}$ (resp. $\theta \leq \tilde{\theta}$) under M^+ (resp. M^-).

(ii) In any stable equilibrium, there is more disclosure ($\tilde{\theta}$ decreases), the higher q is.

Pluralistic ignorance and "social proof". In what precedes, the aggregate preference shock θ and average behavior $\bar{a}_\theta$ 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) at the time of their action choice. 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 may have fairly good observations of the distribution of choices. The idea of pluralistic ignorance however, is that "social proof" (equilibrium behavior $\bar{a}_\theta$) 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

[28] 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)

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

[30] Examples include electricity consumption, recycling, tax compliance, etc. Ali and Bénabou (2020) analyze the "reverse" problem in which the principal seeks to learn about θ, and it is the population who (in the aggregate) has more information about it.

in interpreting $\bar{a}_\theta$.[31] Alternatively, pooling can also make $\bar{a}_\theta$ imperfectly informative, thereby restoring the scope for the principal's disclosure (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 $\theta = \theta_H$ (respectively, $\theta = \theta_L$) there is positive participation, $0 < \overline{a} \leq 1$ (respectively, zero participation, $\bar{a}_\theta = 0$); (ii) the prior probability that $\theta = \theta_L$ is high enough that, when agents are uninformed, no one contributing is the (generically unique) equilibrium.[32] Thus, pluralistic ignorance prevails when agents observe $\bar{a}_\theta = 0$, and dispelling it by (credibly) disclosing that $\theta = \theta_H$ increases participation in the socially desirable activity. This corresponds for instance to the norm-shifting interventions of Prentice and Miller (1993) for alcohol consumption by college students, and of Bursztyn, González and and Yanagizawa-Drott (2020) for Saudi men's allowing their wives to work outside the home. Conversely, Bursztyn, Egorov and Fiorin (2020) show that inducing subjects to think that Donald Trump won the vote in their local area erodes the norm against the expression of xenophobia, making them more prone to direct a donation to an anti-immigrant organization. This corresponds to the case where θ is lower than subjects' priors, so that the principal would want to withhold the information (and individuals, if sophisticated, would interpret such silence skeptically). In Galbiati et al. (2021), pluralistic ignorance was dispelled through expressive law rather than direct communication: introducing (even weakly enforced) lockdown measures against COVID-19 substantially reduced the public's large initial underestimation of the extent of popular support for social distancing.

VI Extensions

A 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.[33] 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 their behavior in other realms. We provide three important applications of this idea.

A.1 "Commodification" and society's resistance to economists' prescriptions

Economists' typical message about the effectiveness and desirable normative properties of incentives often meets with considerable resistance. Examples include tradeable pollution permits, financial incentives for students, teachers or civil servants, unemployment benefits that decrease over time to encourage job search, layoff taxes rather than regulation, taxes rather than prohibition for drugs and prostitution, etc. While misinformation and special-interest considerations

[31] This is done in Ali and Bénabou (2020), with agents receiving noisy idiosyncratic signals about both aggregate shocks, from their own payoffs.

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

[33] For instance, Keizer, Lindenberg and Steg (2008) posted fliers (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.

are surely relevant, 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 light on this phenomenon. Strong or pervasive incentives tend to convey the sense that "society is rotten" - endemic opportunism, corruption, tax evasion, etc.- with everyone primarily looking out for themselves.[34] This dim view, in turn, can be very damaging in other, non-incentivized activities that are mostly norm- and trust-driven ($\Delta'_\theta < 0$). More generally, traditional economics typically brings a message, both positive (empirical studies) and normative (policy recommendations) that is bad news about human motivations, which may encounter resistance, for two reasons. First, individuals, organizations and societies often do not like to hear bad news, preferring to maintain pleasant (albeit costly) illusions about themselves.[35] Second, economists' traditional findings are drawn predominantly from b-type behaviors, where incentives are readily available and the role of social norms limited. Insufficient attention may have been paid to a-type behaviors, in which incentives are unavailable and reliance on social norms important.

A simple example will convey the main insight. Agents' prosociality types are again drawn from a continuous distribution $F(v-\theta)$, with θ taking here only two possible values, θ_H and $\theta_L < \theta_H$, with probabilities ρ and $1-\rho$. Agents can engage in two activities, a and b, both involving 0-1 decisions, with respective externalities ϵ_a, ϵ_b:

(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, $y_a \equiv 0$ and $\mu_a = \mu > 0$. We will assume that, for all θ, the externality ϵ_a is sufficiently large that there is an undersupply of prosocial behavior in activity a, and that it is subject to a norm ($\Delta'_\theta < 0$).

(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, $\mu_b = 0$, and $y_b = y \geq 0$, with or without an associated shadow cost $\lambda \geq 0$.

Agents' cutoff when the principal sets an incentive $y \geq 0$ is $v_b^* = \frac{c_b - y}{e_b}$. Hence, under symmetric information, the incentive y_θ^{SI} is given by maximizing $W_b(y,\theta) \equiv \int_{(c_b-y)/e_b}^{+\infty} (ve_b + \epsilon_b - c_b - \lambda y) f(v-\theta)dv$ over y, with the monotone hazard rate implying that $y_{\theta_H}^{SI} < y_{\theta_L}^{SI}$.

Consider now the non-incentivized activity a. For a given cutoff v_a^*, the principal's welfare is $W_a(v_a^*,\theta) \equiv \int_{v_a^*}^{+\infty} (ve_a + \epsilon_a - c_a) f(v-\theta)dv$. Given any updated beliefs $\hat{\rho}(y) \equiv \Pr(\theta = \theta_H | y)$ from observing incentive y in activity b, the cutoff is $v_a^*(\hat{\rho}(y))$, where for all ρ we define $v_a^*(\rho)$ as the solution to

$$v_a^* e_a - c_a + [\rho \Delta_{\theta_H}(v_a^*) + (1-\rho)\Delta_{\theta_L}(v_a^*)] = 0,$$

The principal's total welfare is thus $W_a(v_a^*(\hat{\rho}(y)),\theta) + W_b(y,\theta)$, clearly showing the expressive

[34] See, e.g., Frey (1997), Bowles (2008), Bowles and Polania-Reyes (2012).
[35] See, e.g., Bénabou and Tirole (2006b) and Bénabou (2013).

spillover from y onto the norm in activity a. As discussed earlier, intuition then suggests that she may want to reduce the power of incentives bearing on activity b to avoid undermining the social norm in activity a. Whether this strategy is incentive-compatible cannot be taken for granted, however, as spillovers introduce a new force toward pooling (relative to Proposition 7). Looking at activity b in isolation, a high-type principal benefits more than a low one from reducing the costly incentive. On the other hand, the low-type principal is more likely to gain more from enlisting the social norm in activity a, to the extent that this affects the behavior of a larger number of marginal agents.[36] These two forces work in opposite directions to determine the set of incentive compatible allocations. The general complexity of second-order conditions with multiple activities is the reason we restrict θ to two values, leading to intuitive results.

Proposition 11 (commodification spillovers) *Suppose that $F_\theta(v) = F(v-\theta)$ has an increasing density and that θ can take two values θ_H and $\theta_L < \theta_H$, with $\theta_H - \theta_L$ not too large. Suppose further that the benefits ϵ_a in the non-incentivized activity are large enough that there is always undersupply of contributions. Then, for λ small enough, soft law prevails: type θ_L setting $y^{SI}_{\theta_L}$ in the controlled activity b and type θ_H setting an appropriate $y^{AI}_{\theta_H} < y^{SI}_{\theta_H}$, together with off-path beliefs $\hat{\rho}(y) = 0$ for all $y > y^{AI}_{\theta_H}$ and $\hat{\rho} = 1$ for all $y \leq y^{AI}_{\theta_H}$, constitutes a least-cost separating equilibrium that is robust to D1.*

A.2 Zero-tolerance policies

Let the two behaviors, a and b, generate externalities $\epsilon_{a,\theta}$ and $\epsilon_{b,\theta}$ that both increase with θ, representing a lower "tolerance" of the planner for this general class of anti-social behaviors, or equivalently her private information about the extent to which the agents whose welfare she maximizes suffer from them. As before, the planner can impose penalties for choosing $b = 0$ or rewards for choosing $b = 1$ at relatively low cost, whereas for a behaviors formal incentives are either not feasible or very costly. An example is where b is petty crime and nuisances in a community (fare evasion, shoplifting, vandalism, public indecency), for which enforcement is easily implemented and observable by fellow citizens on a daily basis, whereas a is more serious crime (theft, drug dealing, violence), which fewer people commit and for which formal enforcement is much more costly ($\lambda_a >> \lambda_b$), unpredictable and remote from public view (long trials in a faraway court, with a high burden of proof). In such cases of "correlated harms" with differentially costly incentives, a higher y_b can convey a signal that not only e_b but also e_a is large, and thereby affect a behavior through two expressive channels.

The first is that of *individual responsibilization*, operating through the intrinsic motivation term $ve_{a,\theta}$. When persuaded that $e_{a,\theta}$ is also important, intrinsically motivated individuals respond by voluntarily lowering their level of a. The second channel, related to the enforcement aspect of broken-windows theory in the crime literature, is that of boosting *social vigilance*. Here, a belief that a behavior is harmful to a community causes its members to pay more attention to, and exert more ostracism against, wrongdoers: $\mu_\theta = \psi(\epsilon_{a,\hat{\theta}})$, with $\psi' > 0$ as discussed in Section II.C. By setting a high y_b on b behavior, the principal can then seek to

[36] To see this, decompose social welfare into $\mathcal{W}(\theta,\hat{\theta},y(\hat{\theta})) = \mathcal{W}_a(\theta,\hat{\theta}) + \mathcal{W}_b(\theta,\hat{\theta},y(\hat{\theta}))$, with $\frac{\partial^2 \mathcal{W}_a}{\partial\theta\partial v_a^*} = (\epsilon_a + v_a^* e_a - c_a) f_\theta'$. Because f_θ' tends to be positive under a norm (and actually is positive for a norm if the distribution is symmetric), this force may cause (SOC_1) to fail and lead to pooling.

convince agents that the community "will not stand" for a-misconduct either, but punish it with stronger social sanctions in lieu of missing or insufficient formal incentives y_a.

A.3 Norm-based interventions and broken-windows theory

A slightly different form of strategic (non)disclosure than the one studied in Section V allows us to capture the idea behind the "broken-windows" theory of local order.[37] Assume that the principal does not have access to material incentives ($y_a = y_b = 0$), and that negative externalities in activity b have been exerted. The principal can, at a cost, undo them (repair the broken windows, clean up the graffiti, etc.), in order to avoid conveying the image that the local community does not really care (low θ), thereby jeopardizing civil behavior in some other activity a. "Repairing" is then a form of "not disclosing" bad news about prosociality, provided agents observe or remember the result (intact windows and buildings) more than the process itself. Let the cost of repairing a fraction of the $1 - \bar{b}_\theta$ "broken windows" or other vandalized public goods be $\xi(b - \bar{b}_\theta,)$ for $b \in [\bar{b}_\theta, 1]$. If activity a is "important" enough, in that, for all θ,

$$\xi(1 - \bar{b}_\theta) \leq \int_{v_a^*(G)}^{v_a^*(\theta^{\min})} \big(v_a e_a + \epsilon_a - c_a\big) f(v - \theta) dv,$$

where $v_a^*(G)$ denotes agents' threshold when their belief about θ is just their prior G, then there exists a full-pooling equilibrium in which all windows are repaired ($b(\theta) = 1$ for all θ), with off-path beliefs $\hat{\theta} = \theta^{\min}$ in case $b < 1$.

B Cruel and unusual punishments

To sanction socially undesirable behaviors, standard economic considerations generally argue for using fines, compensation community service and other "efficient" punishments. These are often politically unpopular, however: large fractions of the electorate demand long and harsh incarcerations, as well as various forms of public humiliation.[38] In many countries, the death penalty and corporal punishments are still the law of the land and, when public, heavily attended. At the same time, a growing number of nations are renouncing what they deem "cruel and unusual" punishments or means of coercion. [39] Such decisions, moreover, are not primarily based on practical considerations of optimal deterrence, but on "what it makes us", what "civilized" people do or don't do –in other words, on *expressive reasons.*

These apparent contradictions can be resolved with a version of expressive law in which the key variable is the prevalence of *vindictiveness*, or even sheer *spitefulness,* in a society:

[37] As summarized in Wikipedia, "In criminology, the Broken Windows Theory states that visible signs of crime, antisocial behavior and civil disorder create an urban environment that encourages further crime and disorder, including serious crimes. The theory suggests that policing methods that target minor crimes, such as vandalism, loitering, public drinking and fare evasion, help to create an atmosphere of order and lawfulness." We capture here the first aspect, and in Section VI.A.2 the second one.

[38] See, e.g., Kahan (1996, 1997) 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.

[39] For instance, the European Community's Charter of Fundamental Rights makes renouncing the death penalty (Article 2) and "inhuman or degrading treatment or punishment" (Article 4) preconditions for membership, and the United States declares debate over exceptions) torture to be "abhorrent both to American law and values and to international norms" (18 U.S.C. §§ 2340-2340A).

some fraction of agents enjoy or easily tolerate cruelty to others, especially those toward whom they feel aggrievement, and do not feel constrained by a notion of respect for universal human dignity. Making criminals suffer intense physical or psychological pain, especially publicly (being a spectator enhances this form of "consumption") is an opportunity, and possibly an excuse, to obtain such enjoyment. At the same time, most people dislike thinking that they live among cruel or vindictive individuals.

Formally, there is a choice of punishment technologies, ranging from ordinary but expensive ones (fines, jail) to cruel but cheap ones (corporal or shaming punishments). Suppose a crime has been committed. Which type of sanction should be used, knowing that civilized ones are costlier –the material cost of the policy is λy, where $\lambda \geq 0$ denotes how "civilized" the punishment is? Note that we are not interested here in the proper *level* y of the punishment, which presumably reflects optimal deterrence. Rather, we focus on the *structure* of the punishment for a given level (incentive power on agents' behavior), and its impact on the perception of societal values.

An agent with type $v \in (-\infty, +\infty)$ has disutility $v\chi(\lambda)$ where χ is positive, decreasing and weakly convex. Suppose that v is distributed according to $F_\theta(v) = F(v-\theta)$, where θ is is drawn from $G(\theta)$ with support $[\theta^{min}, \theta^{max}]$. Agents derive utility $\kappa\hat{\theta}$ from their beliefs about θ –society's general aversion to violence or cruelty– due to either collective self esteem or anticipatory utility with respect to future interactions with others. Normalizing $E_\theta[v] = \theta$, social welfare is:[40]

$$W = \kappa\hat{\theta} - \lambda y - \chi(\lambda)\theta$$

Note that, in contrast to previous sections, the principal internalizes here agents' utility from their beliefs about the type of society they live in, rather than evaluate welfare according to her private knowledge of the true distribution. This, rather than seeking to affect their behavior, is what now gives her an incentive to distort her policies for expressive purposes.

Under symmetric information, $\partial W/\partial\lambda = -y - \chi'(\lambda)\theta$, so the principal chooses $\lambda_\theta^{SI} = 0$ (maximally cheap but cruel punishments) for $\theta \leq -y/\chi'(0) \equiv \theta^*$, and λ_θ^{SI} strictly increasing in θ, given by $-\chi'(\lambda_\theta^{SI}) = y/\theta$, for $\theta > \theta^*$. Under asymmetric information, the (FOC) differential equation writes

$$\kappa\frac{d\theta}{d\lambda} = y + \chi'(\lambda)\theta.$$

Together with the (NDB) condition that $\lambda_0^{AI} = \lambda_0^{SI} = 0$, this defines a unique λ_θ^{AI} that is everywhere strictly increasing and strictly above λ_0^{SI}. Thus, expressive concerns lead to the use of less cruel forms of punishment, in spite of their greater cost. In the linear case where $\chi(\lambda) = \chi_0 - \lambda$ and $\theta^{max} < y$, for instance, we have $\lambda_\theta^{SI} = 0$ for all θ, whereas $\lambda_\theta^{AI} = \kappa\ln[y/(y-\theta)] > 0$.

C Other social payoffs

In the model we have used throughout, agents' social payoffs and the norms they underlie are based on image concerns. In the Online Appendix, we generalize our framework and results to social interactions that operate through channels other than reputation, such as reciprocity, a taste for conformity, or conversely a taste for exclusive status.

[40] As with the case of μ_θ previously, one can think of each agent's aversion to violence being independently drawn from an unknown distribution, with the principal having better information about its society-wide mean θ, over which agents experience anticipatory utility.

VII 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 generate crowding out for rare, admirable behaviors but crowding-in for common, merely respectable ones, their optimal level depends nonmonotonically (hump shape) on the private cost of the behavior and the distribution of intrinsic motivations in society. Second, under asymmetric information, expressive concerns lead in a separating equilibrium to weaker incentives when the principal's information involves the general "goodness" of society (more generally, the strength of social norms), and to stronger ones when it concerns the spillovers created by agents' behavior. We also identify settings in which law cannot be expressive, as equilibrium necessarily involves pooling. Finally, our framework allows us to study norm-based interventions, societies' resistance to economists' prescriptions seen as a general "commodification" of human behavior, and their rejection of cruel but cheap punishments.

There are several directions in which our analysis could be interestingly expanded. First, the law was set here by a single principal (government, firm), taking into account how it interacts with and changes the social norm. In practice, interest groups, activists and norm entrepreneurs will compete to change both the social equilibrium and the law, cognizant of their interactions.

Second, we took 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.

In sum, the coevolution of norms, law, and the social meaning of private and public actions offers a vast and promising topic for research.

REFERENCES

Acemoglu, D and M. Jackson (2015) "History, Expectations, and Leadership in the Evolution of Social Norms," *Review of Economic Studies*, 82: 1-34.

Adriani, F. and S. Sonderegger (2019) "A Theory of Esteem Based Peer Pressure," *Games and Economic Behavior*, 115(C): 314–335.

Alfitian, J., Sliwka, D. and T. Vogelsang (2024) "When Bonuses Backfire: Evidence from the Workplace," *Management Science,* 70(9): 6395–6414.

Alger, I., and J. W. Weibull (2013)."Homo Moralis-Preference Evolution Under Incomplete Information and Assortative Matching", *Econometrica* 81(6): 2269-302.

Ali, N. and Bénabou, R. (2020) "Image versus Information: Changing Societal Norms and Optimal Privacy," *American Economic Journal: Micro*, 12(3): 116-164.

Allcott (2011) "Social Norms and Energy Conservation," *Journal of Public Economics,* 95(9-10): 1082-1095.

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

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.

Ashraf, N. and O. Bandiera (2014) "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery," *Journal of Public Economics,* 120: 1-17.

Ayres, I., Raseman, S. and A. Shih (2013) "Evidence From Two Large Field Experiments that Peer Comparison Feedback Can Reduce Energy Usage," *Journal of Law, Economics, and Organization,* 29(5): 992-1022.

Bar-Isaac, H. (2012) "Transparency, Career Concerns, and Incentives for Acquiring Expertise," *The B.E. Journal of Theoretical Economics,* 12(1): Article 4.

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

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. (2013) "Groupthink: Collective Delusions in Organizations and Markets," *Review of Economic Studies,* 80, 429-462.

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.

—— (2011a) "Identity, Morals and Taboos: Beliefs as Assets," *Quarterly Journal of Economics,*

126, 805–855.

—— (2011b) "Laws and Norms" , NBER Working Paper 17579, November.

—— (2016) "Bonus Culture: Competitive Pay, Screening, and Multitasking," *Journal of Political Economy,"* 124(2): 305-370.

Bernheim, D. (1994) "A Theory of Conformity," *Journal of Political Economy*, 102(5), 842–877.

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

Besley, T., Jensen, A. and T. Persson (2023) "Norms , Enforcement, and Tax Evasion," *Review of Economics and Statistics,* 105(4): 998-1007.

Bicchieri, C. and E. Xiao (2009) "Do the Right Thing: But Only if Others Do So," *Journal of Behavioral Decision Making,* 22: 191-208.

Bodner, R. and Prelec, D. (2003) "Self-Signaling and Self-Control," in G. Loewenstein, D. Read, and R. Baumeister (Eds.), *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice,* Russell Sage Foundation: 277-298.

Bohnet, I., Frey, B., Huck, S. (2001) "More Order with Less Law: on Contract Enforcement, Trust and Crowding," *American Political Science 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.

Bowles, S. and S. Polania-Reyes (2012) "Economic Incentives and Social Preferences: Substitutes or Complements?", *Journal of Economic Literature,* 50(2): 368-425.

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. Khokhlova, E., Suvorov, A. and J. van de Ven (2015) "Bad News: An Experimental Study On The Informational Effects Of Rewards," *Review of Economics and Statistics,* 97(1): 55-70.

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

Bursztyn L., Egorov G., and S. Fiorin (2020) "From Extreme to Mainstream: The Erosion of Social Norms", *American Economic Review,* 110(11): 3522-48.

Bursztyn, L., González, A. and D. Yanagizawa-Drott (2020) "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia," *American Economic Review,* 110(10): 2997-3029.

Butera, L., Metcalfe, R., Morrison, W., and D. Taubinsky (2022) "The Deadweight Loss of Social Recognition," *American Economic Review,* 112(1): 122-168.

Cialdini, R. (1984) *Influence, The Psychology of Persuasion*, New York, NY: William Morrow and Company.

Chen, D. (2016) "The Deterrent Effect of the Death Penalty? Evidence from British Commutations During World War I", TSE Working Paper, n. 16-706, revised February 2020.

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

Corneo, G. and O. Jeanne (1997) "Conspicuous Consumption, Snobbism and Conformism," *Journal of Public Economics*, 66(1): 55-71.

Danilov, A. and D. Sliwka (2017) "Can Contracts Signal Social Norms? Experimental Evidence." *Management Science*, 63(2): 459-476.

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.

Diamond, P. (2006) "Optimal Tax Treatment of Private Contributions for Public Goods With and Without Warm Glow Preferences," *Journal of Public Economics*, 90(4-5): 897-919.

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

Fehr, E. 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," *Science* "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. (2011) "Financial Incentives and Student Achievement: Evidence from Randomized Trials," *Quarterly Journal of Economics,* 126(4): 1755-1798.

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.

Funk, P. (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., Henry, E., Jacquemet, N. and M. Lobeck (2021) How Laws Affect The Perception Of Norms: Empirical Evidence From The Lockdown," *Plos One*, 1-14.

Galbiati, R., Schlag K. and J. van der Weele (2013), "Sanctions that Signal: An Experiment," *Journal of Economic Behavior and Organization,* 94: 34-51.

Galbiati, R. 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 (2000) "Pay Enough or Don't Pay At All," *Quarterly Journal of Economics,* 791-810.

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.

Harbaugh, R. and E. Rasmusen (2018) "Coarse Grades: Informing the Public by Withholding Information," *American Economic Journal: Microeconomics*, 10(1): 210–235.

Huck, S. (1997) "Institutions and Preferences: An Evolutionary Perspective," *Journal of Institutional and Theoretical Economics* 153(4): 771-779.

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.

Jia, R. and T. Persson (2021) "Choosing Ethnicity: The Interplay beteen Individual and Social Motives," *Journal of the European Economic Association,* 19(2): 1203-48.

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.

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

Lane, T., Nosenzo, D. and S. Sonderegger (2023) "Laws and Norms: Empirical Evidence," *American Economic Review,* 113(5), 1255-1293.

Mailath, G. (1987) "Incentive Compatiblity in Signaling Games with a Continuum of Types," *Econometrica*, 55: 1349-1365.

Miller, D., and C. McFarland (1987). "Pluralistic Ignorance: When Similarity is Interpreted as Dissimilarity", *Journal of Personality and Social Psychology,* 53(2): 298-305.

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

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.

Prendergast, C. (2007) "The Motivation and Bias of Bureaucrats," *American Economic Review*, 97(1): 180-196.

Prentice, D., and D. Miller (1993). "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm", *Journal of Personality and Social Psychology,* 64(2): 243-256.

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

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

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.

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

Tabellini, G. (2008) "Institutions and Culture," *Journal of the European Economic Association,* 6(2-3): 255-294.

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

Weele, van der, J. (2012) "The Signalling Power of Sanctions in Social Dilemmas," *Journal of Law, Economics and Organization,* 28(1): 103-126.

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

ONLINE APPENDIX

A. Proofs for Section III: Symmetric Information

Example of a firms' objective function equivalent to (10)

Suppose that agents (employees) do not know their social preference toward the firm (enthusiasm for the job, empathy toward colleagues, identification with the mission) when joining it, but learn it before choosing their effort $a \in \{0, 1\}$. The employer offers a compensation package $\{y, y_0\}$ composed of a fixed benefit y_0 and a performance-based wage y; on the latter it incurs a payroll tax at the rate of λ, whereas on the former (for simplicity) it does not -e.g., working conditions or tax-deductible health benefits. The externality ϵ_θ, finally, represents the extra value reaped by the firm when the agent chooses $a = 1$. The firm's profit is then

$$W_\theta^{SI} = E_\theta[[\epsilon_\theta - (1 + \lambda)y]a_\theta(v, y)].$$

The contract must meet agents' participation constraint, yielding a reservation utility U_0:

$$E_\theta[y_0 + (ve_\theta - c_\theta + y)a_\theta(v, y)] \geq U_0$$

Solving for y_0 under equality and substituting into W_θ^{SI} gives back (10), up to a constant.

Properties of the Δ function. Recall that $\Delta(v) \equiv \Delta_0(v)$ is strictly quasiconcave on its support $V \equiv [v^{min}, v^{max}]$, with an interior minimum by Lemma 1.(ii), which is normalized to be reached at $v = 0$. Thus, Δ 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

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

Note also that

$$\Delta'(v) = \frac{f(v)}{1 - F(v)}[E^+(v) - v] - \frac{f(v)}{F(v)}[v - E^-(v)],$$

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

Proof of Proposition 1. Part (i) was established in the text. For (ii), let y_θ^{SI} be the solution to the implicit equation (15); we take here its existence and uniqueness as given, then establish these properties in Proposition 2. Condition (12) implies, given that (11) ensures an interior equilibrium, $\partial W_\theta^{FB}(y_\theta^{SI}) = \epsilon_\theta + v_\theta^*(y^{SI})e_\theta - c_\theta > 0$, hence $y_\theta^{SI} < y_\theta^{FB}$ by strict quasiconcavity of W_θ^{FB}. Next, since $\partial^2 W_\theta^{SI}/\partial y \partial \lambda = -\partial(y[1 - F_\theta(v_\theta^*(y))]/\partial y < 0$ as the threshold is interior, we have $\partial y_\theta^{SI}/\partial \lambda < 0$. ■

Proof of Proposition 2. Part (i) follows directly from Proposition 1(i) and the properties of Δ. For (ii), we focus to avoid repetitions on the case where θ indexes distributional shifts $F(v - \theta)$ with support $V_\theta \equiv [v^{min} + \theta, v^{max} + \theta]$, while ϵ, c, and μ are fixed (the other cases proceed similarly). Let us first express (15) as $H_\theta(y_\theta^{SI}) = 0$, where

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

with all functions in the bracketed term evaluated at $v_\theta^*(y) = \theta + v_0^*(y + \theta e)$.

1. Existence and uniqueness of y_θ^{SI}. Since h_θ is strictly positive and continuously differentiable (C^1) everywhere, so is H_θ, with

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

Since h_θ and Δ_θ have continuous derivatives, $\chi(v)$ 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, $H_\theta(y)$ 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 $H_\theta(y_\theta^{FB}) < 0$ for all θ. We also have $H_\theta(0) > 0$ if

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

or equivalently by (4) and the identity $v_\theta^*(0) - \theta = v_0^*(\theta)$:

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

From (13), $\epsilon - \Delta(v) > 0$ for all $v \in V$, so this expression is bounded on $V = \left[v^{\min}, v^{\max}\right]$. Therefore

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

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

2. Monotonicity of y_θ^{SI}. We focus here on the case $\theta_1 > \theta_0$; the case $\theta_2 < \theta_0$ can be treated symmetrically. By the implicit function theorem,

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

evaluated at $v_\theta^*(y_\theta^{SI})$. We next show that that $\Delta'_\theta(v_\theta^*(y_\theta^{SI}))$ is negative and bounded away from zero on $[\theta_1, \theta_2]$. First, note that

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

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

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

Since $y^{SI}(\theta) < y^{FB}(\theta) < \epsilon$, when $\lambda < \min\{\lambda_1, \lambda_2, \lambda_3\}$, $v_\theta^*(y^{SI})e - \theta < -\varepsilon''$ for all θ in (θ_1, θ_2). Next, we have by (11) $\theta_0 - \theta \equiv (c-\epsilon)/e - \theta > v_{\min} + \varepsilon$, so there exists $\lambda_4 \in (0, \lambda_3)$ such that for all $\lambda < \lambda_4$, $v^*(y_\theta^{SI})e - \theta > v_{\min} + \varepsilon/2$. Denoting $\varepsilon' \equiv \min\{\varepsilon'', \varepsilon/2\}$, and $\eta' \equiv \eta^*(\varepsilon')$, property (A.1) therefore implies that

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

Finally, let us define

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

Thus, for $\lambda < \bar{\lambda} \equiv \min\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$, $(\partial H_\theta / \partial y)(y_\theta^{SI}) < 0$ and (A.6) implies that y_θ^{SI} is strictly decreasing on $[\theta_1, \theta_2]$, so that $y_\theta'^{SI}$ is bounded away from zero (Property P^-). We denote $\theta^{SI}(y)$ its inverse function. ■

B. Proofs for Section IV: Asymmetric Information

Proof of Proposition 3. When y_θ^{FB} is strictly monotonic, the choice of an incentive y reveals that $\theta = (y_\theta^{AI})^{-1}(y) = (y_\theta^{FB})^{-1}(y)$, with no direct announcement necessary. When it is not, announcing θ and setting $y = y_\theta^{FB}$ is incentive compatible, sustaining a separating equilibrium: if the agents believe the planner to be truthful, the latter achieves first-best welfare level W_θ^{FB}, and thus cannot do better by inducing a different cutoff. ■

B.1 Where is (NDB) satisfied?

The no-distortion-at-the-boundary condition imposed in Proposition 4 required that (NDB) be satisfied at θ_1 under M^+ and θ_2 under M^-. It can, in fact, be weakened to only requiring that there be no distortion ($y_\theta^{AI} = y_\theta^{SI}$) at one of the boundaries, since comonotonicity then implies the claimed pattern.

Lemma 2 *Under(NDB) and (COM), the symmetric-information allocation necessarily prevails at θ_1 under M^+ and at θ_2 under M^-.*

Proof. To prove the lemma under M^+ (the proof is the same under M^-), assume a contrario that (NDB) prevails at θ_2 and so $\mathcal{W}_3(\theta_2, \theta_2, y(\theta_2)) = 0$. From (B.11-B.12), note that

$$\mathcal{W}_3 = \frac{(-\partial v^*/\partial y)}{(-\partial v^*/\partial \hat{\theta})} \mathcal{W}_2 - \lambda \big[1 - F_\theta(v^*)\big], \tag{B.1}$$

so $\mathcal{W}_2$ is bounded away from 0 whenever $\mathcal{W}_3 = 0$. Because $y' = -\mathcal{W}_2/\mathcal{W}_3$, $y'(\theta) = +\infty$ or $-\infty$ at $\theta = \theta_2$. Condition (COM) then implies that $y'(\theta) = +\infty$ if $dy_\theta^{SI}/d\theta > 0\,(P^+)$; and $y'(\theta) = -\infty$ if $dy_\theta^{SI}/d\theta < 0$ (P^-). The quasi-concavity of $\mathcal{W}$ in y then implies that $\mathcal{W}_3 > 0$ if P^+ (and $\mathcal{W}_3 < 0$ if P^- in a neighborhood of θ_2).

The (FOC) $\mathcal{W}_2 + y'\mathcal{W}_3 = 0$, together with (COM), yields $\mathcal{W}_2 < 0$ whether P^+ or P^- holds. But (B.1), together with $-\partial v^*/\partial \hat{\theta} > 0$ (that is M^+), shows that $\mathcal{W}_2 > 0$ in a neighborhood of θ_2, a contradiction. ■

B.2 Proof of Propositions 4 and 5

A separating equilibrium is a function y_θ^{AI} satisfying (FOC), (COM), (NDB) and (SOC). We will again prove the results for the case where θ is a shift in the distribution $F(v-\theta)$. The proofs for the other cases ($\epsilon_\theta, c_\theta, \mu_\theta$, etc.) follow similar steps and are omitted to avoid repetition.

For later reference, it will be useful to use the equilibrium condition $\hat{\theta}(y) = (y_\theta^{AI})^{-1}(y)$ and rewrite (19) as a differential equation in y_θ^{AI}:

$$\left(\frac{\epsilon_\theta + v_\theta^*(y_\theta^{AI})e_\theta - c_\theta - \lambda y_\theta^{AI}}{e_\theta + \Delta'_\theta(v_\theta^*(y_\theta^{AI}))}\right)\left(1 + \frac{v_\theta^*(y_\theta^{AI})\gamma\frac{\partial \epsilon_\theta}{\partial \theta} - \frac{\partial c_\theta}{\partial \theta} + \frac{\partial \Delta_\theta}{\partial \theta}(v_\theta^*(y_\theta^{AI}))}{dy_\theta^{AI}/d\theta}\right) = \frac{\lambda}{h_\theta(v_\theta^*(y_\theta^{AI}))}. \tag{B.2}$$

Let us now fix $(\theta_1, \theta_2) \subset \Theta$, with $\theta_0 < \theta_1$; 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) $y^{SI} : [\theta_1, \theta_2] \to [y_{\theta_2}^{SI}, y_{\theta_1}^{SI}]$ with $y_\theta^{SI} \in (0, y_\theta^{FB})$ that solves the full-information problem, $H_\theta(y_\theta^{SI}) = 0$.

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 (19), which we rewrite as

$$IVP(\lambda): \quad \hat{\theta}'(y) = \Psi_{\hat{\theta}(y)}(y) \tag{B.3}$$

where

$$\Psi_\theta(y) = \frac{H_\theta(y)}{\Delta'_\theta(v_\theta^*(y))[\epsilon + v_\theta^*(y)e - c - \lambda y]}, \tag{B.4}$$

and $H_\theta(y)$ is still given by (A.2), given that we continue to focus on the case where θ indexes a shift in the distribution $F(v - \theta)$.

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_θ^{AI}, 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 $\Psi_\theta(y)$ and its partial derivatives are well-defined and continuous at every point where the denominator terms, $\epsilon + v_\theta^*(y) - c - \lambda y$ and $\Delta'(v_\theta^*(y) - \theta)$, are non-zero. In particular, at $(y_1, \theta_1) = (y_{\theta_1}^{SI}, \theta_1)$ we have $\epsilon + v_{\theta_1}^*(y) - c - \lambda y_1 > 0$ due to (16). Moreover, (A.9) implies that, at $\theta = \theta_1$, $\Delta'_{\theta_1}(v_\theta^*(y_1)) - \theta_1) < -\eta' < 0$. Therefore, $\Psi_\theta(y)$ 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

$$0 < \varepsilon' < \min\left\{\varepsilon,\ \theta_1 - v_{\theta_1}^*(y_1),\ \theta_1 - \theta_0\right\} \tag{B.5}$$

and define

$$\lambda^* \equiv \min\left\{\frac{h(-\varepsilon')\left(\theta_1 - \theta_0 - \varepsilon'\right)}{1 + y_{\theta_1}^{FB} h(-\varepsilon')}, \bar{\lambda}\right\} > 0. \tag{B.6}$$

Lemma 3 *For all $\lambda < \lambda^*$, the function $\hat{\theta}$ has the following properties on its support:*
(i) $b(y) \equiv \epsilon + \hat{v}(y)e - c - \lambda y$ 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

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

Therefore, $b'(y) < 0$ wherever $b(y) > 0$. Since $b(y_1) > 0$ by (16), 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:

$$\begin{aligned}\frac{d[\hat{v}(y) - \hat{\theta}(y)]}{dy} &= \frac{d(v_0^*(y + \hat{\theta}(y)e))}{dy} = -\frac{1 + \hat{\theta}'(y)}{e + \Delta'(\hat{v}(y) - \hat{\theta}(y))} \\ &= \frac{-1}{\Delta'(\hat{v}(y) - \hat{\theta}(y))}\left[1 - \frac{\lambda}{h(\hat{v}(y)) - \hat{\theta}(y))[\epsilon + \hat{v}(y)e - c - \lambda y]}\right], \qquad \text{(B.7)}\end{aligned}$$

by (19) with $\epsilon_\theta = \epsilon, c_\theta = c$, etc. 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,

$$\begin{aligned}& h(v^*_{\hat{\theta}(y')}(y')) - \hat{\theta}(y'))[\epsilon - c - \lambda y' + v^*_{\hat{\theta}(y')}(y')e] \\ = \ & h(-\varepsilon')(-\theta_0 - \lambda y' + \hat{\theta}(y') - \varepsilon') > h(-\varepsilon')\left(\theta_1 - \theta_0 - \varepsilon - \lambda y'\right) \\ > \ & h(-\varepsilon')\left(\theta_1 - \theta_0 - \varepsilon' - \lambda y_1\right) > h(-\varepsilon')\left(\theta_1 - \theta_0 - \varepsilon' - \lambda y^{FB}(\theta_1)\right) > \lambda. \qquad \text{(B.8)}\end{aligned}$$

The bracketed term in (B.7) 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 (B.5), there must exists some $y'' \in (y', y_1)$ where it equals $-\varepsilon'$ again, a contradiction. ■

Lemma 4 *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 showed in Lemma 3 that

$$\epsilon + \hat{v}(y)e - c - \lambda y > 0 > \Delta'_{\hat{\theta}(y)}(\hat{v}(y)). \qquad \text{(B.9)}$$

Equation (19) therefore implies that $\hat{\theta}'(y) \leq 0$ if and only if

$$\frac{\epsilon + \hat{v}(y) - c - \lambda}{e + \Delta'_{\hat{\theta}(y)}(\hat{v}(y))} \geq \frac{\lambda}{h_{\hat{\theta}(y)}(\hat{v}(y))}, \qquad \text{(B.10)}$$

which by (B.2) means that $\partial W^{AI}_{\hat{\theta}}(y)/\partial y \geq 0$ at y. By strict quasiconcavity of $W^{AI}_{\hat{\theta}}(y) = W^{FI}_{\hat{\theta}(y)}(y)$, 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 $\hat{v}(y) - \hat{\theta}(y) = v^*(y + \hat{\theta}(y)e) > v^*(y + \theta^{AI}(y)e) > v^*(y + \theta^{FB}(y)e) = (c-e)/e > v_{\min} + \varepsilon$, 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 (11) together with $\varepsilon' < \varepsilon$. In Lemma 3 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 3 and 4 imply that Ψ_θ 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 (B.4) 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 4(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. ■

B.3 Characterization and existence of a separating allocation satisfying (FOC), (SOC), (NDB) and (COM)

B.3.1 Preliminaries

We collect here a number of assumptions made and properties shown, in Sections II and III, concerning the symmetric-information case, which will be used to study the general asymmetric-information case.

Assumption 1 ***(symmetric information)***

The support of the distribution of principal's type is an interval $\Theta = [\theta_1, \theta_2]$ *such that:*

(i) The distribution F_θ *has a large enough support so that cutoffs are interior: there exists* $\eta > 0$ *such that for all* $\theta \in \Theta$, $F_\theta(v^*_\theta(y^{SI}_\theta)) \in [\eta, 1-\eta]$.

(ii) For all $\theta \in \Theta$, *the density* f_θ, *the responsiveness of the cutoff* v^*_θ *to the material incentive,* $(-\frac{\partial v^*_\theta}{\partial y})$, *and to the principal's type,* $|\frac{\partial v^*_\theta}{\partial \theta}|$, *are bounded away from 0.*

(iii) The slope of the schedule y^{SI}_θ *is bounded away from 0: there exists* $\varepsilon > 0$ *such that for all* $\theta \in \Theta$, $|\frac{dy^{SI}_\theta}{d\theta}| > \varepsilon$.

Discussion. (i) The conditions on the distribution F_θ and its density f_θ are rather weak, and were already identified in Section II. (ii) The results in Sections II and III guarantee that the responsiveness of the cutoff to the material incentive and to the principal's type are always bounded away from 0, except perhaps when θ indexes a shift in the distribution F_θ, in which case $(\partial \Delta_\theta / \partial \theta)(v) = \Delta'(v - \theta)$ changes sign at $v = \theta$; see equations (6) and (8). That is why in Proposition 2, we focused on an interval $\Theta = [\theta_1, \theta_2]$ excluding θ_0, over which either a (strict)

norm or anti-norm prevails under symmetric information. (iii) As shown by (17), the strong monotonicity (slope bounded away from zero, see Definition 2) of the first-best schedule y_θ^{FB} is also ensured on such an interval, and as shown in the proof of Lemma 4 and Proposition 5 (for the case of a uniform shift, but a similar proof applies to the other cases), it extends to the symmetric and asymmetric information ones, y_θ^{SI} and y_θ^{AI}, for λ small enough, which we assume throughout.

Turning now to asymmetric information, let $\mathcal{W}(\theta, \hat{\theta}, y)$ denote the payoff of a type-θ principal when offering incentive y and being perceived as type $\hat{\theta}$, $b(\theta, \hat{\theta}, y)$ the corresponding marginal contribution to social welfare, and $h(\theta, \hat{\theta}, y)$ the hazard rate of the distribution at the cutoff:

$$\begin{aligned}
\mathcal{W}(\theta, \hat{\theta}, y) &\equiv \int_{v_{\hat{\theta}}^*(y)}^{+\infty} [ve_\theta + \epsilon_\theta - c_\theta - \lambda y] f_\theta(v) dv \\
b(\theta, \hat{\theta}, y) &= v_{\hat{\theta}}^*(y) e_\theta + \epsilon_\theta - c_\theta - \lambda y \\
h_\theta(\theta, \hat{\theta}, y) &\equiv \frac{f_\theta(v_{\hat{\theta}}^*(y))}{1 - F_\theta(v_{\hat{\theta}}^*(y))}
\end{aligned}$$

We noted that $\mathcal{W}(\theta, \theta, y)$ is strictly quasi-concave in its third argument for λ not too large; the same property applies to $\mathcal{W}(\theta, \hat{\theta}, y)$ in the asymmetric-information setup. In a separating equilibrium $\{y_\theta^{AI}\}$, the principal's payoff is $\mathcal{W}_\theta^{AI} \equiv \mathcal{W}(\theta, \theta, y_\theta^{AI})$, the social value of a marginal contribution is $b_\theta^{AI} \equiv b(\theta, \theta, y_\theta^{AI})$, and the hazard rate is $h_\theta(y_\theta^{AI}) \equiv h(\theta, \theta, y_\theta^{AI})$. We will denote as $\mathcal{W}_1(\theta, \hat{\theta}, y) \equiv \frac{\partial \mathcal{W}(\theta, \hat{\theta}, y)}{\partial \theta}$, etc, the partial derivatives, and in particular:

$$\mathcal{W}_2(\theta, \hat{\theta}, y(\hat{\theta})) = \left(-\frac{\partial v_{\hat{\theta}}^*(y(\hat{\theta}))}{\partial \hat{\theta}} \right) b(\theta, \hat{\theta}, y(\hat{\theta})) f_\theta(v_{\hat{\theta}}^*(y(\hat{\theta}))), \tag{B.11}$$

$$\mathcal{W}_3(\theta, \hat{\theta}, y(\hat{\theta})) = \left(-\frac{\partial v_{\hat{\theta}}^*(y(\hat{\theta}))}{\partial y} \right) b(\theta, \hat{\theta}, y(\hat{\theta})) f_\theta(v_{\hat{\theta}}^*(y(\hat{\theta}))) - \lambda [1 - F_\theta(v_{\hat{\theta}}^*(y(\hat{\theta})))]. \tag{B.12}$$

The first-order condition (B.2) can then be rewritten as

$$\frac{dy^{AI}}{d\theta} = -\frac{\mathcal{W}_2(\theta, \theta, y_\theta^{AI})}{\mathcal{W}_3(\theta, \theta, y_\theta^{AI})}. \tag{B.13}$$

This is intuitive: restricting here attention to the equilibrium graph, one can indifferently think of the principal choosing y to maximize $\mathcal{W}(\theta, \hat{\theta}(y), y))$, or choosing the induced belief $\hat{\theta}$ to maximize $\mathcal{W}(\theta, \hat{\theta}, y(\hat{\theta}))$.

B.3.2 Comparing the SI and AI allocations

Let us now show that a separating equilibrium must follow one of the four patterns depicted in Figure 3. We do so in the case in which (NDB) holds at θ_1 (that is, under M^+). The same proof can be used when (NDB) holds at θ_2 (that is, under M^-). When (NDB) holds at θ_1, we show that the schedule $\{y(\theta) \equiv y_\theta^{AI}\}$ lies everywhere above $\{y_\theta^{SI}\}$ when P^+ obtains, or everywhere below $\{y_\theta^{SI}\}$ when P^- holds, provided that the asymmetric-information schedule $\{y_\theta^{AI}\}$ inherits from the symmetric-information schedule $\{y_\theta^{SI}\}$ the property that its slope is bounded away from 0 (Assumption 1(i)). Lemma 2 then guarantees that the slope of $\{y_\theta^{AI}\}$ is indeed bounded

away from 0 if λ is not too high. The method of proof is then similar to that used for Lemma 4. It will be useful to rewrite (FOC), namely equation (B.2), as:

$$\mathcal{W}_2 = \frac{\lambda(1-F_\theta)}{1/(\frac{dy_\theta^{AI}}{d\theta}) + (-\frac{\partial v_{\hat\theta}^*}{\partial y})/(-\frac{\partial v_{\hat\theta}^*}{\partial \hat\theta})}. \tag{B.14}$$

(a) Case P^+. Suppose that P^+ obtains and that (NDB) holds at θ_1. (FOC), as expressed in equation (B.14), together with (COM) (i.e., $dy_\theta^{AI}/d\theta > 0$), M^+ (i.e., $-\partial v_{\hat\theta}^*/\partial\hat\theta > 0$), and Assumption 1(iii), implies that $\mathcal{W}_2 > 0$ everywhere and that $\mathcal{W}_2$ is bounded away from 0. At θ_1, $\mathcal{W}_3 = 0$ from (NDB). The slope of y_θ^{AI}, $-\mathcal{W}_2/\mathcal{W}_3$, taken at θ_1 is therefore $+\infty$. This implies that to the right of θ_1, $y_\theta^{AI} > y_\theta^{SI}$ and, from the strict quasi-concavity of $\mathcal{W}$ in its third argument, that $\mathcal{W}_3 < 0$. To show that $y_\theta^{AI} > y_\theta^{SI}$ for all $\theta > \theta_1$, consider the first θ^* such that $y_{\theta^*}^{AI} = y_{\theta^*}^{SI}$. At θ^*, $\mathcal{W}_3 = 0$ and $dy_\theta^{AI}/d\theta = -\mathcal{W}_2/\mathcal{W}_3 = +\infty$ in a left neighborhood of θ^*, provided that the asymmetric-information schedule $\{y_\theta^{AI}\}$'s slope is bounded away from 0 (which it is as we have seen that $\mathcal{W}_2$ is bounded away from 0). Thus the two curves cannot cross.

(b) Case P^-. The proof for P^- is similar, with a nuance: That $\mathcal{W}_2 > 0$ still holds, but does not result directly from (FOC) and (COM) as $\frac{dy_\theta^{AI}}{d\theta} < 0$. However, whenever for some θ^* (equal to, or above θ_1), $\mathcal{W}_3(\theta^*) = 0$, as $\theta \to \theta^*$, $|\ \frac{dy_\theta^{AI}}{d\theta}\ |$ bounded away from 0 implies that $\mathcal{W}_2$ also is and thus $|\ \frac{dy_\theta^{AI}}{d\theta}\ | = |\ -\mathcal{W}_2/\mathcal{W}_3\ | \to +\infty$. The rest of the proof follows the previous lines. $\mathcal{W}_3 = 0$ at θ_1 from (NDB), and $\mathcal{W}_3 > 0$ to the right of θ_1. $\mathcal{W}_2 > 0$ everywhere and $\mathcal{W}_2$ is bounded away from 0. The slope of y_θ^{AI} at θ_1 is therefore $-\infty$. Finally, because the slope of y_θ^{AI} would be equal to $-\infty$ if the two curves crossed at some $\theta^* > \theta_1$, necessarily $y_\theta^{AI} < y_\theta^{SI}$ for all $\theta > \theta_1$.

B.3.3 Proof of Proposition 6 (second-order conditions/SOC)

(a) Mimicking another type (deviations on the equilibrium graph $\mathcal{G}$)

When is a candidate *separating* equilibrium $\{y(\theta) \equiv y_\theta^{AI}\}$ satisfying (FOC), (COM) and (NDB) globally incentive compatible? We will first focus on the graph of the separating equilibrium under investigation: $\mathcal{G} \equiv \{y(\theta)\}_{\theta\in[\theta_1,\theta_2]}$. The following lines build on the proof of Theorem 3 in Mailath (1987), but must account for the complication that in contrast to Mailath's analysis, $\mathcal{W}_2$ need not have a constant sign.

A *sufficient* condition for the allocation to be incentive compatible on graph $\mathcal{G}$ (that is, no deviation by type θ pretending to be type $\hat\theta \neq \theta$ is profitable) is that $\frac{d\mathcal{W}(\theta,\hat\theta,y(\hat\theta))}{d\hat\theta}$ be weakly positive for $\hat\theta < \theta$ and weakly negative for $\hat\theta > \theta$. From the (FOC), i.e. the differential equation (B.2), we have $y'(\hat\theta) = -\frac{\mathcal{W}_2(\hat\theta,\hat\theta,y(\hat\theta))}{\mathcal{W}_3(\hat\theta,\hat\theta,y(\hat\theta))})$ for all $\hat\theta$, while

$$\frac{d\mathcal{W}(\theta,\hat\theta,y(\hat\theta))}{d\hat\theta} = \mathcal{W}_2(\theta,\hat\theta,y(\hat\theta)) - \mathcal{W}_3(\theta,\hat\theta,y(\hat\theta))\frac{\mathcal{W}_2(\hat\theta,\hat\theta,y(\hat\theta))}{\mathcal{W}_3(\hat\theta,\hat\theta,y(\hat\theta))}$$

or

$$\frac{d\mathcal{W}(\theta,\hat\theta,y(\hat\theta))}{d\hat\theta} = \mathcal{W}_2(\theta,\hat\theta,y(\hat\theta))\left(1 - \frac{\mathcal{W}_3(\theta,\hat\theta,y(\hat\theta))}{\mathcal{W}_2(\theta,\hat\theta,y(\hat\theta))}/\frac{\mathcal{W}_3(\hat\theta,\hat\theta,y(\hat\theta))}{\mathcal{W}_2(\hat\theta,\hat\theta,y(\hat\theta))}\right).$$

Since $\frac{d\mathcal{W}(\theta,\hat{\theta},y(\hat{\theta}))}{d\hat{\theta}}\,|_{\hat{\theta}=\theta}=0$, the desired condition is equivalent to

$$\mathcal{A} \equiv \mathcal{W}_2(\theta,\hat{\theta},y(\hat{\theta}))y'(\hat{\theta})\frac{\partial}{\partial\theta}\left(\frac{\mathcal{W}_3(\theta,\hat{\theta},y(\hat{\theta}))}{\mathcal{W}_2(\theta,\hat{\theta},y(\hat{\theta}))}\right)\geq 0. \tag{B.15}$$

When it holds for all $(\theta,\hat{\theta})$, this is a sufficient condition for type θ to prefer setting $y(\theta)$ to mimicking any other type $\theta'\in[\theta_1,\theta_2]$ by setting $y(\theta')$. Taken at $\hat{\theta}=\theta$, the condition is necessary and sufficient for a local maximum.

To simplify the condition, use (B.11)-(B.12) to compute $\mathcal{W}_2/\mathcal{W}_3$ and note that, since $-\frac{\partial v^*_{\hat{\theta}}}{\partial\hat{\theta}}$ and $-\frac{\partial v^*_{\hat{\theta}}}{\partial y}>0$ do not depend on θ, we have

$$sgn\left(\frac{\partial}{\partial\theta}\left(\frac{\mathcal{W}_3(\theta,\hat{\theta},y(\hat{\theta}))}{\mathcal{W}_2(\theta,\hat{\theta},y(\hat{\theta}))}\right)\right)=sgn\left(-\frac{\partial v^*_{\hat{\theta}}}{\partial\hat{\theta}}\right)sgn\left(\frac{\partial}{\partial\theta}\left(b(\theta,\hat{\theta},y(\hat{\theta}))h_\theta(v^*_{\hat{\theta}}(y(\hat{\theta})))\right)\right)$$

where $h_\theta=f_\theta/[1-F_\theta]$ is the hazard rate. Using again (B.11), we obtain:

$$\mathcal{A}=y'b\frac{\partial(bh_\theta)}{\partial\theta}\geq 0$$

for all $(\theta,\hat{\theta})$ as a sufficient condition for on-path incentive compatibility. It is important to note that this condition is sufficient, but not necessary.

(b) Deviations outside the equilibrium graph

Next, we need to consider deviations to incentives y that lie outside the equilibrium graph $\mathcal{G}$. In the standard model considered by Mailath, off-path beliefs are trivially selected as "adverse beliefs" (say, $\hat{\theta}$ is then the lowest θ in the support of the distribution), which are the worst possible perception regardless of the sender's type. This cannot be done here as the function $\mathcal{W}_2$ in general changes sign. Under P^+ for example, the principal might want to induce a perception $\hat{\theta}>\theta$ in order to enlist more contributions and thereby make up for the costly material incentives; if $\hat{\theta}$ is much higher than θ, however, the principal may "overshoot" and induce too much prosocial behavior given the low stake: the perception of high type is beneficial locally, but may be worse than being perceived as a low type if that perception is too inflated.

We will therefore provide an alternative proof method to rule out deviations that are not on the equilibrium path, based on the following strategy: we will presume monotonic beliefs, both on-path (as required by equilibrium behavior) and off-path. Namely when $y'(\theta)>0$, we will posit off-path beliefs: $\hat{\theta}=\theta_1$ for $y<y(\theta_1)$ and $\hat{\theta}=\theta_2$ for $y>y(\theta_2)$; and symmetrically when $y'(\theta)<0$. We will then check whether the single-crossing condition

$$\mathcal{B}\equiv y'(\hat{\theta})\frac{\partial^2\mathcal{W}(\theta,\hat{\theta},y)}{\partial\theta\partial y}\geq 0, \tag{B.16}$$

also holds, at least at $\hat{\theta}=\theta_1$ and $\hat{\theta}=\theta_2$, which is what (SOC_2) requires. This implies, for example when $y'(\theta)>0$, that reducing y below $y(\theta_1)$ is more costly for an arbitrary type θ than for type θ_1. To show that (B.16) rules out deviations outside the equilibrium path, let us focus on this case (the one where both terms in $\mathcal{B}$ are negative is treated symmetrically). Specify beliefs $\hat{\theta}=\theta_1$ for $y<y_1$ and $\hat{\theta}=\theta_2$ for $y>y_2$, and note that:

(i) From equilibrium behavior when policies are restricted to lie in $[\theta_1, \theta_2]$, type θ prefers (at least weakly) selecting $y(\theta)$ rather than y_1:

$$\mathcal{W}(\theta, \theta, y(\theta)) - \mathcal{W}(\theta, \theta_1, y_1) \geq 0.$$

(ii) Since $y_1 = y_{\theta_1}^{SI}$ (NDB condition), type θ_1 prefers y_1 to any $y < y_1$:

$$\mathcal{W}(\theta_1, \theta_1, y_1) - \mathcal{W}(\theta_1, \theta_1, y) \geq 0.$$

(iii) Because $\partial^2 \mathcal{W}(\theta, \hat{\theta}, y)/\partial\theta\partial y \geq 0$, and in particular $\partial^2 \mathcal{W}(\theta, \theta_1, y)/\partial\theta\partial y \geq 0$, which is all that is needed, a fortiori so does type θ:

$$\mathcal{W}(\theta, \theta_1, y_1) - \mathcal{W}(\theta, \theta_1, y) \geq \mathcal{W}(\theta_1, \theta_1, y_1) - \mathcal{W}(\theta_1, \theta_1, y).$$

Summing the three inequalities yields $\mathcal{W}(\theta, \theta, y(\theta)) \geq \mathcal{W}(\theta, \theta_1, y)$, meaning that deviating to y is not profitable for type θ.

The case for $y > y_2$ proceeds similarly, except that in the second step $\mathcal{W}(\theta_2, \theta_2, y_2) - \mathcal{W}(\theta_2, \theta_2, y) \geq 0$ follows from strict quasi-concavity of the principal's objective function under symmetric information, given that $y > y_2 = y_2^{AI} > y_2^{SI}$.

To check (B.16) in the various applications, we make the cross-derivative within it explicit: from (B.12), we have

$$\frac{\partial^2 \mathcal{W}(\theta, \hat{\theta}, y)}{\partial\theta\partial y} = \left(-\frac{\partial v_{\hat{\theta}}^*(y)}{\partial y}\right)\left(\frac{\partial b}{\partial\theta}(\theta, \hat{\theta}, y) f_\theta(v_{\hat{\theta}}^*(y)) + b(\theta, \hat{\theta}, y)\frac{\partial f_\theta(v_{\hat{\theta}}^*(y))}{\partial\theta}\right) + \lambda\frac{\partial F_\theta(v_{\hat{\theta}}^*(y))}{\partial\theta}. \tag{B.17}$$

B.4 Applications: separating equilibria

Proof of Proposition 7. For each case in which θ indexes a shift or truncation in the distribution $F_\theta(v)$, the intensity of social monitoring μ_θ, the externality ϵ_θ, or the contribution cost c_θ, we will now examine: (i) whether the condition $\mathcal{A} \geq 0$ (SOC_1) holds, either everywhere or over a range to which the search for a global optimum to the principal's problem can be restricted; (ii) if it does, whether the condition $\mathcal{B} \geq 0$ (SOC_2) also holds, either everywhere or over a range to which attention can be restricted, given the off-path beliefs specified in Part (b) of the proof of Proposition 6. The two conditions together ensure the existence of a separating equilibrium, and the results in Section B.3.2 then determine whether it involves weak or strong law ($y_\theta^{AI} \lesseqgtr y_\theta^{SI}$). When $\mathcal{A} < 0$, the equilibrium must involve some pooling, and when $\mathcal{A} \equiv 0$ everywhere pooling is also the natural outcome, as analyzed in Section B.5 below.

Recalling that

$$b(\theta, \hat{\theta}, y(\hat{\theta})) = v_{\hat{\theta}}^*(y(\hat{\theta}))e_\theta + \epsilon_\theta - c_\theta - \lambda y(\hat{\theta}),$$

which will be abbreviated as b when no confusion results, we can rewrite the (FOC) as:

$$b(\hat{\theta}, \hat{\theta}, y(\hat{\theta}))\left(-\frac{\partial v_{\hat{\theta}}^*}{\partial\hat{\theta}}\frac{1}{y'(\hat{\theta})} - \frac{\partial v_{\hat{\theta}}^*}{\partial y}\right) = \frac{\lambda}{h_\theta(v_{\hat{\theta}}^*(y(\hat{\theta})))} \tag{B.18}$$

Note also that, on-path, $b(\theta, \theta, y^{AI}(\theta)) > 0$ for all $\theta \in [\theta_1, \theta_2]$ when $\lambda > 0$; this is obvious from (B.18) in the cases (M^+, P^+) and (M^-, P^-), and was shown directly in Lemma 3 for the case

where θ indexes $F(v-\theta)$. For $\lambda = 0$ (first best), $b(\theta, \theta, y(\theta)) = 0$, but $b(\theta, \hat{\theta}, y(\theta)) \neq 0$ for all $\hat{\theta} \neq \theta$.

1. Shift in image concerns

Let the parameter unknown to the agents affect the image incentive Δ_θ: θ indexes a uniform shift or a truncation parameter of the distribution F_θ, or else the intensity μ_θ of image concerns (see Section II). In all these cases, b is independent of θ, so

$$\mathcal{A} = y' b^2 \frac{\partial h_\theta}{\partial \theta}.$$

(a) Norm ($\Delta'_\theta < 0$). Since $\frac{dy_\theta^{AI}}{d\theta} < 0$ in this case, and $\frac{\partial h_\theta}{\partial \theta} < 0$ from the monotone-hazard-rate property , we have $\mathcal{A} \geq 0$, with strict inequality at $\hat{\theta} = \theta$ when $\lambda > 0$, and at all $\hat{\theta} \neq \theta$ when $\lambda = 0$. Thus (SOC_1) is satisfied everywhere and type θ strictly prefers setting $y(\theta)$ to mimicking any other type $\hat{\theta} \in [\theta_1, \theta_2]$.

Turning next to (SOC_2) and off-path deviations, since b does not depend on θ and $F_\theta(v) = F(v-\theta)$, we have:

$$\frac{\partial^2 \mathcal{W}(\theta, \hat{\theta}, y)}{\partial \theta \partial y} = \left(-\frac{\partial v^*_{\hat{\theta}}(y)}{\partial y} \right) b(\hat{\theta}, y)(-f'_\theta) - \lambda f_\theta. \tag{B.19}$$

Since $y' < 0$, $f'_\theta \geq 0$ is thus a sufficient condition for $\mathcal{B}(\theta, \hat{\theta}, y) \geq 0$, provided that $b \geq 0$. We saw that $b(\theta, \theta, y_\theta^{AI}) > 0$ for all $\theta \in [\theta_1, \theta_2]$ when $\lambda > 0$, and $b(\theta, \theta, y_\theta^{FB}) = 0$ when $\lambda = 0$, but it could be that $b(\theta, \hat{\theta}, y) < 0$ for some values of y off the equilibrium path and associated belief $\hat{\theta}$. To deal with this possibility, let us fix θ and define the functions

$$b_k(y) \equiv v^*_{\theta_k}(y)e + \epsilon - c - \lambda y, \text{for } k = 1, 2,$$

noting that both are always strictly decreasing.

Consider first a potential off-path deviation to $y < y_2$, leading to belief $\hat{\theta} = \theta_2$. Since $b_2(y_2) \geq 0$ by the (FOC) and $b'_2(y) < 0$, we have $b_2(y) > 0$ for any $y < y_2$, and therefore $\frac{\partial^2 \mathcal{W}(\theta, \theta_2, y)}{\partial y \partial \theta} < 0$ over that range: type θ suffers more than type θ_2 from a reduction of y below y_2, fixing belief at θ_2. But type θ_2 prefers y_2 to y (by strict concavity of $\mathcal{W}(\theta_2, \theta_2, y)$) and type θ prefers $y(\theta)$ to y_2: the reasoning in part (b) of the proof of Proposition 6 applies here unchanged, ruling out type θ deviating to $y < y_2$.

Consider next a potential off-path deviation to $y > y_1$, leading to belief $\hat{\theta} = \theta_1$. Given that $b_1(y_1) \geq 0$ (with strict inequality unless $\lambda = 0$ and $\hat{\theta} = \theta$) by the (FOC), and $b'_1(y) < 0$, either $b_1(y) > 0$ for all $y \geq y_1$ (which can only occur when $\lambda = 0$), or else there is a unique point $\bar{y} \geq y_1$ such that $b_1(y)$ goes from positive to negative at $\bar{y}$. In the first case, let $\bar{y} \equiv +\infty$, so that it is subsumed in the second case. Because $b_1(y) > 0$ on $(y_1, \bar{y})$, we have $\partial^2 \mathcal{W}(\theta, \theta_1, y)/\partial \theta \partial y < 0$ on that interval, and by the same reasoning as before this implies that $\mathcal{W}(\theta, \theta, y(\theta)) > \mathcal{W}(\theta, \theta_1, y)$ for any $y \in (y_1, \bar{y})$.

Next, for any $y \geq \bar{y}$, we can write:

$$\begin{aligned}
\mathcal{W}(\theta, \theta_1, y) &= \int_{v^*_{\theta_1}(y)}^{v^*_{\theta_1}(\bar{y})} (ve + \epsilon - c - \lambda y) f_\theta(v)\, dv + \int_{v^*_{\theta_1}(\bar{y})}^{+\infty} (ve + \epsilon - c - \lambda y) f_\theta(v)\, dv \\
&\leq \int_{v^*_{\theta_1}(y)}^{v^*_{\theta_1}(\bar{y})} (v^*_{\theta_1}(\bar{y})e + \epsilon - c - \lambda \bar{y}) f_\theta(v)\, dv + \int_{v^*_{\theta_1}(\bar{y})}^{+\infty} (ve + \epsilon - c - \lambda \bar{y}) f_\theta(v)\, dv,
\end{aligned}$$

with strict inequality when $\lambda > 0$. The first term is zero by definition, while the second one equals $\mathcal{W}(\theta, \theta_1, \bar{y}) \leq \mathcal{W}(\theta, \theta, y(\theta))$, hence again deviating from $y(\theta)$ to y is unprofitable.

(b) Anti-norm ($\Delta'_\theta > 0$.) Under (COM), $\frac{dy^{AI}_\theta}{d\theta} > 0$ if an anti-norm prevails ($\Delta'_\theta > 0$). Thus $\mathcal{A}(\theta, \hat{\theta}, y(\hat{\theta})) \leq 0$, with strict inequality at $\hat{\theta} = \theta$ when $\lambda > 0$, by (B.18), and at all $\hat{\theta} \neq \theta$ when $\lambda = 0$. The global second-order condition is not satisfied even locally, and so there is no separating equilibrium.

(c) Right-truncation: Truncate the cumulative distribution so that only types $v \leq v^{\max} - \theta$ remain: $F_\theta(v) = F(v)/F(v^{\max} - \theta)$ on $[v^{\min}, v^{\max} - \theta]$. A truncation corresponds to an increase in θ, and so to a decrease in glory (but not in shame, for a given cutoff v^*) and thereby in image incentives. The hazard rate $h_\theta = f_\theta(v^*)/[1 - F_\theta(v^*)]$ is increasing in θ, and (COM) implies that $dy^{AI}_\theta/d\theta \geq 0$ (the reduced glory requires more extrinsic motivation). Thus, we again have $\mathcal{A} \geq 0$ (SOC_1), with strict inequality at $\hat{\theta} = \theta$ when $\lambda > 0$, and at all $\hat{\theta} \neq \theta$ when $\lambda = 0$. Next,

$$\mathcal{W}(\theta, \hat{\theta}, y) = \frac{\mathcal{W}(0, \hat{\theta}, y)}{F(v^{\max} - \theta)} = \frac{F(v^{\max} - \theta_2)}{F(v^{\max} - \theta)} \mathcal{W}(\theta_2, \hat{\theta}, y),$$

Consequently,

$$sgn(\mathcal{B}(\theta, \theta_2)) = sgn\left(\frac{\partial \mathcal{W}}{\partial y}(\theta_2, \theta_2, y(\theta_2)\right),$$

which is positive since the strictly quasiconcave function $\mathcal{W}(\theta_2, \theta_2, y)$ is maximized at $y_2 = y^{SI}_{\theta_2} > y^{AI}_{\theta_2} = y(\theta_2)$. Thus, ($\text{SOC}_2$) is also satisfied. A similar proof applies for $\mathcal{B}(\theta, \theta_1)$.

(d) Left-truncation and social vigilance. Two environments share the knife-edge property that $\mathcal{A} \equiv 0$ for all $\{\theta, \hat{\theta}\}$, as well as an invariance (up to an affine transformation) of the principal's objective function with respect to θ, as she cares only about agents' perception $\hat{\theta}$. an increase in θ weakens image incentives but leaves both the benefit b and the hazard rate $h_\theta = h_0$ unchanged. Similarly for social vigilance μ_θ, $\partial(bh_\theta)/\partial\theta = 0$. In both cases there is no sorting condition, and we will see that the most natural equilibrium often involves full pooling.

2. Shift in externality or cost. In both these cases h is independent of θ, so

$$\mathcal{A} = y' b \frac{\partial b}{\partial \theta} h.$$

(a) Externality: When θ affects ϵ_θ, (COM) implies that $\frac{dy^{AI}_\theta}{d\theta} > 0$ and $\frac{\partial b}{\partial \theta} > 0$. Thus $sgn(\mathcal{A}) = sgn(b)$. As noted above, we have $b(\hat{\theta}, \hat{\theta}, y(\hat{\theta})) \geq 0$, with strict inequality unless $\lambda = 0$. Therefore, locally, i.e. for $\hat{\theta}$ near θ, (SOC_1) is satisfied. To show global on-path optimality, note first that type θ would never gain from mimicking a type $\hat{\theta}$ such that $b(\theta, \hat{\theta}, y(\hat{\theta})) \leq 0$. Indeed,

$$\frac{d\mathcal{W}(\theta, \hat{\theta}, y(\hat{\theta}))}{d\hat{\theta}} = \left(-\frac{\partial v^*_{\hat{\theta}}}{\partial \hat{\theta}}\right) b f_\theta + \left(-\frac{\partial v^*_{\hat{\theta}}}{\partial y}\right) b f_\theta y'(\hat{\theta}) - \lambda(1 - F_\theta) y'(\hat{\theta}),$$

so if $b(\theta, \hat{\theta}, y(\hat{\theta})) \leq 0$, then $\frac{d\mathcal{W}(\theta, \hat{\theta}, y(\hat{\theta}))}{d\hat{\theta}} \leq 0$, with strict inequality when $\lambda > 0$. Intuitively, reducing y slightly would reduce both the cost of incentives (if $\lambda > 0$) and (if $b < 0$) the excessively high level of prosocial behavior.[41] Next, note that the function

$$b(\theta, \hat{\theta}, y(\hat{\theta})) \equiv v^*_{\hat{\theta}}(y(\hat{\theta}))e_\theta + \epsilon_\theta - c_\theta - \lambda y(\hat{\theta})$$

[41] If $\hat{\theta} = \theta_1$, the strict quasi-concavity of $\mathcal{W}(\theta, \hat{\theta}, y(\hat{\theta}))$ in $\hat{\theta}$ contradicts the fact that $\frac{d\mathcal{W}(\theta, \hat{\theta}, y(\hat{\theta}))}{d\hat{\theta}} |_{\theta = \hat{\theta}} = 0$.

is decreasing in $\hat{\theta}$, since $y' > 0$ and $v^*_{\hat{\theta}}$ is strictly decreasing in $\hat{\theta}$. Therefore, the set

$$B^+_\theta \equiv \{\hat{\theta} \in [\theta_1, \theta_2] \mid b(\theta, \hat{\theta}, y(\hat{\theta})) \geq 0\}$$

is an interval $[\theta_1, \theta^*]$ that contains θ, and the inequality in its definition is strict on $[\theta_1, \theta^*)$ when $\lambda > 0$. It must also contain any other local maximum of $\mathcal{W}(\theta, \hat{\theta}, y(\hat{\theta}))$, since $d\mathcal{W}/d\hat{\theta} < 0$ where $b < 0$. On that interval, we therefore have $\mathcal{A} \geq 0$, which by the proof of Proposition 6 implies that $\mathcal{W}$ is, at least weakly, increasing to the left of θ and decreasing to the right. Therefore θ weakly dominates any possible other local maximum on $[\theta_1, \theta_2]$, and so it must be a global maximum on that interval. For $\lambda > 0$, furthermore, $\mathcal{A} > 0$ so the monotonicities are strict and θ is the unique maximum.

Turning next to off-path deviations, we have

$$\frac{\partial^2 \mathcal{W}(\theta, \hat{\theta}, y)}{\partial \theta \partial y} = \left(-\frac{\partial v^*_{\hat{\theta}}(y)}{\partial y} \right) \frac{\partial b}{\partial \theta}(\theta, \hat{\theta}, y) f_\theta(v^*_{\hat{\theta}}(y)) > 0,$$

by (B.17) and the fact that $\frac{\partial b}{\partial \theta} > 0$, $\frac{\partial f_\theta}{\partial \theta} = 0$, and $\frac{\partial F_\theta}{\partial \theta} = 0$. Thus ($\text{SOC}_2$) is satisfied everywhere, which as we showed rules out off-path deviations.

(b) Cost: When θ affects c_θ, the sign of $\mathcal{A}$ is that of $-y'(\theta)b$, since $\partial b/\partial \theta < 0$. For an *antinorm* we have $y' < 0$, so here again $sgn(\mathcal{A}) = sgn(b)$. In this case the function $b(\theta, \hat{\theta}, y(\hat{\theta}))$ is increasing in $\hat{\theta}$, since $y(\hat{\theta})$ is decreasing, and $v_{\hat{\theta}}$ increasing, in θ. Therefore B^+_θ is again an interval that contains θ and any other possible local maximum, implying that $\mathcal{A} \geq 0$ between θ and any potential local maximum to its left or to its right, with strict inequality when $\lambda = 0$. So once again, θ must be the global maximum of $\mathcal{W}(\theta, \hat{\theta}, y(\hat{\theta}))$. To rule out off-path deviations, finally, note that the expression for $\mathcal{B}$ is the same as in the previous case; the term $\partial b/\partial \theta$ is now negative, but since $y' < 0$, $\mathcal{B}$ is again positive everywhere.

With a *norm,* in contrast, $y' > 0$ so $\mathcal{A} \leq 0$, with strict inequality at $\hat{\theta} = \theta$ when $\lambda > 0$ and everywhere else when $\lambda = 0$. Thus (SOC_1) fails to hold even locally, implying that there can be no separating equilibrium. ■

B.5 Applications: pooling equilibria

Proof of Proposition 8

We will first focus on a full-pooling equilibrium in which all principal types offer the same incentive $y^\sharp$, which maximizes the full-pooling payoff (of all types, as they have identical preferences; this can be called the "best full-pooling outcome"), and later show that this equilibrium is Pareto-dominant, and so arguably focal, in case other equilibria exist. In any equilibrium (full-pooling or not), we will let $G = G(.)$ denote the prior distribution over $\theta \in [\theta_1, \theta_2]$, $G(. \mid v)$ the posterior distribution conditional on having type v, $G(. \mid y)$ the posterior distribution conditional on observing y,[42], and $G(. \mid y, v)$ the posterior distribution conditional on observing both $\{y, v\}$.

Given an (on- or off-path) offer y, let us ensure the uniqueness of a cutoff v^* by assuming that image concerns are not too strong; more formally, given type v 's posterior beliefs $G(\theta \mid y, v)$, let

[42] So $G(. \mid y) \equiv G(.)$ for a full-pooling equilibrium offer y.

$\Delta_{G(.|y,v)}(v^*, v) \equiv \int_{\theta_1}^{\theta_2} \Delta_\theta(v^*) dG(\theta \mid y, v)$ denote the average image concern when the cutoff type is v^*, given conditional distribution $G(\theta \mid y, v)$. We assume that the function $ve + \Delta_{G(.|y,v)}(v^*, v)$ is strictly increasing in v (which is the case if the intensity of image concerns is not too large), ensuring the existence of a unique cutoff $v^*(y)$, defined by the following fixed-point equation:

$$v^*(y)e - c + y + \Delta_{G(.|y,v^*(y))}(v^*(y), v^*(y)) = 0, \tag{B.20}$$

as we will assume all along that the cutoffs are interior. Next, except in a separating equilibrium, beliefs are not point expectations, even on-path.[43], whereas the principal's objective function $\mathcal{W}(\theta, \hat{\theta}, y)$ was defined so far as a function of such point expectations $\hat{\theta}$. In the image-concerns/social monitoring case (Case 1 below), agents' behavior (on- and off-path) depends only on mean beliefs about μ_θ, so (after a relabeling, $\mu_\theta \equiv \theta$, without loss of generality) we will be back to a scalar $\hat{\theta}$. In contrast, an additional assumption will be needed for a *left-truncation parameter* (Case 2); there, we will posit a uniform distribution F, so as to similarly make the entire conditional distribution of θ matter only through its conditional expected mean.

The intuition behind the Pareto-dominance of the full-pooling equilibrium in both cases is that (i) the principal, who does not directly care about the unknown parameter, always benefits from a higher value of the parameter θ (if M^+, say), and (ii) in an alternative equilibrium, the martingale property of beliefs applied to the mean value of the parameter θ implies that at least one induced mean belief in the support of the equilibrium distribution of mean beliefs is smaller than the prior mean, and so the welfare in the alternative equilibrium cannot exceed that in the full-pooling one.

1. Image concerns. For image concerns, agents' behavior depends only on the average μ_θ (and the principal's payoff does not directly depend on θ). Let $\bar{\mu}$ denote the prior expectation of μ: $\bar{\mu} \equiv E_G[\mu_\theta]$. In a full-pooling equilibrium, condition (1) can be rewritten as $v^*(y)e - c + y + \bar{\mu}[E^+[v^*(y)] - E^-[v^*(y)]] = 0$; relabeling G, there is no loss of generality in assuming that $\mu_\theta \equiv \theta$ for all θ, in which case $\bar{\mu} = \bar{\theta} \equiv E_G[\theta]$. More generally, with this relabeling any posterior distribution about θ can be reduced to a scalar variable equal to the expected θ given incentive y , namely $\hat{\theta}(y) \equiv E_{G(.|y)}[\theta]$. For a given incentive y and induced beliefs $\hat{\theta}$, one can then define

$$\mathcal{W}(\theta, \hat{\theta}, y) \equiv \int_{v^*_{\hat{\theta}}(y)}^{v^{max}} (ve + \epsilon - c - \lambda y) f(v) dv$$

where $v^* = v^*_{\hat{\theta}}(y)$ is given by (B.20) applied to image concerns,

$$v^* e - c + y + \hat{\theta}[E^+[v^*] - E^-[v^*]] = 0 \tag{B.21}$$

As usual, to the extent that image is positional, its intensity does not enter the social welfare function, so that $\mathcal{W}(\theta, \hat{\theta}, y)$ above is in fact a function of $\hat{\theta}$ and y only, which will be denoted $\mathcal{Z}(\hat{\theta}, y)$. Furthermore, $\frac{\partial b}{\partial \theta} = 0$ and $\frac{\partial h_\theta}{\partial \theta} = 0$, so $\mathcal{A} \equiv 0$.

Whether the equilibrium is a full pooling one or not, all equilibrium policies must yield the same payoff to the principal. The full-pooling equilibrium we focus on is one in which the principal, regardless of her type, picks (the unique) $y^\sharp = \arg\max_{y'}\{Z(\bar{\theta}, y')\}$. There may

[43] Even in a separating equilibrium, off-path beliefs need not be point expectations, but we have been so far able to establish equilibrium existence limiting attention to such beliefs.

also exist other equilibria, with multiple pooling offers y_i that covary with perceived social monitoring $\hat{\theta}_i$. Full pooling is the unique equilibrium if we make the "Markovian" assumption that the principal's strategy does not depend on θ, which is payoff irrelevant for her. Even without this assumption, we verify below that, for the social-vigilance application (and later the left-truncation one), the full-pooling equilibrium is the Pareto-optimal one –it is strictly preferred by all types of principal to any other equilibrium

Consider thus an alternative equilibrium to full-pooling equilibrium at $y^\sharp$. In this alternative equilibrium, let Ω denote the set of mean beliefs $\hat{\theta}$ that arise for some equilibrium offer. The martingale property implies that $inf_\Omega(\hat{\theta}) \leq \bar{\theta}$. Put differently, in the equilibrium graph of this alternative equilibrium, either $\hat{\theta} = \bar{\theta}$ for all $\hat{\theta}$ in the graph, and then the equilibrium is outcome-equivalent to the full-pooling one (agents' and principal's utilities are the same as under full pooling); or, more interestingly, there exists at least two on-path points $\{\hat{\theta}_1, y_1\}$ and $\{\hat{\theta}_2, y_2\}$ with $\hat{\theta}_1 < \bar{\theta} < \hat{\theta}_2$. Furthermore, for each equilibrium $\hat{\theta}$, the principal's welfare is equal to $Z(\hat{\theta}_1, y_1)$, which is at most equal to $\max_y\{Z(\hat{\theta}_1, y)\}$ Because $max_y\{Z(\hat{\theta}, y)\}$ is increasing in $\hat{\theta}$[44], the principal's payoff is higher in the full-pooling equilibrium at $y^\sharp$ than in the alternative equilibrium when making offer y_1 (and by implication when making any other equilibrium offer).

2. Left-truncation parameter Truncate the cumulative distribution so that only types $v \geq v^{min} + \theta$ remain: the new distribution is $F_\theta(v) = [F(v) - F(v^{min} + \theta)]/[1 - F(v^{min} + \theta)]$ on $[v^{min} + \theta, v^{\max}]$. A truncation corresponds to an increase in θ, and so to a decrease in image incentives. The hazard rate is invariant to left truncations, $\frac{\partial h_\theta}{\partial \theta} = 0$ as long as the cutoff v^*_θ is interior, and so a shift in θ does not affect the ratio of the marginal cost and benefit of enlisting agents at the margin. Furthermore, $\partial b/\partial\theta = 0$, as well, and so $\mathcal{A} = 0$.

Let now F be uniform on $[v^{min} + \theta, v^{\max}]$; for a cutoff v^* interior in this interval,

$$E^-_\theta(v^*) = \frac{v^* + v^{min} + \theta}{2}.$$

So, letting $\bar{\theta} \equiv E_G[\theta]$,

$$E^-_G(v^*) = \frac{v^* + v^{min} + \bar{\theta}}{2}$$

More generally, we focus on mean beliefs, on- and off-path. Letting, for an arbitrary y , $\hat{\theta}(y)$ denote the mean of θ given distribution $G(. \mid y)$,

$$E_{G(.|y)}[E^-_\theta(v^*)] = \frac{v^* + v^{min} + \hat{\theta}(y)}{2}$$

The ranking of these truncated expectations is independent of v^*. Note also that $E^+(v^*) = \frac{v^* + v^{max}}{2}$ is independent of θ for left truncations. The (interior) cutoff $v^*_{\hat{\theta}}(y)$ is now defined by:

$$v^*_{\hat{\theta}}(y)e - c + y + \mu\left(\frac{v^{max} - v^{min} - \hat{\theta}}{2}\right) = 0,$$

and the principal's welfare equals

$$\mathcal{W}(\theta, \hat{\theta}, y) \equiv \left(\frac{1}{1 - F(v^{min} + \theta)}\right)\int_{v^*_{\hat{\theta}}(y)}^{+\infty}[ve + \epsilon - c - \lambda y]f(v)dv.$$

[44] $Z(\hat{\theta}_1, y_1) \leq Z(\hat{\theta}_1, argmax_{y'}\{Z(\hat{\theta}_1, y'\})$. Furthermore, the envelope theorem implies that $\frac{\partial Z(\hat{\theta}_1, argmax_{y'}\{Z(\hat{\theta}_1, y'\})}{\partial \hat{\theta}_1} = (-\frac{\partial v^*_{\hat{\theta}}}{\partial \hat{\theta}})b > 0$ as $(-\frac{\partial v^*_{\hat{\theta}}}{\partial y})b - \lambda(1 - F) = 0$.

All types θ have the same normalized payoff $[1 - F(\theta + v^{min})]\mathcal{W}(\theta, \hat{\theta}(y), y)$. We look for a full-pooling equilibrium, in which all types offer the same incentive $y^\sharp$ and agents with $v \geq v^*_{\bar{\theta}}(y^\sharp)$ contribute, and which is best for all principal types. The jointly-optimal pooling $y^\sharp$ offer maximizes over y the payoff $\int_{v^*_{\bar{\theta}}(y)}^{+\infty}(ve+\epsilon-c-\lambda y)f(v)dv$, subject to $v^*_{\bar{\theta}}(y)e-c+y+(\mu/2)(v^{max}-v^{min}-\bar{\theta})=0$. The off-path beliefs can be taken to be passive: following any incentive y, agents believe that it is a full-pooling offer, i.e., that it comes from the distribution G; thus, $\hat{\theta}(y) = \bar{\theta}$ for all y .

Let us now show that this optimal equilibrium among full-pooling equilibria dominates all other equilibria (if any). Let $\sigma(y)$ denote the probability that the principal plays y on the equilibrium path (not necessarily a full-pooling-equilibrium one), with $\int \sigma(y)dy = 1$. All incentives in the support of σ necessarily yield the same payoff to the principal. Recalling that G denotes the prior cumulative distribution of θ , the martingale property (viewed from the point of view of a third-party observer) writes, for any given v^*:

$$E^-_G(v^*) = \int \sigma(y) E_{G(.|y)}[E^-_\theta(v^*)]dy$$

Select an on-path incentive y such that $E^-_{G(.|y)}(v^*) \leq E^-_G(v^*)$. This property is independent of v^* for a uniform distribution, as we just saw; it follows from the martingale property. So, for such a y, image incentives are always smaller (and the cutoff larger) under posterior beliefs $G(. \mid y)$ than under the prior beliefs G: for any v^*,

$$v^*e - c + y + \frac{v^{max} - v^{min} - \bar{\theta}_{G(.|y,v^*)}}{2} \leq v^*e - c + y + \frac{v^{max} - v^{min} - \bar{\theta}_{G(.|v^*)}}{2.}$$

Therefore, $v^*_{\hat{\theta}(y)}(y) \geq v^*_{\bar{\theta}}(y)$ and

$$\begin{aligned} max_y \int_{v^*_{\hat{\theta}(y)}(y))}^{+\infty}(ve+\epsilon-c-\lambda y)f(v)dv &\leq max_y \int_{v^*_{\bar{\theta}}(y)}^{+\infty}(ve+\epsilon-c-\lambda y)f(v)dv \\ &= \int_{v^*_{\bar{\theta}}(y^\sharp)}^{+\infty}(ve+\epsilon-c-\lambda y^\sharp)f(v)dv \end{aligned}$$

Thus, the payoff cannot exceed that under full pooling.

C. Proofs for Section V: Norm-Based Interventions

Proof of Proposition 10. In the absence of disclosure, let $G(\theta|v, \mathcal{P})$ denote the updated distribution of beliefs about θ of an agent with intrinsic motivation v, when the principal uses some equilibrium disclosure strategy denoted by $\mathcal{P}$. Because the equilibrium is no longer separating, each agent now also learns about θ through introspection, i.e., through his own type v. The agent contributes if $ve - c + E_{G(\theta|v,\mathcal{P})}[E_\theta[\tilde{v}|a = 1] - E_\theta[\tilde{v}|a = 0]] \geq 0$. Provided that reputational concerns μ are not too large, which we will assume, the function on the left-hand side increases in v, so equilibrium in agents' choices is again defined by a cutoff, which we will denote $v^*_\emptyset$ and take to be interior without loss of generality:

$$m_{G(\theta|v^*_\emptyset,\mathcal{P})}(v^*_\emptyset, v^*_\emptyset) = v^*_\emptyset e - c + E_{G(\theta|v^*_\emptyset,\mathcal{P})}[\Delta_\theta(v^*_\emptyset)] = 0, \tag{C.1}$$

which has a unique solution for μ not too large. As explained in the text, the principal discloses if and only if $v_\theta^* \leq v_\emptyset$, which defines a cutoff rule for θ.

Turning to the second part of the proposition, we focus on the case where M^+ holds; that of M^- proceeds similarly. Let $G(\theta|v)$ denote the conditional distribution of θ knowing v only and $G^R_{\tilde\theta}(\theta|v)$ its right-truncation at $\tilde\theta$:

$$G^R_{\tilde\theta}(\theta|v) \equiv \begin{cases} \frac{G(\theta|v)}{qG(\tilde\theta|v)+1-q} & \text{for } \theta \leq \tilde\theta \\ \\ \frac{G(\tilde\theta|v)+[G(\theta|v)-G(\widetilde{\theta}|v)](1-q)}{qG(\widetilde{\theta}|v)+1-q} & \text{for } \theta \geq \widetilde{\theta}. \end{cases}$$

In an equilibrium with agent cutoff $v_\emptyset^*$ and principal's threshold $\tilde\theta$, we have $G(\theta|v,\mathcal{P}) = G^R_{\tilde\theta}(\theta|v)$. Now, as q rises $G^R_{\tilde\theta}(\theta|v)$ increases for all $(\tilde\theta, v)$, and thus so does $G(\theta|v,\mathcal{P})$: agents' non-disclosure beliefs about θ worsen, in the first-order stochastic sense. Given that $\Delta_\theta(v)$ increases with θ under a norm, its expectation under $G(\theta|v,\mathcal{P})$ therefore falls with q, so by (C.1) $v_\emptyset^*$ must rise. Since the principal discloses if and only if $v_\theta^* \leq v_\emptyset^*$, this implies more disclosure. ■

D. Extensions

Proof of Proposition 11. First, if $W_a(v_a^*(1),\theta_L) - W_a(v_a^*(0),\theta_L) > W_b(y^{SI}_{\theta_L},\theta_L) - W_b(y^{SI}_{\theta_H},\theta_L)$, which can be ensured by taking θ_H and θ_L not too different, the θ_H type cannot chose her symmetric-information incentive $y^{SI}_{\theta_H}$, as the θ_L type would then want to mimic. In order to separate, the θ_H type must choose a lower level of y; in particular, $y^{AI}_{\theta_H}$ defined by $W_a(v_a^*(1),\theta_L) - W_a(v_a^*(0),\theta_L) \equiv W_b(y^{SI}_{\theta_L},\theta_L) - W_b(y^{AI}_{\theta_H},\theta_L)$ achieves least-cost separation.

Next, note that $\partial^2 W_b/\partial\theta\partial y = [\epsilon_b - (1+\lambda)y)] - f'_\theta - \lambda f_\theta$; as λ becomes small, all terms vanish except $-f' < 0$, so the cross derivative is negative. For the a activity, $\partial^2 W_a/\partial\theta\partial v_a^* = -(v_a^* e_a + \epsilon_a - c_a)f'(v-\theta) < 0$, given our assumptions. Checking D1 consists in looking at the strictly beneficial beliefs sets $[\rho^*_\theta(y),1]$ following any off-path deviation y. For $y < y^{AI}_{\theta_H}$, type θ_H's payoff W_a is unchanged while W_b decreases since $y < y^{AI}_{\theta_H} < y^{SI}_{\theta_H}$, so there is a net loss; for type θ_L, W_a increases to $W_a(v_a^*(1),\theta_L)$ but W_b falls to a lower level than $W_b(y^{AI}_{\theta_H})$, so her total payoff is less than $W_a(v_a^*(1),\theta_L) + W_b(y^{AI}_{\theta_H},\theta_L)$, which equals her equilibrium payoff, so here again the deviation is not profitable. For $y > y^{AI}_{\theta_H}$, the set is larger for θ_L (unless both are empty). To show this, let $\hat\rho$ be the belief induced by the deviation. Concerning W_a, we have

$$\mathcal{W}_a(v_a^*(1),\theta_H) - \mathcal{W}_a(v_a^*(0),\theta_L) > \mathcal{W}_a(v_a^*(0),\theta_H) - \mathcal{W}_a(v_a^*(0),\theta_L) > \mathcal{W}_a(v_a^*(\hat\rho),\theta_H) - \mathcal{W}_a(v_a^*(\hat\rho),\theta_L),$$

where the last inequality follows from $\partial^2 W_a/\partial\theta\partial v_a^* < 0$. Therefore,

$$\mathcal{W}_a(v_a^*(\hat\rho),\theta_L) - \mathcal{W}_a(v_a^*(0),\theta_L) > \mathcal{W}_a(v_a^*(\hat\rho),\theta_H) - \mathcal{W}_a(v_a^*(1),\theta_H),$$

Concerning W_b, $\partial^2 W_b/\partial\theta\partial y < 0$ and $y > y^{AI}_{\theta_H} > y^{SI}_{\theta_H L}$ implies

$$W_b(y,\theta_L) - W_b(y^{SI}_{\theta_L},\theta_L) \equiv W_b(y,\theta_L) - W_b(y^{AI}_{\theta_H},\theta_L) > W_b(y,\theta_H) - W_b(y^{AI}_{\theta_H},\theta_H).$$

Thus, if the deviation raises $W_a + W_b$ for type θ_H, it raises it by strictly more for type θ_L. D1 beliefs following y should therefore be $\hat\rho = 0$, and so the separating equilibrium satisfies D1. ■

D.1 Allowing for other social payoffs

We now generalize agents' social payoffs beyond image concerns to include other sources of norms, both reputational and non-reputational, such as reciprocity, conformity and other club-type effects. Under symmetric information about θ, let agents' preferences be of the form

$$U = [m_\theta(v, v^*) + y]a + n_\theta(v, v^*), \tag{D.2}$$

where m_θ and n_θ are assumed to be $\mathcal{C}^2$ in all arguments. The term $m_\theta + y$ is the agent's perceived *net return* to choosing $a = 1$, comprising (i) any monetary or other extrinsic incentive y (we again assume quasi-linear preferences for simplicity), and (ii) all other sources (and costs) of motivation, $m_\theta \gtrless 0$. Individual motivation increases with the agent's type v, ensuring a cutoff v^* for choosing $a = 1$. It also depends on where that cutoff lies in the distribution (e.g., on how many people participate). In the image-concern model for instance, $m_\theta(v, v^*) = ve_\theta - c_\theta + \Delta_\theta(v^*)$.

The final term in (D.2) is the *unconditional*, "baseline" component $n_\theta(v, v^*)$ of the agent's utility when not acting ($a = 0$), which incorporates all external benefits received by type v when the set choosing $a = 1$ is $[v^*, +\infty)$ and the environment is described by θ. In the image-concern model, these benefits (or costs) were equated to $n_\theta(v, v^*) = [1 - F_\theta(v^*)]\epsilon_\theta + \mu_\theta\left[E_\theta^-(v^*) - \bar{v}_\theta\right]$.

To avoid issues of multiplicity we assume that the function $m_\theta(v^*, v^*)$ is strictly increasing in v^*. Agents' behavior given a belief $\hat{\theta}$ (equal to θ under symmetric information and to $\hat{\theta}(y)$ under asymmetric information) is then entirely summarized by the equation determining the equilibrium cutoff $v^*_{\hat{\theta}}(y)$, is given (when interior) by

$$m_{\hat{\theta}}(v^*_{\hat{\theta}}(y), v^*_{\hat{\theta}}(y)) + y = 0. \tag{D.3}$$

The principal's objective function is, as before, agents' average utility, evaluated according to the objective θ, minus the opportunity cost of funds. Let us define, for any (v^*, y, θ), the functions

$$W^{FB}_\theta(v^*) \equiv \int_{v^*}^{+\infty} m_\theta(v, v^*) f_\theta(v)\, dv + \int_{-\infty}^{+\infty} n_\theta(v, v^*) f_\theta(v) dv \tag{D.4}$$

$$W_\theta(v^*, y) \equiv W^{FB}_\theta(v^*) - \lambda y[1 - F_\theta(v^*)] \tag{D.5}$$

$$\mathcal{W}(\theta, \hat{\theta}, y) \equiv W_\theta(v^*_{\hat{\theta}}(y), y) \tag{D.6}$$

representing respectively the principal's welfare in the first-best setting, under symmetric-information with costly incentives, and under asymmetric information with agents' belief $\hat{\theta}$.

Before stating results for this general framework, we illustrate it with four applications. For conciseness they involve uniform shifts in the goodness of society, $F_\theta(v) = F(v - \theta)$, but our generalization applies more broadly. The first two applications involve image concerns that differ from those of the basic framework laid out in Section II. The last two replace image concerns with other forms of socially determined preferences. In each setting we indicate: (i) the corresponding case of Table 1, and the resulting direction in which expressive concerns will distort incentives in a separating equilibrium, if there is one; (ii) whether or not the (unchanged) second-order conditions hold, ensuring that such an equilibrium exists, or if some pooling must arise.[45]

[45] As in Section IV.E, we focus the exposition on the main condition $\mathcal{A} \geq 0$ (SOC_1), and verify in Section D.3 that, when it does, $\mathcal{B} \geq 0$ (SOC_2) also holds, over the relevant range to which an optimal y must belong.

1. Differentially valued audiences (M^+, P^-)**.** Ellingsen and Johannesson (2008) emphasize that people typically care more about being esteemed by types who are themselves highly respected. Formally, an agent assigns weight $\rho(v_j)$ to the image he has in the eyes of individual j with type v_j, with $\rho' > 0$. The framework is the same as in the core model, with now

$$\mu_\theta \equiv \int_{v^{\min}}^{v^{\max}} \rho(v) f_\theta(v) dv,$$

Given that $F_\theta(v) = F(v-\theta)$, μ_θ is increasing in θ, and so is $\Delta_\theta(v^*) = \mu_\theta[E^+(v^*-\theta) - E^-(v^*-\theta)]$ provided that $(E^+ - E^-)'$ is negative, or not too positive (e.g., with a uniform distribution, it is constant). Through this mechanism, a society with more intrinsically motivated individuals also benefits from stronger social norms. Image concerns are again zero-sum and therefore vanish in the social welfare function. The analysis is identical to that of a norm with unweighted image concerns, with $y_\theta^{FB} = \epsilon - \mu_\theta \Delta_\theta(\frac{c-\epsilon}{e})$ and $\mathcal{A} > 0$, except that under asymmetric information $(-\partial v_{\hat\theta}^*/\partial\hat\theta)$ tends to be higher for weighted image concerns, making expressive law, obtained from (B.2), even softer.[46]

2. Endogenous visibility of contributions (M^+, P^-)**.** Keeping with zero-sum image concerns, another factor contributing to the emergence of a norm, or strengthening an existing one, arises when whether one contributes or not is more likely to be observed if the number of contributors increases. For instance, if only contributors observe other people's choices (say, in volunteering for a cause) and θ is a parameter of shift of the distribution,[47]

$$\Delta_\theta(v^*) = [1 - F_\theta(v^*)]\mu[E^+(v^* - \theta) - E^-(v^* - \theta)].$$

Here again, the new term entering $\partial v^*/\partial\theta$ increases the extent to which a higher θ is motivation-enhancing (or can even change its impact from M^- to M^+).[48] It also makes $y_\theta^{FB} = \epsilon - \mu_\theta \Delta_\theta(\frac{c-\epsilon}{e})$ decline faster with θ and leads, through (B.2), to a further weakening of incentives y_θ^{AI} in the separating equilibrium, which obtains since $\mathcal{A} > 0$.

3. Conformism (M^+, P^+) **and anticonformism** (M^-, P^-)**.** A pure preference for conformity (e.g., Bernheim 1994) or a "strength in numbers" externality delivers payoffs increasing in the size of one's group , such that $m_\theta(v, v^*) = ve - c + \alpha[1 - 2F_\theta(v^*)]$ and $n_\theta(v, v^*) = \alpha F_\theta(v^*) + [1 - F_\theta(v^*)]\epsilon$, where $\alpha \geq 0$ captures the strength of the conformity motive.[49] Conversely, the case $\alpha < 0$ corresponds to a search for "exclusivity." Social welfare equals

$$\mathcal{W}(\theta, \hat\theta, y) = \int_{v_{\hat\theta}^*}^{+\infty} (ve + \epsilon - c - \lambda y) f_\theta(v) dv + \alpha\left[[1 - F_\theta(v_{\hat\theta}^*(y))]^2 + [F_\theta(v_{\hat\theta}^*(y))]^2\right] \tag{D.7}$$

and the first-best incentive is $y_\theta^{FB} = \epsilon + \alpha[1 - 2F(v_\theta^* - \theta)]$, so that P^+ obtains under conformity and P^- under anti-conformity or exclusivity seeking. The second externality from participation

[46] In this case, $m_\theta(v, v^*) = ve - c + \mu_\theta[E_{v^*}^+(\tilde v) - E_{v^*}^+(\tilde v)]$ and $n_\theta(v, v^*) = \mu[E_{v^*}^-(\tilde v) - \bar v_\theta] + [1 - F(v^*)]\epsilon$. The claim concerning $-\partial v_{\hat\theta}^*/\partial\hat\theta$ is formally true as long as the function ρ is weakly concave (or not too convex). In this case, $\mu_\theta = E_\theta[\rho(v)] \leq \rho(E_\theta[v]) = \rho(\bar v) \equiv \mu^\dagger$ for all θ, hence

$$-\frac{\partial v_\theta^*}{\partial\theta} = \frac{\mu_\theta'(E_\theta^+ - E_\theta^-)(v_\theta^*) - \mu_\theta(E_\theta^+ - E_\theta^-)'(v_\theta^*)}{e_\theta + \mu_\theta(E_\theta^+ - E_\theta^-)'(v_\theta^*)} > \frac{-(E_\theta^+ - E_\theta^-)'(v_\theta^*)}{e_\theta/\mu_\theta + (E_\theta^+ - E_\theta^-)'(v_\theta^*)} \geq \frac{-\mu^\dagger(E_\theta^+ - E_\theta^-)'(v_\theta^*)}{e_\theta/\mu^\dagger + (E_\theta^+ - E_\theta^-)'(v_\theta^*)}.$$

[47] Conversely, it may be the case that norm violators are the only agents observing who obeys or does not obey the norm. Asymmetric observability then tends to make participants' behaviors strategic substitutes.

[48] In this case, $m_\theta(v, v^*) = ve - c + \mu[1 - F_\theta(v^*)][E_\theta^+(v^*) - E_\theta^+(v^*)]$ and $n_\theta(v, v^*) = \mu[E_\theta^-(v^*) - \bar v_\theta] + [1 - F_\theta(v^*)]\epsilon$.

[49] To ensure the uniqueness of the cutoff, we assume that $e > 2\alpha f_\theta$.

is here not image stealing but changing the relative sizes of the two groups, which is no longer a zero-sum game. In Section D.3 below, we show that when $\alpha > 0$ we have $\mathcal{A} > 0$ provided the hazard rate of F is not too increasing, leading to a separating equilibrium, and conversely $\mathcal{A} < 0$ when α is small enough, in which case some pooling must occur. For anticonformity, $\alpha < 0$, on the other hand (which implies that $\partial y^{FB}/\partial\theta < 0$), $\mathcal{A} > 0$ holds for $|\alpha|$ small enough. In this case a separating equilibrium obtains, in which incentives are stronger under asymmetric information, $y_\theta^{AI} > y_\theta^{SI}$.

4. Forms of prosociality. In all that follows, let $n_\theta(v, v^*) = \epsilon[1-F_\theta(v^*)]$, so that participation is unambiguously a prosocial action. *Unconditional altruism* then corresponds to $m_{\hat\theta}(v, v^*) = ve - c$ and $n_\theta(v, v^*) = \epsilon[1-F_\theta(v^*)]$ for all $\hat\theta$, which is our benchmark model when image concerns are absent. More interestingly, indirect *type-based reciprocity*, that is, altruism conditional on the "goodness" of others (as in Levine 1998 or van der Weele 2012) corresponds to m_θ increasing in θ (for example $m_{\hat\theta}(v, v^*) = (v + \alpha\hat\theta)e - c$ with $\alpha > 0$), implying M^+). The first-best incentive is type independent, $y_\theta^{FB} = \epsilon$, because the social preference term, $\alpha\hat\theta e$, is internalized by the principal. With even a small cost of incentives, however, the symmetric-information solution y_θ^{SI} satisfies P^- when $f'_\theta > 0$ and α is not too large, as we show in Section D.3. Under asymmetric information, there are two opposing effects at play. On the one hand, a higher θ type pays y to more people (the usual effect), making the principal less eager to enlist the norm to boost participation. On the other, such a type also gains more in conformity benefits from raising participation (since $\partial F_\theta(v_\theta^*)/\partial\theta < 0$). We show in Section D.3 that when α is relatively small the first effect dominates and there is a separating equilibrium, in which incentives are weakened, $y_\theta^{AI} < y_\theta^{SI}$. Otherwise, some pooling may occur.

Indirect *action-based reciprocity,* finally, is based not on who others are, but on what they do: $m_{\hat\theta}(v, v^*)$ depends on v^* through $\bar a_{\hat\theta}$, e.g., $m_{\hat\theta}(v, v^*) = v\bar a_{\hat\theta}^\alpha e - c$, $\alpha > 0$. The interior cutoff v_θ^* is thus implicitly defined by $v_\theta^*[1 - F(v_\theta^* - \theta)]^\alpha = c - y$, and at that point the left-hand side expression is increasing in v_θ^*, implying that $\partial v_\theta^*/\partial\theta < 0$, so M^+ obtains, while $\partial[1 - F(v_\theta^* - \theta)]/\partial\theta > 0$.[50] Concerning y^{FB}, we show in Section D.3 that

$$y^{FB} = \epsilon + \alpha[1 - F(v_\theta^* - \theta)]^\alpha,$$

implying that P^+ obtains. Finally, in Section D.3 we also check that $\mathcal{A} > 0$ when the hazard rate of F is not too strongly increasing, relative to α. There is then a separating equilibrium, which involves tough law.

Related reciprocity and conformity concerns were studied by Sliwka (2008) and van der Weele (2012). In Van der Weele (2012) there are non-conformist selfish types and conditional cooperators, who are willing to contribute if and only the number of those who do exceeds some threshold (a form of conformism). In Sliwka (2008) there are two non-conformist types (selfish and prosocial) with steadfast preferences, and a conformist type who also follows a rule of contributing when the contributions of others exceed a cutoff. Both papers identify conditions under which a separating equilibrium obtains, in which an informed principal sets low incentives to convince agents that there is a high fraction of conditional cooperators in the population.[51]

[50] There is also a corner equilibrium with zero contributions at $v_\theta^* = v^{\max} + \theta$ (but it is worse for all principal types) and an unstable one in-between these two, see Section D.3. We focus on the stable equilibrium with positive participation.

[51] While similar in spirit, there are several differences between these models and ours, in which conformism

D.2 Optimal incentives and expressiveness in the general framework

First best. When the principal has access to costless incentives, her objective function reduces to $\bar{U}_\theta = W_\theta^{FB}(v_\theta^*(y))$. Maximizing over y is equivalent to directly choosing the participation cutoff v_θ^{FB} as the solution (which we assume interior) to:

$$\frac{\partial W_\theta^{FB}}{\partial v^*}\left(v_\theta^{FB}\right) = 0, \tag{D.8}$$

and (using $m_\theta(v^*, v^*) + y = 0$) the corresponding incentive is given by

$$y_\theta^{FB} = \frac{\int_{-\infty}^{+\infty} -\frac{\partial n_\theta}{\partial v^*}(v, v_\theta^{FB}) f_\theta(v) dv + \int_{v_\theta^{FB}}^{+\infty} -\frac{\partial m_\theta}{\partial v^*}(v, v_\theta^{FB}) f_\theta(v) dv}{f_\theta(v_\theta^{FB})}. \tag{D.9}$$

The first term in the numerator of (D.9) is the generalized Pigouvian wedge, summing over gainers and losers in society all the external effects of a marginal participant–including, say, the prestige-stealing externality in the image-concern model. The second one captures the fact that when aggregate participation changes, so do the non-monetary (intrinsic, social or "club") benefits of participation, in a way that can be either motivation-enhancing (M^+) or motivation-reducing (M^-), as illustrated above. The response of the first-best incentive to a change in the environment is given by differentiating (D.8) with $v_\theta^{FB} = v_\theta^*(y^{FB})$,

$$\frac{dy_\theta^{FB}}{d\theta} = \frac{-1}{\partial v_\theta^*/\partial y}\left(\frac{\partial v_\theta^*}{\partial \theta} + \frac{\partial^2 W_\theta^{FB}/\partial\theta\partial v_\theta^*}{\partial^2 W_\theta^{FB}/\partial v_\theta^{*2}}\right) \tag{D.10}$$

which can be decomposed into two effects. First, the principal offsets the influence of the change, if any, in the environment on agents' behavior: this is the "leaning against the wind" term $(\partial v_\theta^*/\partial\theta)/s_\theta$. Second, a change in θ can also affect the principal's demand for compliance, as reflected in the complementarity/substitutability term $\partial^2 W_\theta^{FB}/\partial\theta\partial v_\theta^*$. Absent such a direct sorting condition, we have:[52]

Corollary 1 *If $\partial^2 W_\theta^{FB}/\partial\theta\partial v_\theta^* = 0$, then $M^+ \iff P^-$ and $M^- \iff P^+$.*

Before turning to the study of expressive law, we note that Proposition 3 and its proof apply verbatim to this environment: when extrinsic incentives are costless to the principal, her first-best policy is implementable whether or not agents know the underlying parameter θ.

Symmetric information. When $\lambda > 0$ and θ is common knowledge, the principal chooses y to maximize $W_\theta^{SI}(y) \equiv \mathcal{W}(\theta, \theta, y) \equiv W_\theta(v_\theta^*(y), y)$, yielding the first-order condition:

$$\frac{\partial W_\theta}{\partial v^*}\frac{\partial v_\theta^*}{\partial y} + \frac{\partial W_\theta}{\partial y} = 0. \tag{D.11}$$

and reciprocity can generate weaker incentives, stronger ones, or pooling, depending on the relevant case. First, discrete distributions do not satisfy the monotone-hazard-rate property, making it difficult to compare second-order conditions, which are key to the existence of a separating equilibrium. Second, in these models agents are heterogeneous in their degrees of conformism, whereas in ours all have the same conformity or reciprocal preferences but differ in altruism. We view the two types of analyses as complementary in showing the rich set out outcomes that can arise when a principal has private information about some distribution of social preferences.

[52] When $F_\theta(v) = F(v - \theta)$, we have $\partial^2 W_\theta^{FB}/\partial v_\theta^*\partial\theta = f'(v^* - \theta)(\epsilon + v_\theta^* e)e - c)$, which for an arbitrary cutoff v_θ^* may be positive or negative. At the first best, however, $\epsilon + v_\theta^* e = c$, so Corollary 1 applies.

As before, since $\partial W_\theta^{SI}/\partial y = -\lambda\big[1 - F_\theta(v_\theta^*(y))\big]$ and $\partial v_\theta^*/\partial y < 0$, this implies that the principal chooses weaker incentives, the more costly they are for him to provide: y_θ^{SI} is decreasing in λ, and in particular $y_\theta^{SI} \leq y_\theta^{FB}$ for all θ since the first-best is achieved for $\lambda = 0$. Accordingly, the principal would benefit from some other source of motivation, such as persuasion, that raises participation: $(\partial W_\theta/\partial v^*)(v_\theta^*(y_\theta^{SI}), y_\theta^{SI}) < 0$. For $\lambda > 0$ small enough, finally, y_θ^{SI} shares the same comparative statics as y_θ^{FB}, which will again underpin agents' inferences under asymmetric information.

Asymmetric information. When θ is private information of the principal, we look again for a separating equilibrium (satisfying (NDB)) in which her policy is a strictly monotonic y_θ^{AI}, which agents invert to form their belief $\hat{\theta}(y)$, while conversely she sets y_θ^{AI} to maximize $\mathcal{W}(\theta, \hat{\theta}(y), y) = W_\theta(v^*_{\hat{\theta}(y)}(y), y)$. As in the baseline model, agents' beliefs matter here only through the determination of the cutoff $v_{\hat{\theta}}(y)$, whereas all final payoffs m_θ and n_θ are evaluated by the principal based on her knowledge of the true θ. Assuming again an interior cutoff, this leads to the differential equation in y_θ

$$\frac{\partial W_\theta}{\partial y}(v_\theta^*(y_\theta), y_\theta) + \frac{\partial W_\theta}{\partial v^*}(v_\theta^*(y_\theta), y_\theta)\left(\frac{\partial v_\theta^*}{\partial y}(y_\theta) + \frac{\partial v^*_{\hat{\theta}}}{\partial \hat{\theta}}(y_\theta) \cdot \frac{1}{dy_\theta^{AI}/d\theta}\right) = 0, \tag{D.12}$$

where $v_\theta^*(y_\theta)$ is given by (D.3) taken at $\hat{\theta} = \theta$.

In the core image-concern model, we showed the existence (for λ small enough) of a unique solution to the differential equation (D.12), co-monotonic with y_θ^{FB} and satisfying (NDB). In what follows we will take existence as given (it could be shown through similar steps) and consider directly the second-order conditions. Recalling the definitions of $W_\theta(v^*, y)$ and $\mathcal{W}(\theta, \hat{\theta}, y)$ from (D.5)-(D.6), denote

$$b_\theta(v^*, y) \equiv -\frac{\partial W_\theta(v^*, y)}{\partial v^*}\frac{1}{f_\theta(v^*)} \tag{D.13}$$

$$= \frac{\int_{v^*}^{+\infty} -\frac{\partial m_\theta(v, v^*)}{\partial v^*} f_\theta(v)dv + \int_{-\infty}^{+\infty} -\frac{\partial n_\theta(v, v^*)}{\partial v^*} f_\theta(v)dv}{f_\theta(v^*)} - (1+\lambda)y \tag{D.14}$$

the marginal value of a contribution given cutoff v^* (equal to zero in the first-best, $\lambda = 0$ and $y = y_\theta^{FB}$), and $b(\theta, \hat{\theta}, y) \equiv b_\theta(v^*_{\hat{\theta}}(y), y)$ that marginal value given belief $\hat{\theta}$ determining the cutoff. With these more general definitions of $W_\theta, \mathcal{W}$ and b entering $\mathcal{A}$ and $\mathcal{B}$, the second-order conditions (SOC_1)-(SOC_2) remain unchanged.

In Section D.3, we will examine these conditions for each of the four new applications listed earlier, verifying the claimed results about separation with weak or strong law, or the necessity of pooling. More generally, we have:

Proposition 12 (soft or tough expressive law) *Assume that: (i) $W_\theta(v_\theta^*(y), y)$ is strictly quasi-concave; (ii) there is an undersupply of prosocial behavior under symmetric information: $\frac{\partial}{\partial v^*}(W_\theta(v_\theta^*(y_\theta^{SI}), y_\theta^{SI})) < 0$; (iii) the differential equation (D.12) admits a monotonic solution $y^{AI}(\cdot)$ satisfying (NDB) and (iv) the (SOC) hold. Then*

(i) Soft law: If either M^+ and P^-, or M^- and P^+ obtain, then $y_\theta^{AI} \leq y_\theta^{SI}$ for all θ, with strict inequality, except at θ_1 in the former case, and at θ_2 in the latter case.

(ii) Tough law: If either M^+ and P^+, or M^- and P^- obtains, then $y_\theta^{AI} \geq y_\theta^{SI}$ for all θ, with strict inequality, except at θ_1 in the former case, and at θ_2 in the latter case.

Proof. The proof of Proposition 6 applies here as well. ■

D.3 Second-order conditions for the applications in Section D.1

We examine here the second-order conditions for each of these applications, verifying the claims ade earlier.

1. Differentially valued audiences (M^+, P^-)**.** In this case θ affects only the distribution f_θ and not the marginal social benefit b, so the sign of $\mathcal{A}$ is that of $b^2 y' \partial h_\theta / \partial \theta \geq 0$, since $y' < 0$. Similarly, the sign of $\mathcal{B}$ remains given by (B.19), and under the maintained assumptions that $f'_\theta \geq 0$ the same reasoning as in the basic image-concern case (Section B.4 of this Appendix) shows that off-path deviations can be ruled out. Thus, a separating equilibrium exists.

2. Endogenous visibility of contributions (M^+, P^-)**.** Here again b is independent of θ and $y' < 0$, so $\mathcal{A} > 0$ and the same reasoning on $\mathcal{B}$ rules out off-path deviations.

3. Conformism (M^+, P^+) **and anticonformism** (M^-, P^-)**.** In this setting,

$$b(\theta, \hat{\theta}, y) = v^*_{\hat{\theta}}(y)e + \epsilon - c - \lambda y + 2\alpha[1 - 2F_\theta(v^*_{\hat{\theta}}(y))].$$

When $\alpha > 0$ this function is increasing in θ, so the sign of $\partial(bh)/\partial\theta$ is generally ambiguous. If α is small and h_θ sufficiently decreasing, the sign is negative, so $\mathcal{A} < 0$ since $y' > 0$. If, on the other hand, the hazard rate is only weakly decreasing, then the effect of α dominates, so $\mathcal{A} > 0$. Turning next to $\mathcal{B}$, (B.17) takes the form:

$$\frac{\partial^2 \mathcal{W}}{\partial\theta\partial y} = \frac{1}{e + 2\alpha f_\theta}\left[4\alpha f_\theta(v_{\hat{\theta}}(y)) - b(\theta, \hat{\theta}, y) f'_\theta(v_{\hat{\theta}}(y))\right] - \lambda f_\theta(v^*_{\hat{\theta}}(y)). \tag{D.15}$$

Fixing α, as λ tends to zero, so does b (approaching the first best), therefore the expression is positive and so is $\mathcal{B}$. Thus, when the conditions above ensuring $\mathcal{A} > 0$ are also satisfied, a separating equilibrium exists.

With $\alpha < 0$, finally, b and therefore hb are decreasing in θ, while $y' < 0$; thus $\mathcal{A} > 0$. As to $\mathcal{B}$, $\partial^2 \mathcal{W}/\partial\theta\partial y < 0$ wherever $b \leq 0$; the reasoning in Section B.4 of this Appendix (externality case) can then by applied again to rule out off-path deviations.

4. Forms of prosociality. Consider first *type-based reciprocity.* First, agents' cutoff is $v^*_\theta = (c - y)/e - \alpha\theta$, and the first-order condition determining y^{SI}_θ thus takes the form

$$[\epsilon - (1 + \lambda)y)] f\left(\frac{c - y}{e} - (1 + \alpha)\theta\right) - \lambda e\left[1 - F\left(\frac{c - y}{e} - (1 + \alpha\theta)\right)\right] = 0.$$

For $\lambda > 0$, the left-hand side is decreasing in θ, provided that $f' > 0$. Since the objective function is strictly quasiconcave in y (for small enough λ, as usual), this implies that $\partial y^{SI}_\theta / \partial\theta < 0$, so P^- obtains. Turning to the second-order condition, since $b(\theta, \hat{\theta}, y) = \epsilon - (1 + \lambda)y$ is independent of θ we have $\mathcal{A} = y' b^2 (\partial h_\theta / \partial\theta) \geq 0$, given the hazard rate condition. Finally, $\partial^2 \mathcal{W}/\partial\theta\partial y$ is given by (B.19), and the same reasoning as the one following that equation can again be applied to rule out off-path deviations.

Consider now *action-based reciprocity.* The function $\chi(v^*) \equiv v^*[1 - F(v^* - \theta)]^\alpha$ is strictly quasi-concave due to the increasing hazard rate, and such that $\chi(v^{min} + \theta) = v^{min} + \theta$ and

$\chi(v^{max}+\theta)=0$. Let v_θ^{**} denote the point where it reaches its maximum, and focus on the relevant case where $y<c$. If $\chi(v_\theta^{**})<c-y$, the only equilibrium involves zero participation, $v_\theta^*=v^{max}+\theta$. Otherwise, there is also another stable equilibrium (where $\chi'>0$) at some v_θ^* in the interval, plus an unstable one (where $\chi'<0$) in between. Furthermore, v_θ^* is interior when $v^{min}+\theta<c-y$. We will focus on the interesting case where $v^{min}+\theta<c-y<\chi(v_\theta^{**})$, and on the stable, interior equilibrium with positive-participation v_θ^* which then solves $v_\theta^*[1-F(v_\theta^*-\theta)]^\alpha=c-y$. Because $\chi'>0$ at that point, $\partial v_\theta^*/\partial\theta<0<\partial[1-F(v_\theta^*-\theta)]/\partial\theta$, as claimed in the text. Turning now to the first-best incentive, we have $-\partial m_\theta/\partial v^*=\alpha v f_\theta(v^*)[1-F(v^*-\theta)]^{(\alpha-1)}$, therefore $\int_{v^*}^\infty(-\frac{\partial m_\theta}{\partial v^*})=\alpha f_\theta(v^*)[1-F(v^*-\theta)]^\alpha$, so applying (D.9) yields the claimed expression for y^{FB}. Consider now (SOC_1): equation (D.14) yields

$$b(\theta,\hat\theta,y)=\alpha e[1-F(v_{\hat\theta}^*(y)-\theta)]^\alpha+\epsilon-(1+\lambda)y,$$

which is increasing in θ. If the hazard rate of F is not increasing too strongly, so that h_θ does not decrease too fast, then $\partial(bh_\theta)/\partial\theta>0$, so $\mathcal{A}>0$ since $y'>0$. Turning next to (SOC_2),

$$\frac{\partial^2\mathcal{W}}{\partial\theta\partial y}=\frac{-\partial v_{\hat\theta}^*(y)}{\partial y}\left[\alpha e[1-F(v_{\hat\theta}^*(y)-\theta)]^{\alpha-1}f_\theta(v_{\hat\theta}(y))-b(\theta,\hat\theta,y)f_\theta'(v_{\hat\theta}(y))\right]-\lambda f_\theta(v_{\hat\theta}^*(y)) \quad \text{(D.16)}$$

is positive for λ and hence b small enough, ensuring that $\mathcal{B}>0$. ■