Laws and Norms

Roland Bénabou

Princeton University, National Bureau of Economic Research, Centre for Economic Policy Research, and Institute of Labor Economics

Jean Tirole

Toulouse School of Economics and Institute for Advanced Study in Toulouse

We analyze how private decisions and optimal public policies are shaped by personal and societal preferences, material incentives, and social norms. We show how incentives interact with honor and stigma, deriving optimal taxation. We then analyze the expressive role of law as embodying society's values and when it calls for a weakening or a strengthening of incentives. The law should be softened when it signals agents' 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.

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 are also

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 as well as to Andrei Rachkov and Edoardo Grillo for superb research assistance. Bénabou gratefully acknowledges support from the Canadian Institute for Advanced Research. Both authors gratefully acknowledge support from the European Union through ERC grant HWS–grant agreement 10109831 (and previously

Electronically published December 12, 2025

Journal of Political Economy, volume 134, number 2, February 2026.
© 2025 The University of Chicago. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published by The University of Chicago Press.
https://doi.org/10.1086/738343

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 substitutabilities in agents' actions (social multiplier)—and by conveying private information she may have about the environment in which they operate

by grant COGNITION 249429 and grant MARKLIM 669217) and from Toulouse School of Economics partnership Chaire Association Finance Durable et Investissement Responsable (AFG). This paper was edited by Leonardo Bursztyn.

[1] Examples of such puzzles include, e.g., Gneezy and Rustichini (2000), Knez and Simester (2001), Fehr and Gächter (2002), Fehr and Rockenbach (2003), Karlan and List (2007), Galbiati and Vertova (2008), Ariely, Bracha, and Meier (2009), Funk (2010), Fryer (2011), Ashraf, Bandiera, and Jack (2014), and Alfitian, Sliwka, and Vogelsang (2024). See, e.g., Bowles (2008) and Bowles and Polania-Reyes (2012) for surveys and Gibbons (1997) and Prendergast (1999) for the more classical literature on incentives in organizations.

(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 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 induce maximal stigma and maximal honor, respectively) and highest for behaviors in the gray 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 (1) when it encourages or discourages agents to engage in prosocial behavior and (2) 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 antisocial 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 nonmonetary 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 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 formulas 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 to 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, in section V we allow the policymaker 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; conversely, 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 in the other one. The consequence of this expressive spillover is a lower use (in optimal policymaking) 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 appendix (available online), 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 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, Snorre, and Nyborg (2003), Besley and Ghatak (2005), and Bénabou and Tirole (2006b).[2] We provide here a comprehensive framework that generates both new testable predictions and optimal tax formulas. Chen (2016), Jia and Persson (2021), and Besley, Jensen, and Persson (2023) build on it to study empirically the interaction of norms and incentives in the context of military desertions, ethnic identity choice, and tax evasion, respectively. Lane, Nosenzo, 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, 2000b), and McAdams (2000). Our signaling approach is most closely related to the one informally advocated by Posner (1998, 2000a, 2000b) and McAdams (2000).[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,

[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 as a tax and a subsidy, respectively) 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) that the principal cannot tell apart but that are subject to separate social norms among agents, giving rise to different social multipliers.

[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; Guiso, Sapienza, and Zingales 2008; Tabellini 2008; Greif and Tadelis 2010; Acemoglu and Jackson 2015).

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 Sliwka (2008), social complementarities operate through conformist types, whose preference is to mimic whatever action the majority chooses. In van der Weele (2012), 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 formulas 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 contribution 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 the difficulty of a task and the previous effort norm among a set of agents, respectively.

Finally, the analysis of direct disclosure connects the paper to the literature on norms-based interventions and pluralistic ignorance (e.g., Cialdini 1984; Miller and McFarland 1987; 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), Allcott (2011), and Ayres, Raseman, and Shih (2013) demonstrate these effects for electricity conservation, and Lefebvre et al. (2015) demonstrates similar ones 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 presidential 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. Schroeder and Prentice (2006) 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 and 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, and contributing—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 coworkers, and so on, and y represents 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:

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

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, and so on. 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]

[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 (2006b) allow for heterogeneous marginal utilities of money and reputational concerns. We abstract here from the overjustification and full crowding-out

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 $E_\theta[\tilde{v}|a]$ 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, and so on. In the social sphere, people perceived as generous, public minded, good citizens, and so on 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 (see Bem 1972; Smith 1759; Bodner and Prelec 2003; Bénabou and Tirole 2004, 2011a).

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

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

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 to feelings of pride and shame, respectively. The net reputational incentive to contribute is

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

For simplicity, we shall focus 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:[6]

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

Note that reputation is here a positional good: $E_\theta[E_\theta(\tilde{v}|a, y)] = \bar{v}_\theta$.[7] Agents' average utility is thus

---

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] 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 $ve_\theta + \Delta_\theta(v)$.

[7] 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).

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

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 fig. 1*A*), 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 *antinorm* effect.

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.[8] 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:

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

Intuition suggests that honor concerns will dominate when people who "do the right thing" ($a = 1$) are fairly rare and that stigma considerations will 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 2004; Harbaugh and Rasmusen 2018; Adriani and Sonderegger 2019).

i. If f_θ is everywhere decreasing (increasing), then Δ_θ is everywhere increasing (decreasing).
ii. If f_θ is unimodal (strictly quasiconcave), then Δ_θ is strictly quasiconvex. 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 quasiconcave. Its maximum is interior provided that $f_\theta(v_\theta^{\min})$ and $f_\theta(v_\theta^{\max})$ are sufficiently large.

[8] The fact that $|\Delta_\theta'|$ is bounded is shown in the appendix. Bénabou and Tirole (2006b) 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 nonreputational norms and incentives, see Weibull and Villa (2005).

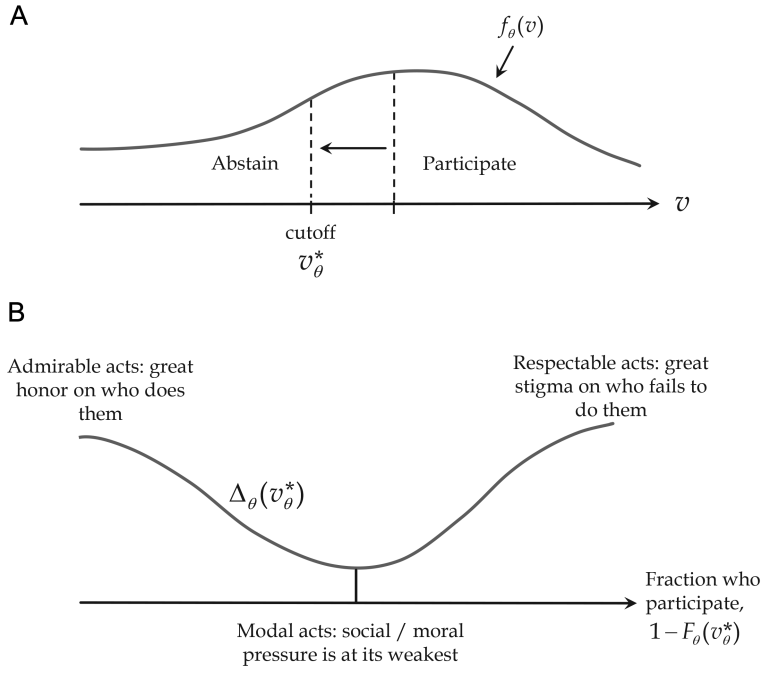

FIG. 1.—*A*, Distribution of preferences. The line $f_\theta(v)$ plots the distribution of intrinsic preferences v. The participation cutoff v^*_θ separates agents who choose the prosocial action ($a = 1$) and those who do not ($a = 0$). The shift (arrow) illustrates the changes in the sets of agents choosing each action as the participation cutoff decreases. *B*, Reputational return. The line $\Delta_\theta(v^*_\theta)$ plots an individual's net image return to choosing the prosocial action ($a = 1$) as a fraction of the equilibrium fraction $1 - F_\theta(v^*_\theta)$ of agents choosing that same action.

iv. In general, the extrema of f_θ and Δ_θ do not coincide. If—in addition to case 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^*.[9]

We shall focus on the unimodal case ii, which is empirically the most natural and allows for both strategic substitutability and complementarity, with case i being a subcase (see fig. 1*B*). 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

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

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.[10]

It is worth noting that the model generates endogenously the two types of signaling motives that 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).[11]

C. *Comparative Statics and Empirics*

Comparative statics.—Let $m_\theta(v, v^*)$ denote agent v's nonextrinsic 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 motivation-reducing parameter). A parameter θ is (in equilibrium)

- motivation enhancing (M^+) if $(\partial m_\theta/\partial\theta)(v_\theta^*(y), v_\theta^*(y)) > 0$;
- motivation reducing (M^-) if $(\partial m_\theta/\partial\theta)(v_\theta^*(y), v_\theta^*(y)) < 0$.

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

[10] 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 2006b).

[11] Brennan and Brooks (2007) do not formulate a signaling model but postulate, on the basis of 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 that the distribution of types is unimodal.

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

$$\frac{\partial v_\theta^*(y)}{\partial\theta} = -\frac{v_\theta^*(y)\gamma(\partial\epsilon_\theta/\partial\theta) - (\partial c_\theta/\partial\theta) + (\partial\Delta_\theta/\partial\theta)(v_\theta^*(y))}{e_\theta + \Delta_\theta'(v_\theta^*(y))}. \quad (8)$$

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 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. Finally, all these effects 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. Conditional 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 occurs at $v = \theta$.[12] Assuming as before that the

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

equilibrium cutoff $v_\theta^*(y)$ is always interior and thus given by (4), it is easily seen that

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

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 that she can achieve credibility.

2. *Shifts affecting the tails.*—Adriani and Sonderegger (2019) extend the analysis in 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.*—Similarly, signaling incentives intensify when the population becomes more diverse in the sense of second-order stochastic dominance. If $\int_{v^{\min}}^{v^{\max}} (\partial F_\theta(v)/\partial\theta)\, dv > 0$ for all $v \in (v^{\min}, v^{\max})$ while $\int_{v^{\min}}^{v^{\max}} (\partial F_\theta(v)/\partial\theta)\, dv = 0$, 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 Welsh 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 (1990–93) 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 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 (2021) 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).[13] 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 and 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

[13] 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 Jia and Persson [2021] assume), however, the sign of $\partial^2 \bar{a}_\theta / \partial y \partial \theta$ is primarily governed by that of Δ''.

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.[14] The principal's objective function is

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

A first interpretation of (10) is that of a *social planner* facing a cost $\lambda \geq 0$ of public funds.[15] 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' equilibrium utilities; $\bar{a}_\theta$ is their total contribution; and $\lambda \bar{a}_\theta y$ is the deadweight loss from the required taxation.[16]

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 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 quasiconcave in y (such is clearly the case provided that λ is

[14] 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 Fischer 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 (2006b), Daughety and Reinganum (2010), Bar-Isaac (2012), and Ali and Bénabou (2020).

[15] 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 affects only the definition of regions in which there is an over- or underprovision of prosocial behavior. That is, we could alternatively assume that $W_\theta^{SI} = \int_{v_\theta^*(y)}^{+\infty} (\epsilon_\theta - c_\theta - \lambda y) f_\theta(v)\, dv$.

[16] 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 $(1 + \lambda) y F_\theta(v^*)$.

small enough)[17] and the equilibrium cutoff interior. To ensure the latter, we restrict the model's parameters to satisfy, for all $\theta \in \Theta$,

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

for some fixed, arbitrarily small $\varepsilon > 0$.[18] 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, $\lambda[1 - F_\theta(v_\theta^*(y))]dy$.

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 $y_\theta^{FB} = \epsilon_\theta - \Delta_\theta((c_\theta - \epsilon_\theta)/e_\theta)$, 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,

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

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,

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

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

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

[18] Condition (11) means that it is socially inefficient (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 n. 6 is satisfied).

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

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 fig. 2).

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 fig. 2). Furthermore, with a positive shadow cost of providing material incentives, there is always underprovision of prosocial behavior: by (12),

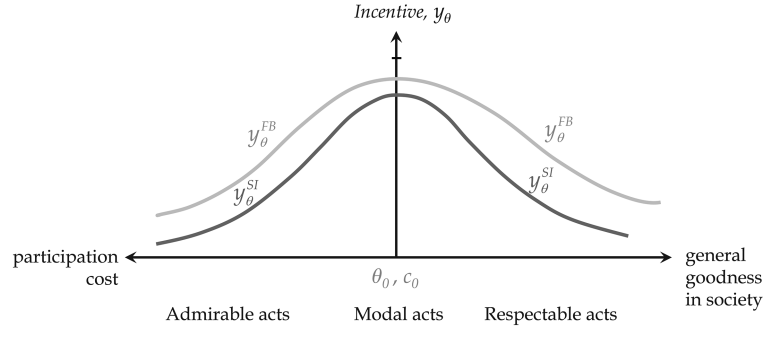

FIG. 2.—Formal incentives, societal values, and participation costs. The two lines plot the optimal incentive rate for each level of the distributional shift θ in societal preferences or the participation cost c under symmetric information between principal and agents. The top line, y_θ^{FB}, corresponds to the first-best situation in which incentives are costless ($\lambda = 0$); the bottom line, y_θ^{SI}, obtains when they are costly.

$$\varepsilon_\theta + v_\theta^*(y_\theta^{SI})e_\theta - c_\theta - \lambda y_\theta^{SI} > 0, \tag{16}$$

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 to contribute, making a relatively low y optimal.[19] 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, while ϵ, c, and v are constant: $F_\theta(v) = F(v - \theta)$, so $y_\theta^{FB} = e - \Delta((c - \varepsilon/e) - \theta)$, 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_θ^{FB} is single peaked with respect to θ and c and maximized at $\theta_0 \equiv c - \varepsilon/e$ and $c_{0,\theta} = \varepsilon + \theta e$, respectively.
ii. For any $\varepsilon > 0$, there exists $\bar{\lambda} > 0$ such that for all $\lambda < \bar{\lambda}$, the symmetric information policy y_θ^{SI} is uniquely defined by (15), strictly increasing for $\theta < \theta_0 - \varepsilon$, and strictly decreasing for $\theta > \theta_0 + \varepsilon$.

Implications.—Here we list some of the implications of proposition 2.

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, and so on. Similarly, agencies computing corporate social responsibility indexes 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

[19] 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.

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, Bracha, and Meier 2009) and toward fair-trade coffee compared with food kitchens.

3. Consider a new environment-friendly technology, such as electric vehicles, that diffuses more widely as its cost c_θ falls because of technological progress. The optimal subsidy rate should first rise and 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 $1/e_\theta$.[20] Incentives are much more effective (the multiplier exceeds $1/e_\theta$) 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),

$$\frac{dy_\theta^{FB}}{d\theta} = \left(1+(\gamma e_\theta + 1 - \gamma)\frac{\Delta_\theta'}{e_\theta^2}\right)\frac{\partial \varepsilon_\theta}{\partial \theta} - \frac{\Delta_\theta'}{e_\theta}\frac{\partial c_\theta}{\partial \theta} - \frac{\partial \Delta_\theta}{\partial \theta}. \tag{17}$$

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_θ^{FB} grows with the size of the externality, decreases with image concerns, and increases (decreases) with the private cost of prosocial behavior in the case of a norm (an antinorm). By continuity, the same properties

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

will hold for the second-best level of incentives y_θ^{SI} 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_θ^{SI} is differentiable, we shall say that condition P^+ (P^-) holds if there exists $\varepsilon > 0$ such that $dy_\theta^{SI}/d\theta > \varepsilon$ ($dy_\theta^{SI}/d\theta < -\varepsilon$) for all θ.

Table 1 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, mean-preserving spread, 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.
- By contrast, when θ measures the externality ϵ_θ (and $\gamma > 0$ so that this affects e_θ), (M^+, P^+) holds: a higher externality intrinsically motivates agents to comply (as long as they are somewhat consequentialist)

TABLE 1
COMPARATIVE STATICS OF INDIVIDUAL MOTIVATION AND OPTIMAL INCENTIVES

	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 mean-preserving spread; Distribution F_θ when θ is a parameter of goodness and $\Delta'_\theta < 0$	Agent's cost ($\theta = c$) when $\Delta' > 0$

NOTE.—The table displays the comparative statics of individual motivation ($-v_\theta^*$) and the optimal incentive rate (y_θ^{SI}) for the different parameters of the economic environment that θ can represent. Each quadrant shows the cases in which an increase in θ is motivation enhancing or reducing (M^+/M^-) and calls for stronger or weaker incentives (P^+/P^-) under symmetric information.

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 $\Delta' > 0$ (antinorm) and noncompliance more respectable when $\Delta' < 0$ (norm), affecting optimal taxation accordingly.

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.[21] Keeping with the normative focus of section III, 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 (SOC): a principal of type θ must not want to induce in agents a belief $\hat{\theta} = \theta$ (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

[21] 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.

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 FOC 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 know only 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 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 noncontributors 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 by 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 unique solution $\hat{\theta}(y) \in [\theta_1, \theta_2]$ to $y_{\hat{\theta}(y)}^{AI} \equiv y$. 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

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

The FOC for maximizing $W_\theta^{AI}(y)$ 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 from (12)—that is, the second term in parentheses on the left-hand side of (19)—thus reflects the way the principal takes into account that (1) agents will draw inferences from her policy choice, as captured by the term $\hat{\theta}'(y) = 1/(y_\theta^{AI})'$, the sign of which is governed by condition P^+/P^-, provided that the incentive varies with θ in the same way for symmetric and asymmetric information, that is, the policy schedules are comonotonic (COM); (2) 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.[22] 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*

Properties of separating equilibria.—Here again, the case of no deadweight loss provides a useful benchmark.

PROPOSITION 3 (Costless incentives.) Let $\lambda = 0$. If the first-best solution $y_\theta^{FB} = \epsilon_\theta - \Delta_\theta(c_\theta - \epsilon_\theta/e_\theta)$ 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 $y = y_\theta^{FB}$.

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 FOC for the optimal policy.

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},

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

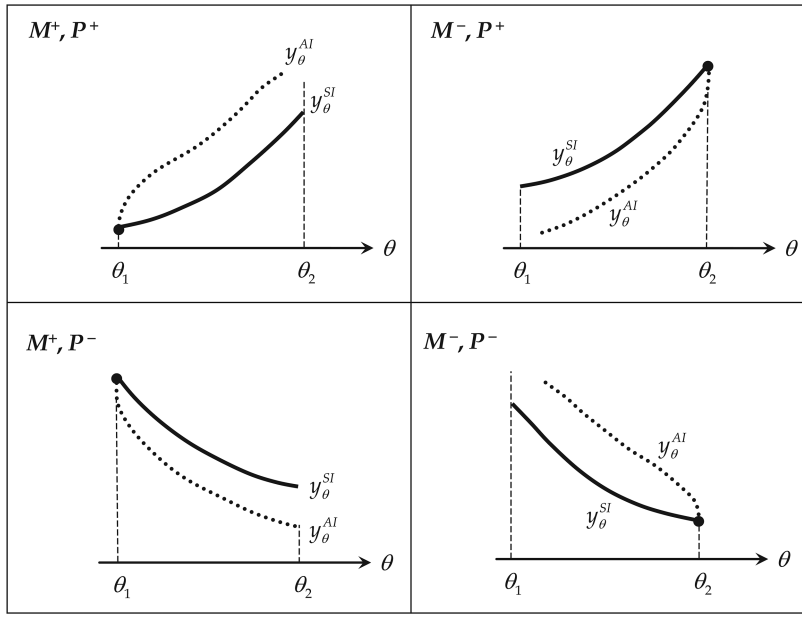

FIG. 3.—Equilibrium incentives under symmetric and asymmetric information. Each quadrant plots the optimal incentive rate as a function of the generic parameter θ. Lines denoted y_θ^{SI} obtain when there is symmetric information between principal and agents about θ; lines denoted y_θ^{AI} obtain when the principal has private information about it. The four quadrants correspond to different cases of what θ can represent: they map into the four quadrants in table 1, in which a higher θ can be either motivation enhancing or motivation reducing (M^+/M^-) and can call for either stronger or weaker incentives (P^+/P^-) under symmetric information.

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 COM with the symmetric information policy: $(y_\theta^{AI})'(y_\theta^{SI})' > 0$.[23]

The next proposition (illustrated in fig. 3) focuses on this solution (thus neglecting the SOC, 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.

[23] The appendix shows that an allocation that satisfies COM as well as no distortion at one of the boundaries necessarily satisfies the NDB at θ_1 under M^+ and at θ_2 under M^-.

PROPOSITION 5 (Determinants of soft or tough law). When the global SOC is satisfied so that the necessary condition (19) indeed defines a separating equilibrium:

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: $b_\theta^{AI} \equiv v_\theta^{*\prime}(y_\theta^{AI})e_\theta + \epsilon_\theta - c_\theta - \lambda y_\theta^{AI} > 0$.

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 (19): 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 (19). This is obvious when M and P are of the same sign; indeed, if b_θ^{AI} were negative, the principal could 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 lemma 3 in the appendix 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 message that "everyone does it, except disreputable people, who suffer substantial stigma; that is why we do not need to provide very strong extrinsic incentives."

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 and credibly signals 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's (1987) 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 nontrivial extension of Mailath's pioneering work. Even under his regularity conditions, for instance, our SOCs need not be satisfied by the solution to the 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, (SOC_1) and (SOC_2), which guarantee that the solution y_θ^{AI} to the 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 $\mathcal{G} \equiv \{y_\theta^{AI}\}_{\theta \in [\theta_1, \theta_2]}$) or 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 (SOCs). Let $\lambda > 0$ be small enough and consider $y(\theta)$, a differentiable and strictly monotonic function satisfying the FOC, COM, and NDB. The condition

$$\mathcal{A}(\theta, \hat{\theta}) \equiv y'(\hat{\theta})\, b(\theta, \hat{\theta}, y(\hat{\theta}))\, \frac{\partial (b(\theta, \hat{\theta}, y(\hat{\theta}))\, h_\theta(v_{\hat{\theta}}^*(y(\hat{\theta}))))}{\partial \theta} \geq 0, \tag{SOC_1}$$

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

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

holds at $\hat{\theta} \in \{\theta_1, \theta_2\}$ for all y.[24]

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 ϵ_θ or participation cost c_θ). In the 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.[25] 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 the FOC satisfying COM and NDB also satisfies (SOC_1) and (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, $\Delta'_\theta < 0$;[26] 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
d. affects the contribution cost c_θ (with $\partial c_\theta/\partial\theta > 0$) and an antinorm prevails, $\Delta' > 0$.

[24] 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.

[25] The four cases correspond to the southwest, northeast, northwest, and southeast quadrants.

[26] 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 ($F_\theta(v^*_\theta(y^{AI}_\theta))$ small enough), $f'_\theta(v^*_\theta(y^{AI}_\theta)) > 0$ as well.

The equilibrium involves soft law ($y_\theta^{AI} < y_\theta^{SI}$) when θ shifts societal values (cases a and 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 antinorm (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 antinorm ($\Delta' > 0$) 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 that 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 antinorm ($\Delta'_\theta > 0$), 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_θ ("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 $\bar{a}_\theta$, ϵ, or μ_θ.[27] 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.[28] 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 $\bar{a}_\theta$.[29] 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 that μ is not too large) agents' equilibrium choices are again defined by a cutoff, denoted $v_\varnothing^*$. Since greater participation increases social welfare, the principal discloses if and only if $v_\theta^* \leq v_\varnothing$. Recalling that v_θ^* is decreasing (increasing) in θ under M^+ (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 $\theta \geq \tilde{\theta}$ ($\theta \leq \tilde{\theta}$) under M^+ (M^-).

[27] When the descriptive and injunctive norms visibly diverge, the former tends to trumps the latter (e.g., Tyran and Feld 2006; Bicchieri and Xiao 2009).

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

[29] 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 that (in the aggregate) has more information about it.

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 $\bar{a}_\theta$ 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 coworkers, 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 in interpreting $\bar{a}_\theta$.[30] 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 (1) when agents know that $\theta = \theta_H$ ($\theta = \theta_L$), there is positive participation, $0 < \bar{a} \leq 1$ (zero participation, $\bar{a}_\theta = 0$); and (2) the prior probability that $\theta = \theta_L$ is high enough that when agents are uninformed, no one contributing is the (generically unique) equilibrium.[31] 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 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 the principal would want to withhold the information (and individuals, if sophisticated, would interpret such silence skeptically).

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

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

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.[32] 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.

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; and so on. While misinformation and special interest considerations 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, and so on—with everyone primarily looking out for themselves (see, e.g., Frey 1997; Bowles 2008; Bowles and Polania-Reyes 2012). This dim view in turn can be very damaging in other nonincentivized 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

[32] 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.

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 (see, e.g., Bénabou and Tirole 2006a; Bénabou 2013). 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, and so on. 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, and so on. 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?). 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 $v_b^* = (c_b - y)/e_b$. 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 nonincentivized 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 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-type 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.[33] These two forces work in opposite directions to determine the set of incentive compatible allocations. The general complexity of SOCs 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 nonincentivized activity are large enough that there is always undersupply of contributions. Then, for λ small enough, soft law prevails: type θ_L setting $y_{\theta_L}^{SI}$ in the controlled activity b and type θ_H setting an appropriate $y_{\theta_H}^{AI} < y_{\theta_H}^{SI}$, together with off-path beliefs $\hat{\rho}(y) = 0$ for all $y > y_{\theta_H}^{AI}$ and $\hat{\rho}(y) = 1$ for all $y \leq y_{\theta_H}^{AI}$, constitutes a least cost separating equilibrium that is robust to D1.

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 antisocial 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),

[33] 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 $\partial^2 \mathcal{W}_a / \partial\theta \partial v_a^* = (\epsilon_a + v_a^* e_a - c_a) f_\theta'$. 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.

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 ϵ_b but also ϵ_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 sec. II.C). By setting a high y_b on b behavior, the principal can then seek to convince agents that the community "will not stand" for a misconduct either but will punish it with stronger social sanctions in lieu of missing or insufficient formal incentives y_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.[34] 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 that 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,

$$\xi(1 - \bar{b}_\theta) \leq \int_{v_a^*(G)}^{v_a^*(\theta_1)} (v_a e_a + \epsilon_a - c_a) f(v - \theta)\, dv,$$

[34] 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 sec. VI.A.2 the second one.

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 $\hat{\theta} = \theta_1$ 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.[35] 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.[36] Such decisions, moreover, are not primarily based on practical considerations of optimal deterrence but on "what it makes us," what civilized people do or do not 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: 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 that 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

[35] 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 US states (internet postings, compulsory lawn signs, license plates, etc.), better satisfy this demand.

[36] 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 that "torture is abhorrent both to American law and values and to international norms" (18 U.S.C. §§ 2340–2340A).

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 drawn from $G(\theta)$ with support $[\theta_1, \theta_2]$. Agents derive utility $\kappa\hat{\theta}$ from their beliefs about θ—society's general aversion to violence or cruelty—because of either collective self-esteem or anticipatory utility with respect to future interactions with others. Normalizing $E_\theta[v] = \theta$, social welfare is[37]

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

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 evaluating 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 $\theta_2 < y$, 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 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.

[37] 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 subunits 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, Daron, and Matthew O. Jackson. 2015. "History, Expectations, and Leadership in the Evolution of Social Norms." *Rev. Econ. Studies* 82 (2): 423–56.

Adriani, Fabrizio, and Silvia Sonderegger. 2019. "A Theory of Esteem Based Peer Pressure." *Games and Econ. Behavior* 115 (C): 314–35.

Alfitian, Jakob, Dirk Sliwka, and Timo Vogelsang. 2024. "When Bonuses Backfire: Evidence from the Workplace." *Management Sci.* 70 (9): 6395–414.
Alger, Ingela, and Jörgen W. Weibull. 2013."Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching." *Econometrica* 81 (6): 2269–302.
Ali, Nageeb, and Roland Bénabou. 2020. "Image versus Information: Changing Societal Norms and Optimal Privacy." *American Econ. J. Microeconomics* 12 (3): 116–64.
Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *J. Public Econ.* 95 (9): 1082–95.
Andreoni, James. 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *J.P.E.* 97 (6): 1447–58.
Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *A.E.R.* 99 (1): 544–55.
Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery." *J. Public Econ.* 120:1–17.
Ayres, Ian, Sophie Raseman, and Alice Shih. 2013. "Evidence from Two Large Field Experiments That Peer Comparison Feedback Can Reduce Residential Energy Usage." *J. Law, Econ., and Org.*. 29 (5): 992–1022.
Bar-Gill, Oren, and Chaim Fershtman. 2004. "Law and Preferences." *J. Law, Econ., and Org.* 20 (2): 331–52.
Bar-Isaac, Heski. 2012. "Transparency, Career Concerns, and Incentives for Acquiring Expertise." *B. E. J. Theoretical Econ.* 12 (1).
Bem, Daryl. 1972. "Self-Perception Theory." In *Advances in Experimental Social Psychology*, vol. 6, edited by L. Berkowitz, 1–62. New York: Academic Press.
Bénabou, Roland. 2013. "Groupthink: Collective Delusions in Organizations and Markets." *Rev. Econ. Studies* 80:429–62.
Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Rev. Econ. Studies* 70 (3): 489–520.
———. 2004. "Willpower and Personal Rules." *J.P.E.* 112 (4): 848–86.
———. 2006a. "Belief in a Just World and Redistributive Politics." *Q.J.E.* 121 (2): 699–746.
———. 2006b. "Incentives and Prosocial Behavior." *A.E.R.* 96 (5): 1652–78.
———. 2011a. "Identity, Morals and Taboos: Beliefs as Assets." *Q.J.E.* 126:805–55.
———. 2011b. "Laws and Norms." Working Paper no. 17579 (November), NBER, Cambridge, MA.
———. 2016. "Bonus Culture: Competitive Pay, Screening, and Multitasking." *J.P.E.* 124 (2): 305–70.
Bernheim, Douglas. 1994. "A Theory of Conformity." *J.P.E.* 102 (5): 842–77.
Besley, Timothy, and Maitreesh Ghatak. 2005. "Competition and Incentives with Motivated Agents." *A.E.R.* 95 (3): 616–36.
Besley, Timothy, Anders Jensen, and Torsten Persson. 2023. "Norms, Enforcement, and Tax Evasion." *Rev. Econ. and Statis.* 105 (4): 998–1007.
Bicchieri, Cristina, and Erte Xiao. 2009. "Do the Right Thing: But Only If Others Do So." *J. Behavioral Decision Making* 22 (2): 191–208.
Bodner, Ronit, and Drazen Prelec. 2003. "Self-Signaling and Self-Control." In *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, edited by G. Loewenstein, D. Read, and R. Baumeister, 277–98. New York: Russell Sage Found.

Bohnet, Iris, Bruno Frey, and Steffen Huck. 2001 "More Order with Less Law: On Contract Enforcement, Trust and Crowding." *American Polit. Sci. Rev.* 95 (1): 131–44.

Bowles, Samuel. 2008. "Policies Designed for Self-Interested Citizens May Undermine 'The Moral Sentiments': Evidence from Economic Experiments." *Science* 320:1605–9.

Bowles, Samuel, and Sandra Polania-Reyes. 2012 "Economic Incentives and Social Preferences: Substitutes or Complements?" *J. Econ. Literature* 50 (2): 368–425.

Brekke, Kjell Arne, Kverndokk Snorre, and Karine Nyborg. 2003. "An Economic Model of Moral Motivation." *J. Public Econ.* 87 (9–10): 1967–83.

Bremzeny, Andrei, Elena Khokhlova, Anton Suvorov, and Jeroen van de Ven. 2015. "Bad News: An Experimental Study on the Informational Effects of Rewards." *Rev. Econ. and Statis.* 97 (1): 55–70.

Brennan, Geoffrey, and Michael Brooks. 2007. "Esteem, Norms of Participation and Public Goods Supply." In *Public Economics and Public Choice*, edited by P. Baake and R. Borck, 63–80. Berlin: Springer.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From Extreme to Mainstream: The Erosion of Social Norms." *A.E.R.* 110 (11): 3522–48.

Bursztyn, Leonardo, Alessandra González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *A.E.R.* 110 (10): 2997–3029.

Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky. 2022. "Measuring the Welfare Effects of Shame and Pride." *A.E.R.* 112 (1): 122–68.

Chen, Daniel. 2016. "The Deterrent Effect of the Death Penalty? Evidence from British Commutations during World War I." Working Paper no. 16-706, Toulouse School Econ.

Cialdini, Robert. 1984. *Influence, the Psychology of Persuasion* New York: William Morrow.

Cooter, Robert. 1998. "Expressive Law and Economics." *J. Legal Studies* 27 (S2): 585–607.

Corneo, Giacomo, and Olivier Jeanne. 1997. "Conspicuous Consumption, Snobbism and Conformism." *J. Public Econ.* 66 (1): 55–71.

Danilov, Anastasia, and Dirk Sliwka. 2017. "Can Contracts Signal Social Norms? Experimental Evidence." *Management Sci.* 63 (2): 459–76.

Daughety, Andrew, and Jennifer Reinganum. 2010. "Public Goods, Social Pressure, and the Choice between Privacy and Publicity." *American Econ. J. Microeconomics* 2 (2): 191–221.

Diamond, Peter. 2006. "Optimal Tax Treatment of Private Contributions for Public Goods with and without Warm Glow Preferences." *J. Public Econ.* 90 (4–5): 897–919.

Ellickson, R. C. 1998. "Law and Economics Discovers Social Norms." *J. Legal Studies* 27 (S2): 537–52.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory." *A.E.R.* 98 (3): 990–1008.

Fehr, Ernst, and Simon Gächter. 2002. "Do Incentive Contracts Undermine Voluntary Cooperation?" Working Paper no. 34, Inst. Empirical Res. Econ., Zurich Univ.

Fehr, Ernst, and Bettina Rockenbach. 2003. "Detrimental Effects of Sanctions on Human Altruism." *Nature* 422:137–40.

Fischer, Paul, and Steven Huddart. 2008. "Optimal Contracting with Endogenous Social Norms." *A.E.R.* 98 (4): 1459–75.

Frey, Bruno S. 1997. *Not Just for the Money: An Economic Theory of Personal Motivation*. Cheltenham: Edward Elgar.
Fryer, Roland. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Q.J.E.* 126 (4): 1755–98.
Funk, Patricia. 2007. "Is There an Expressive Function of Law? An Empirical Analysis of Voting Laws with Symbolic Fines." *American Law and Econ. Rev.* 9 (1): 135–59.
———. 2010. "Social Incentives and Voter Turnout: Evidence from the Swiss Mail Ballot System." *J. European Econ. Assoc.* 8 (5): 1077–103.
Galbiati, Roberto, Emeric Henry, Nicolas Jacquemet, and Max Lobeck. 2021. "How Laws Affect the Perception of Norms: Empirical Evidence from the Lockdown." *PLOS One* 16 (9): 1–14.
Galbiati, Roberto, Karl Schlag, and Joël van der Weele. 2013. "Sanctions that Signal: An Experiment." *J. Econ. Behavior and Org.* 94:34–51.
Galbiati, Roberto, and Pietro Vertova. 2008. "Obligations and Cooperative Behaviour in Public Good Games." *Games and Econ. Behavior* 64 (1): 146–70.
Gibbons, Robert. 1997. "Incentives and Careers in Organizations." In *Advances in Economics and Econometrics: Theory and Applications*, edited by David M. Kreps and Kenneth F. Wallis, chap. 1. Cambridge: Cambridge Univ. Press.
Gneezy, Uri, and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." *Q.J.E.* 115 (3): 791–810.
Greif, Avner, and Steven Tadelis. 2010. "A Theory of Moral Persistence: Crypto-Morality and Political Legitimacy." *J. Comparative Econ.* 38 (3): 229–44.
Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2008. "Social Capital as Good Culture." *J. European Econ. Assoc.* 6 (2–3): 295–320.
Harbaugh, Rick, and Eric Rasmusen. 2018. "Coarse Grades: Informing the Public by Withholding Information." *American Econ. J. Microeconomics* 10 (1): 210–35.
Herold, Florian. 2010. "Contractual Incompleteness as a Signal of Trust." *Games and Econ. Behavior* 68 (1): 180–91.
Huck, Steffen. 1997. "Institutions and Preferences: An Evolutionary Perspective." *J. Inst. and Theoretical Econ.* 153 (4): 771–79.
Jewitt, Ian. 2004. "Notes on the 'Shape' of Distributions." Working paper, Oxford Univ.
Jia, Ruixue, and Torsten Persson. 2021. "Choosing Ethnicity: The Interplay between Individual and Social Motives." *J. European Econ. Assoc.* 19 (2): 1203–48.
Kahan, Dan. 1996. "What Do Alternative Sanctions Mean?" *Univ. Chicago Law Rev.* 63 (2): 591–653.
———. 1997. "Between Economics and Sociology: The New Path of Deterrence." *Michigan Law Rev.* 95 (8): 2477–97.
Kaplow, Louis, and Steven Shavell. 2007. "Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System." *J.P.E.* 115 (3): 494–514.
Karlan, D., and J. A. List. 2007. "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment." *A.E.R.* 97 (5): 1774–93.
Keizer, Kess, Siegwart Lindenberg, and Linda Steg. 2008. "The Spreading of Disorder." *Science* 322 (5908): 1681–85.
Knez, M., and D. Simester. 2001. "Firm-Wide Incentives and Mutual Monitoring at Continental Airlines." *J. Labor Econ.* 19 (4): 743–72.
Lane, Tom, Daniele Nosenzo, and Silvia Sonderegger. 2023. "Laws and Norms: Empirical Evidence." *A.E.R.* 113 (5): 1255–93.
Lefebvre, M., P. Pestieau, A. Riedl, and M.-C. Villeval. 2015. "Tax Evasion and Social Information: An Experiment in Belgium, France, and the Netherlands." *Internat. Tax and Public Finance* 22 (3): 401–25.

Lessig, L. 1998. "The New Chicago School." *J. Legal Studies* 27 (2): 661–91.
Mailath, George. 1987. "Incentive Compatibility in Signaling Games with a Continuum of Types." *Econometrica* 55 (6): 1349–65.
McAdams, R. H. 2000. "A Focal Point Theory of Expressive Law." *Virginia Law Rev.* 86 (8): 1649–729.
McAdams, R. H., and E. B. Rasmusen. 2007. "Norms and the Law." In *Handbook of Law and Economics*, vol. 2, edited by A. M. Polinsky and S. Shavell, 1573–618. Amsterdam: Elsevier.
Miller, Dale, and Cathy McFarland. 1987. "Pluralistic Ignorance: When Similarity Is Interpreted as Dissimilarity." *J. Personality and Soc. Psychology* 53 (2): 298–305.
Pesendorfer, Wolfgang. 1995. "Design Innovation and Fashion Cycles." *A.E.R.* 85 (4): 771–92.
Posner, E. A. 1998. "Symbols, Signals, and Social Norms in Politics and the Law." *J. Legal Studies* 27 (2): 765–98.
———. 2000a. *Law and Social Norms*. Cambridge, MA: Harvard Univ. Press.
———. 2000b. "Law and Social Norms: The Case of Tax Compliance." *Virginia Law Rev.* 86 (8): 1781–819.
Prat, Andrea. 2005. "The Wrong Kind of Transparency." *A.E.R.* 95 (3): 862–77.
Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *J. Econ. Literature* 37 (1): 7–63.
———. 2007. "The Motivation and Bias of Bureaucrats." *A.E.R.* 97 (1): 180–96.
Prentice, Deborah, and Dale Miller. 1993. "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *J. Personality and Soc. Psychology* 64 (2): 243–56.
Rasmusen, Eric. 1996. "Stigma and Self-Fulfilling Expectations of Criminality." *J. Law and Econ.* 39 (2): 519–43.
Rotemberg, Julio. 2008. "Minimally Acceptable Altruism and the Ultimatum Game." *J. Econ. Behavior and Org.* 66 (3–4): 457–76.
Schroeder, Christine M., and Deborah A. Prentice. 2006. "Exposing Pluralistic Ignorance to Reduce Alcohol Use among College Students." *J. Appl. Soc. Psychology* 28 (23): 2150–80.
Schultz, Wesley, Jessica Nolan, Robert Cialdini, Noah Goldstein, and Vladas Griskevicius. 2007. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Sci.* 18 (5): 429–33.
Shavell, Steven. 2002. "Law versus Morality as Regulators of Conduct." *American Law and Econ. Rev.* 4 (2): 227–57.
Sliwka, Dirk. 2008. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *A.E.R.* 97 (3): 999–1012.
Smith, Adam. (1759) 1997. *The Theory of Moral Sentiments*. Washington, DC: Regnery.
Sunstein, Cass. 1996. "On the Expressive Function of Law." *Univ. Pennsylvania Law Rev.* 144 (5): 2021–53.
Tabellini, Guido. 2008. "Institutions and Culture." *J. European Econ. Assoc.* 6 (2–3): 255–94.
Tyran, Jean-Robert, and Lars Feld. 2006. "Achieving Compliance When Legal Sanctions Are Non-Deterrent." *Scandinavian J. Econ.* 108 (1): 135–56.
van der Weele, Joël. 2012. "The Signaling Power of Sanctions in Social Dilemmas." *J. Law, Econ., and Org.* 28 (1): 103–26.
Weibull, Jörgen, and Edgar Villa. 2005. "Crime, Punishment and Social Norms." SSE/EFI Working Paper no. 610, Econ. Res. Inst., Stockholm School Econ.