American Economic Journal: Microeconomics 2020, 12(3): 116–164
https://doi.org/10.1257/mic.20180052

Image versus Information: Changing Societal Norms and Optimal Privacy[†]

By S. Nageeb Ali and Roland Bénabou*

We analyze the costs and benefits of using social image to foster desirable behaviors. Each agent acts based on his intrinsic motivation, private assessment of the public good, and reputational concern for appearing prosocial. A Principal sets the general degree of privacy, observes the social outcome, and implements a policy: investment, subsidy, law, etc. Individual visibility reduces free riding but makes aggregate behavior ("descriptive norm") less informative about societal preferences ("prescriptive norm"). We derive the level of privacy (and material incentives) that optimally trades off social enforcement and learning, and we characterize its variations with the economy's stochastic and informational structure. (*JEL* D82, D83, D91, Z13)

Why Privacy?—

> *If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place.*
>
> Google CEO Eric Schmidt, CNBC (December 8, 2009, interview)

> *The trend toward elevating personal and downgrading organizational privacy is mysterious to the economist. … Secrecy is an important method of appropriating social benefits to the entrepreneur who creates them, while in private life it is more likely to conceal discreditable facts. … The economic case for according legal protection to such information is no better than that for permitting fraud in the sale of goods.*
>
> Judge Richard Posner, "The Right of Privacy" (1978, 401–405)

*Ali: Pennsylvania State University, 411 Kern Building, University Park, PA 16802 (email: nageeb@psu.edu); Bénabou: Princeton University, 286 Julis Romo Rabinowitz, Princeton, NJ 08544, NBER, CEPR, IZA, CIFAR, BREAD, THRED, and BRIQ (email: rbenabou@princeton.edu). Michael Ostrovsky was coeditor for this article. We are grateful for helpful comments to Alberto Alesina, Jim Andreoni, Gabrielle Demange, Navin Kartik, Gilat Levy, Raphael Levy, Alessandro Lizzeri, Kristof Madarasz, David Martimort, Max Mihm, Stephen Morris, Casey Mulligan, Justin Rao, Joel Sobel, Jean Tirole, Pierre-Luc Vautrey, as well as to participants in many seminars and conferences. We also thank three anonymous referees for very valuable suggestions. Edoardo Grillo, Tetsuya Hoshino, Charles Lin, Pellumb Reshidi, and Ben Young provided superb research assistance. Ali gratefully acknowledges financial support from the NSF (SES-1530639). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Bénabou gratefully acknowledges financial support from the Canadian Institute for Advanced Research.

[†]Go to https://doi.org/10.1257/mic.20180052 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

Social visibility is a powerful incentive. When people know that others will learn of their actions, they contribute more to public goods and charities and are more likely to vote, give blood, or save energy. Conversely, they are less likely to lie, cheat, pollute, make offensive jokes, or engage in other antisocial behaviors.[1] Compared to other incentives such as financial rewards, fines, and incarceration, publicity (good or bad) is also extremely cheap. So indeed, following the implicit logic of Google's CEO and a number of scholars, why not publicize all aspects of individuals' behavior that have important external effects, leveraging the ubiquitous desire for social esteem to achieve better social outcomes?

Personal data should of course be protected from the eyes of parties with malicious intent: authoritarian governments, firms tracking consumers' habits to exploit them, hackers and identity thieves, rivals seeking trade secrets, etc. Our concern is with a very different notion of privacy—how much citizens know about each other's behaviors—and with identifying the costs of social transparency arising from *evolving social norms* and the required *adaptation of formal institutions.* As we shall see, these imply that even when the Principal is fully benevolent, incurs no direct cost of publicizing behaviors, and in doing so leads agents to provide needed public goods, it is desirable to maintain a certain degree of "horizontal" privacy. This remains a fortiori true under less ideal conditions.[2]

These issues are of growing policy relevance. Many public and private entities already use esteem as a motivator: the military awards medals for valor, businesses recognize the "employee of the month," and nonprofits publicize donors' names on buildings and plaques. On the sanctions side, many US states and towns use updated forms of the pillory, whereby people arrested for certain offenses (drunk driving, tax or child support delinquency, drug possession) are publicly shamed on television or the internet, and judges sometimes sentence offenders to "advertise" their deeds with special clothing, signs, or newspaper ads. While less common in other advanced countries, such "shaming punishments" are on the rise there as well, as tax authorities, regulators and the public come to perceive the legal system as unable to discipline major tax evaders and rogue financiers.[3]

With advances in "big data," face recognition, automated licence plate readers, and other tracking technologies, the cost of widely disseminating what someone did, gave, took, or even just said is rapidly falling to zero—it is in fact maintaining privacy and anonymity that is becoming increasingly expensive.[4] The trends

[1] On public goods, see, e.g., Ariely, Bracha, and Meier (2009); Linardi and McConnell (2011); DellaVigna, List, and Malmendier (2012); Ashraf, Bandiera, and Jack (2014); or Algan et al. (2013). On voting, see Gerber, Green, and Larimer (2008), and on blood donors, see Lacetera and Macis (2010).

[2] For a survey of privacy issues involving actors who seek to misuse individuals' data, see Acquisti, Taylor, and Wagman (2016). We shall also abstract (for similar reasons) from concerns that public shaming constitutes "cruel and unusual punishment," negating important societal values such as human dignity; see, e.g., Posner (1998) for discussions and Bénabou and Tirole (2011) for a formal analysis of expressive law.

[3] In Peru, businesses convicted of tax evasion can be shut down, with a sign plastered on the door; conversely, municipalities publish an "honor list" of households who have always paid their property taxes on time (Del Carpio 2013). Shaming is also often organized by activists, as with the Occupy Wall Street and #metoo movements.

[4] A flourishing image-ransoming industry is even developing in the United States. These "shame entrepreneurs" operate by reposting on high-visibility websites the official arrest "mugshots" from police departments all across the country, then asking the people involved for a hefty fee in order to take them down (Segal 2013).

described above are therefore likely to accentuate, whether impulsed by public authorities, activist groups, or individual whistleblowers.[5]

A number of scholars in law, economics, and philosophy have in fact long argued for a more systematic recourse to public marks of honor (Cooter 2004, Brennan and Pettit 1990, 2004, Frey 2007) and shame (Kahan 1997, Kahan and Posner 1999, Reeves 2013, Jacquet 2015) on grounds of both efficiency and expressive justice. R. Posner (1978, 1979) carries this logic the furthest, arguing that people should have essentially zero property rights over facts concerning them, whatever their nature (e.g., sexual behaviors, religious or political opinions, decades-old offenses, or medical conditions), and no right not to self-incriminate.[6]

There remains, however, substantial unease at the idea of shaming as a policy tool and, more generally, a widespread view that a society with zero privacy is undesirable. Since the foundational article of Warren and Brandeis (1890), a broad right of privacy has progressively been enshrined in most countries' constitutions, though its practical content varies. Besides the attachment to anonymous voting as indispensable to democracy, there are many instances where social institutions preserve privacy even though publicity could help curb free riding and other socially disapproved behaviors.[7] Is there, then, contra the earlier quotations, an "*economic case for according legal protection to such information*"?

Ideas and Framework.—The paper's objective is threefold. First, we develop a tractable framework in which the interplay of *social norms* and *social learning* can be studied. We build here on Bénabou and Tirole (2006), to which we add both individual and aggregate preference uncertainty. Agents thus choose their actions anticipating that they will be assessed against an endogenous social norm that is yet to emerge from their collective behavior, and that they must therefore try to forecast. Second, we further expand this framework to study the *costs and benefits of privacy* in a society where preferences may be *changing* unpredictably, and how a (benevolent or selfish) Principal should optimally set its level. Third, we take up the novel problem of how *monetary and reputational incentives* should jointly be used, and fully solve for the optimal policy mix.

[5] China is implementing a "Social Credit System" in which most behaviors of every citizen, firm, and public entity will be rated, with scores conditioning access to credit and private and public services (housing, travel, welfare), as well as public praise and shame lists, in order to build a nationwide "culture of sincerity." It builds on the Alibaba/Ant Financial Group's "Sesame Credit" system, in which "higher scores have already become a status symbol, with … people bragging about their scores on Weibo. … A citizen's score can even affect their odds of getting a date, or a marriage partner" (Botsman 2017).

[6] In this view, market forces will ensure that mistakes and "irrational" discriminations are quickly eliminated, leaving only efficient uses of the information that create reputational incentives for socially beneficial behavior. One could surely dispute the first assumption, but our purpose lies instead in finding rationales for privacy that do not rely on the presence of observational mistakes, irrational inferences, or expectational coordination failures.

[7] During episodes of energy or water rationing, local authorities do not publish lists of overusers (the media, on the other hand, often reports on the most egregious cases). In publicly funded health care, there is no policy to "out" those who impose high costs through behaviors such as smoking, poor diet, or addiction. On the contrary, there are strong legal protections for patient confidentiality. Governments often expunge criminal records after some time or conceal them from private view (e.g., prohibiting credit bureaus from reporting arrest records), and a major debate over the "*right to be forgotten*" is ongoing with search engine and social media companies.

The paper's underlying theme is that while publicity is a powerful and cheap instrument of control, it is also a blunt one, generating substantial uncertainty both for those *subject to it* and for those who *wield it*. This involves two parallel mechanisms:

(i) *Variability in the Power of Social Image*.—The rewards and sanctions generated by publicizing someone's actions stem from the reactions that this elicits from his family, peers, neighbors, customers, etc., making their severity hard to predict and fine-tune (Whitman 1998, E. Posner 2002). Depending on a host of circumstances, it can range from mild ostracism to mob action, be easy or hard to avoid, etc.[8] Variability in concerns about social sanctions, in turn, generates inefficient variations in behavior, which become amplified as individuals' actions are made more visible or salient.[9]

(ii) *Misguided Policies*.—Because societal preferences constantly *evolve*, legislators and courts must keep learning how policies and institutions should be adapted from prevailing mores and "community standards." Overt racism, sexism, and domestic violence went from "normal" to deeply scorned within a decade or two, while divorce, cohabitation ("living in sin"), and homosexuality shifted from intensely stigmatized to broadly acceptable. Inevitably, *some* conducts seen today as abhorrent will become mundane over time, and vice versa.[10] If low privacy makes people too worried about social stigma, they will distort their actions, and these shifts will remain obscured, resulting in rigidification and maladaptation not just of *private conduct* but also of *public policy*. This learning issue remains even for behaviors that remain unambiguously bad or good from a social point view (drunk driving, tax compliance, etc.): as long as the *relative* importance of different public goods for social welfare evolves, policymakers will need to appropriately redirect limited financial or enforcement resources.

Formally, we consider a Principal interacting with a continuum of agents in a canonical context of public-goods provision or externalities. Agents have private signals about the quality of some public good, corresponding to a common-value setting, or alternatively they derive private values from the prevalence of some behavior in society. Each individual chooses how to act based on his own mix of public spiritedness, information about the social value of the conduct in question, and concern for appearing prosocial. The Principal can amplify or dampen reputational payoffs by making individual contributions more or less visible to the community. While this entails little direct cost (none, for simplicity), two sources of losses arise in equilibrium. The first is an excessive image-driven *variance* in contributions, both across individuals and in total. The second and most important

[8] On instability and indeterminacy in collective action outcomes, see Lohmann (1994) and Kuran (1997). The explosive growth of shaming on social media is a good example of this variability, with the ultimate costs to the punished party (loss of job and family, suicide) wildly disproportionate to the perceived offense (Ronson 2015).

[9] Similar inefficiencies occur if social sanctioning involves (convex) resource costs or agents are risk averse.

[10] Uncertainty lies only in which ones (e.g., blood and organ sales, prostitution, consuming animals, gestational surrogacy, human enhancement) and which way the cursor will move. For trends, see Elías et al. (2017).

one is an *information distortion* in the Principal's own inference problem. Because societal preferences change, she only imperfectly knows the social value of the public good (and the importance attached by agents to social esteem or sanctions), which is critical for choosing her own contribution, matching rate, or legal policy. Hence, there is a trade-off between using *image as an incentive* and gaining better *information* on societal preferences: if she dampens signaling motivations, she can reliably infer societal preferences from agents' aggregate behavior, but people will free ride substantially, leaving her most of the burden of public good provision. Conversely, if she leverages social image to spur compliance, she exacerbates her own signal extraction problem by making collective behavior more reflective of variations in the importance of social payoffs.[11]

We analyze these trade-offs and derive the optimal privacy and contributions-matching policies for the Principal. We then derive comparative statics predictions with respect to all the economy's primitives, especially the various sources of individual and aggregate uncertainty. Finally, we show that the same trade-offs emerge (and remain solvable) even when the Principal can combine reputational and material incentives, or when she uses observed social norms to inform a subsequent choice of legal mandates, taxes, or subsidies.

Main Applications.—

Public Goods Provision.—We cast the basic model in terms of a classical benchmark: providing the "right kind" of public goods in a cost-effective manner. Community leaders, philanthropists, and foundations often rely on constituents' and activists' degree of involvement to identify the value of investing in schools, parks, transportation, or development projects in remote countries.[12] Publicly recognizing and honoring the efforts of individuals or NGOs encourages commitment, but also makes it a less precise signal of true social value. Governments can similarly learn about the public's perceived *legitimacy* of certain mandates or prohibitions by observing the general level of (non)compliance: alcohol or drug use, electoral turnout, tax evasion or "morale," etc.

In the private sphere, firms are pressured or shamed by activists, and consumers by each other, into taking "social responsibility" for the environment, fair trade, workplace safety, animals' welfare, etc. To the extent that such reputational incentives make up for deficient regulation or taxation, they are beneficial, but the strong signaling effects they create make it hard to know which practices are truly socially valuable and which are just "greenwashing."

[11] The point applies more generally to any incentive to which agents respond strongly on average (effectiveness), but to a degree that is hard to predict ex ante and parse out ex post (uncertainty). As discussed earlier, this is much more a feature of social sanctions than of monetary incentives, on which many trade-offs are observable. Thus, it is arguably easier to estimate a stable response of tax compliance to fines and audit probabilities than to posting the names of evaders on a shame list. Absent such an asymmetry between formal and informal incentives, our model provides further reasons why high-powered incentives of any kind can be counterproductive.

[12] This is also why the practice of matching individual contributions is common among sponsors, as are "leadership" gifts used as signals of worth for subsequent donors (Vesterlund 2003, Andreoni 2006).

From Social Norms to Formal Institutions.—Laws and institutions most often crystallize from preexisting community standards, norms, and practices, which inform designers about what behaviors are generally deemed to generate positive or negative externalities. These views change over time, sometimes quite radically. Consider, for instance, the opinion of the Supreme Court legalizing same-sex marriage:

> Under the centuries-old doctrine of coverture, a married man and woman were treated by the State as a single, male-dominated legal entity. ... As women gained legal, political, and property rights, and as society began to understand that women have their own equal dignity, the law of coverture was abandoned. … Changed understandings of marriage are characteristic of a Nation where new dimensions of freedom become apparent to new generations. … Well into the twentieth century, many States condemned same-sex intimacy as immoral, and homosexuality was treated as an illness. Later in the century, cultural and political developments allowed same-sex couples to lead more open and public lives. Extensive public and private dialogue followed, along with shifts in public attitudes. … The Court, in this decision, holds same-sex couples may exercise the fundamental right to marry in all States.
>
> Supreme Court of the United States, Obergefell v. Hodges, 576 U.S. 644 (2015)

The latest examples of evolutions in social practices informing subsequent changes in the law include cannabis consumption, physical punishment of children, and sexual harassment. Where behavior is highly constrained by the fear of social stigma, conversely, assessing social preferences by what people do—the "*descriptive norm*"—is a poor indicator of what they really value—the "*prescriptive norm*"; laws and other policies will lag far behind societal values.[13]

The first class of applications above correspond best to a *common-value* setting, involving some general public good of imperfectly known quality. For public goods with heterogenous incidence, and for many societal norms falling into the second class, a more appropriate setting is that of *private values*, in which people differ in their fundamental attitudes toward some externality-generating behavior. Our analysis will integrate both cases.

Road Map.—The rest of this section discusses related literature. Section I presents the core model. Section II derives, for any given level of visibility, agents' equilibrium behavior and a benchmarking result for social inferences. Section III analyzes the Principal's resulting trade-offs, optimal publicity level, and contributions-matching rate; Section IV then characterizes their full comparative statics. Section V extends all results to the optimal mix of monetary and social incentives. Section VI applies the model to the design of fixed mandates and

[13] Related issues arise in the debate over freedom of speech versus "political correctness" (Loury 1994, Morris 2001), with activists commonly using publicity to curtail "offensive" acts and words.

marginal incentives. Section VII concludes. Proofs are in the Appendix, extensions in an online Appendix.

Related Literature.—Several lines of work examine the impact of transparency on individual and collective decision-making. A first strand focuses on signaling, especially in a public goods context.[14] Our model builds on Bénabou and Tirole (2006), who study how extrinsic incentives can undermine the reputational returns derived from a prosocial activity. We develop this framework in three directions novel to the literature. First, a Principal explicitly chooses how much agents know about each other's behavior, internalizing their equilibrium responses. Second, both agents and Principal are imperfectly informed about the social value of the activity, generating a social-learning problem for the former and a trade-off between image incentives and information aggregation for the latter. Third, the Principal may optimally combine standard material incentives and reputational ones—a feature unique to this paper.

Signaling or career concerns often lead agents to exert wasteful effort (e.g., Holmström 1999). Relatedly, Daughety and Reinganum (2010) shows how making actions fully public can result in the overprovision of public goods, whereas making them fully private can result in underprovision. The mechanisms we explore differ from these in several important ways. First, there is no excess effort or investment: in equilibrium, the social marginal value of contributions is always positive. Second, there is an optimizing Principal who continuously adjusts how much privacy to accord individuals, faces uncertainty about how they will respond to it, and cares not about who does what, but about the informational content of the collective behavior.

Transparency is also a central issue when experts, judges, or committee members have reputational concerns over the quality of their information, as they may distort their advice or actions in order to appear more competent. A first effect, working toward conformity or "conservatism," arises when agents have no private knowledge of their own ability: they will then make forecasts and choices that aim to be in line with the Principal's prior (Prendergast 1993, Prat 2005, Bar-Isaac 2012) or those of more knowledgeable "senior" agents (Ottaviani and Sørensen 2001), or to indicate a broad consensus (Visser and Swank 2007, Swank and Visser 2013).[15] When competence is a private type, on the other hand, the incentive to signal it generates "anticonformist" or activist tendencies: agents will overreact to their own signals, reverse precedents, etc.; which of the two forces dominates then depends on the details of the game's information and strategic structure (Levy 2005, 2007).[16] In our framework, agents' incentives to signal their types simultaneously have positive (mean contribution) and negative (excessive variance and information

[14] See, e.g., Bernheim (1994), Corneo (1997), Harbaugh (1998), Ellingsen and Johannesson (2008), and Andreoni and Bernheim (2009).

[15] In similar spirit, Swank and Visser (2015) examine whether agents' reputations should be assessed "locally" (based on each one's own action) or "globally" (relative to the actions of others).

[16] On the normative side, whether the Principal prefers (full) transparency or (full) anonymity for the agents turns on how her loss function weighs "getting things wrong" in more likely states of the world versus more rare ones (Fox and Van Weelden 2012, Fehrler and Hughes 2018).

garbling) effects. A fundamental difference is that the strength of image motives, which is common knowledge in nearly all of the signaling and career concerns literatures, is here one of the key sources of uncertainty.[17] In emphasizing how laws emerge from evolving social norms, finally, we relate to a growing literature on how formal and informal institutions shape each other (Bénabou and Tirole 2011; Jia and Persson 2017; Besley, Jensen, and Persson 2014; Acemoglu and Jackson 2017).

While also concerned with the broad issue of information aggregation, our mechanism is entirely distinct from that of "global games" or expectations coordination models (e.g., Morris and Shin 2002 and subsequent literature). That entire line of work centers on how the availability of a public signal leads private agents to put too little weight on their own information when choosing their actions, resulting in informationally inefficient herding and, if there are coordination externalities, social welfare losses. In our model: (i) There is *no strategic interdependence* in payoffs. (ii) Individuals act based *solely on their private information* (which is multidimensional) before observing any public variable. The equilibrium norm $(\bar{a})$ is learnt only afterward, and what matters is not the "macro" visibility of this or any other statistic but instead the "micro" visibility of personal, *idiosyncratic* choices. (iii) This degree of individual privacy (x) has no direct impact on the aggregate signal, and in equilibrium it *reduces* its precision; by contrast, the above literature is about the effects of exogenous increases in common knowledge. (iv) What is essential there is that agents *observe* an *actual* common signal, whereas here it is that each one *thinks*, rightly or wrongly, that he *might* be *observed* by others.

I. Model

We study the interaction between a continuum of small agents $\left(i \in [0,1]\right)$ and a single large Principal (P), each of whom chooses how much to contribute (in time, effort, or money) to a public good.[18] Depending on the context, these actors may correspond to (i) a government and its citizens; (ii) a charitable organization and potential donors; or (iii) a profit-maximizing firm and workers who care to some degree about how well it is doing, whether out of pure loyalty or because they have a stake in its long-run survival.

[17] Bénabou and Tirole (2006) study signaling agents with heterogenous (privately known) image concerns, and Fischer and Verrecchia (2000) and Frankel and Kartik (2019) study agents with heterogenous payoffs to misrepresenting their actions. In such settings, greater visibility makes each individual's observed behavior less informative about his true motivations. In none of these papers is there any aggregate uncertainty (hence also no social learning thereof), nor a Principal who seeks to incentivize agents through publicity (and payments) and learn from their behavior to make her own decision (investment, tax or subsidy, law, etc.).

[18] We shall follow the common practice of applying the law of large numbers to a continuum of i.i.d. random variables. This is known (e.g., Judd 1985) to be inappropriate when working within the usual probability space on realizations of a continuum of draws. Several solutions to the underlying nonmeasurability problems exist, however. Uhlig (1996) shows that it suffices to redefine all integrals over the continuum as L_2-Riemann integrals. Al-Najjar (2004) shows how to approximate the continuum model with finitely additive distributions on a countable set of agents. Sun (2006) shows that one can extend the usual product space to what are called (rich) Fubini extensions, in which (essentially pairwise) independence ensures that all law of large numbers and sample distribution equalities hold exactly.

Agents.—Each agent i selects a contribution level $a_i \in \mathbb{R}$, at cost $C(a_i) \equiv a_i^2/2$. An individual's utility depends on his own contribution, from which he derives some intrinsic satisfaction (or "joy of giving"), on the total provision of the public good, which has quality or social usefulness indexed by θ, and on the reputational rewards attached to contributing. Given total private contributions $\bar{a}$ and the Principal contributing a_P, Agent i's direct (nonreputational) payoff is

$$(1) \qquad U_i(v_i, \theta, w; a_i, \bar{a}, a_P) \equiv (v_i + \theta)\, a_i + (w + \theta)(\bar{a} + a_P) - C(a_i).$$

The first term corresponds to his *intrinsic motivation*, which includes both an idiosyncratic component v_i and the common shift factor θ, reflecting the idea that people like to contribute more to socially valuable projects than to less useful ones. Agent i's baseline valuation v_i is distributed as $N(\bar{v}, s_v^2)$ and privately known to him. The second term in (1) is the *value derived from the public good*, which we take to be similar across individuals, without loss of generality. We assume $\bar{v} < w$, ensuring that intrinsic motivations alone do not solve the free rider problem, namely are insufficient to equalize the (average) marginal values of contributing and receiving a unit of public good.

The quality or social value of the public good is a priori uncertain, with agents and the Principal starting with a common prior belief that θ is distributed as $N(\bar{\theta}, \sigma_\theta^2)$. Each agent i receives a private noisy signal, $\theta_i \equiv \theta + \epsilon_i$, in which the error is distributed as $N(0, s_\theta^2)$ independently of the signals of others.[19] Here and throughout the paper, we use the following mnemonics: *aggregate* variabilities are denoted as $\sigma_\cdot^2$, *cross-sectional* dispersions as $s_\cdot^2$.

Each agent also cares about the inferences that members of his social and economic networks will draw about his intrinsic motivation, v_i: he wishes to appear prosocial, a good citizen rather than a free rider, dedicated to his work, etc.[20] The value of reputation varies across individuals, communities, and periods; it is greater, for instance, where people are engaged in long-run relationships based on trust than where exchange occurs through impersonal markets and complete contracts. Social enforcement also relies on the intensity and contagion of emotional responses (e.g., via social media) and on offenders' vulnerability to them, generating further individual and aggregate variability. We denote the strength of agent i's reputational concerns as μ_i (specifying below how it affects his payoffs) and allow it to be distributed cross-sectionally as $N(\mu, s_\mu^2)$ around the group average μ, which itself varies as $N(\bar{\mu}, \sigma_\mu^2)$ around a common prior $\bar{\mu}$ held by agents and Principal alike. We assume that $\bar{\mu}$ is large enough that with very high probability, the fraction of agents who desire a positive reputation is close to 1.

Formally, an agent i's complete type is a triplet (v_i, θ_i, μ_i), where for tractability we take the three components to be mutually independent, and his contribution a_i is (or might be) observed by a representative sample of the population.[21] An agent j

[19] Alternatively, each may have his own genuine valuation θ_i for it: see Section IIIE for this private values-case.

[20] These concerns may be instrumental (appearing as a more desirable employee, mate, business partner, or public official), hedonic (feeling pride rather than shame, basking in social esteem), or a combination of both.

[21] If μ_i was correlated with v_i or θ_i, the inference problems of agents and Principal would become nonlinear.

observing i's choice of a_i does not know to what extent it was motivated intrinsically (high v_i), by a high signal about the value of the public good (high θ_i), or by a strong image motive (high μ_i). He can, however, use his own signal θ_j and reputational concern μ_j (since (θ_i, θ_j) and (μ_i, μ_j) are correlated), as well as the realized average contribution $\bar{a}$, to form his assessment $E[v_i | a_i, \bar{a}, \theta_j, \mu_j]$ of player i. Thinking ahead, agent i uses his ex ante information to forecast how he will be judged by others. The average *social image* that he can anticipate if he contributes $a_i = a$ is thus

$$R(a, \theta_i, \mu_i) \equiv E_{\bar{a}, \theta_{-i}, \mu_{-i}}\left[\int_0^1 E[v_i | a, \bar{a}, \theta_j, \mu_j]\, dj \,\middle|\, \theta_i, \mu_i\right]. \tag{2}$$

We assume that a social image $R(a, \theta_i, \mu_i)$ yields for agent i a (normalized) payoff of $\mu_i x[R(a, \theta_i, \mu_i) - \bar{v}]$, where μ_i reflects his baseline concern for social esteem and $x \geq 0$ parametrizes the degree of visibility and memorability of individual actions, which can be exogenous or under the Principal's control. Accounting for both direct and image-based payoffs, agent i chooses a_i to solve

$$\max_{a_i \in \mathbb{R}} \{E[U_i(v_i, \theta, w; a_i, \bar{a}, a_P) | \theta_i] + x\mu_i[R(a_i, \theta_i, \mu_i) - \bar{v}]\}. \tag{3}$$

Principal.—The Principal's objective function is a convex combination of agents' total utility and her private payoffs from the overall supply of the (quality-adjusted) public good:

$$\begin{aligned} V(\bar{a}, a_P, \theta) \equiv \lambda\Big[&(w + \theta)(\bar{a} + a_P) - \int_0^1 C(a_i)\, di \\ &+ \alpha \int_0^1 (v_i + \theta) a_i\, di + \tilde{\alpha} \int_0^1 x\mu_i [R(a_i, \theta_i, \mu_i) - \bar{v}]\, di\Big] \\ &+ (1 - \lambda)[b(w + \theta)(\bar{a} + a_P) - k_P C(a_P)]. \end{aligned} \tag{4}$$

The first line captures agents' standard costs and benefits from public goods provision. In the second line, $\alpha \in [0, 1]$ measures the extent to which the Principal internalizes their intrinsic "joy of giving," and $\tilde{\alpha}$ that to which she internalizes their gains and losses in social image. In the last line, k_P is the Principal's cost of directly contributing, relative to that of agents, while $b \in \mathbb{R}$ represents any private benefits she may derive from the total supply of public good. It will be useful to denote

$$\varphi \equiv \lambda + (1 - \lambda) b, \tag{5}$$

$$\omega \equiv (w + \bar{\theta})\varphi - \lambda(1 - \alpha)(\bar{v} + \bar{\theta}). \tag{6}$$

The coefficient φ is the Principal's total gain per (efficiency) unit added to the total supply of public good $\bar{a} + a_P$, whatever its source. The coefficient ω is her *net expected utility* from each marginal unit of the good provided specifically by the agents, taking into account that when $\lambda > 0$, she internalizes (i) a fraction $\lambda\alpha$ of their intrinsic satisfaction from doing so and (ii) a fraction λ of their marginal

contribution cost $\int_0^1 C'(a_i)di = \bar{a}$, which, absent reputational incentives, they would equate to their intrinsic marginal benefit, $\bar{v} + \theta$.

Put differently, ω represents the *wedge* between the Principal's *expected value* of agents' contributions and the latter's *expected willingness* to contribute spontaneously. To make the problem nontrivial, we shall assume that $\omega > 0$: on average, the Principal wants to increase private contributions (or norm compliance). To cut down on the number of cases, we shall focus the exposition on that where $b > 0$, which in turn implies that $\varphi > 0$ and $\partial\omega/\partial\bar{\theta} = \lambda\alpha + (1-\lambda)b > 0$: ex ante, the Principal's preferences over the quality of the public good are congruent with those of the agents, even though her preferences over the level and sharing of its supply may be very different.[22]

Our framework includes as special cases:

(i) For $\lambda = \alpha = \tilde{\alpha} = 1$, a purely altruistic, "selfless" Principal.

(ii) For $\lambda = 1/2$ and $b = \alpha = \tilde{\alpha} = 0$, a standard social planner who values equally agents' and her own costs of provision. The latter could also be those incurred by the rest of society, e.g., due to a shadow price of public funds.

(iii) For $\lambda = 0$, a purely selfish Principal, such as a profit-maximizing firm that uses image to elicit effort from its employees.

To set her own provision a_P efficiently, the Principal must learn about θ. A key piece of data that she observes is the aggregate contribution or *compliance* rate $\bar{a}$, which embodies information about both aggregate shocks, θ and μ, generating a signal-extraction problem. The Principal shares agents' prior $\theta \sim N(\bar{\theta}, \sigma_\theta^2)$ about the quality of the public good and obtains an independent signal $\theta_P \sim N(\theta, s_P^2)$. Her prior for the importance of image is $N(\bar{\mu}, \sigma_\mu^2)$. These beliefs incorporate all the information previously obtained by the Principal, for instance by polling agents about the quality of the public good or the importance of social image.[23]

Timing.—The game unfolds as follows:

(i) The Principal chooses the level of observability of individual behavior, x, that will prevail among agents. Conversely, $1/x$ represents the degree of *privacy.*

(ii) Each agent learns his private type (v_i, θ_i, μ_i), then chooses his contribution a_i.

(iii) The aggregate contribution $\bar{a}$ is publicly observed.

[22] The model and all analytical results also allow for $b < 0$ (even potentially $\varphi < 0$, $\omega < 0$ and $\partial\omega/\partial\bar{\theta} < 0$), however. This corresponds to a Principal who intrinsically *dislikes* an activity that most agents consider socially appropriate: political opposition, cultural resistance, racial or sexual discrimination, etc.

[23] This information is typically limited: polling is costly (see Auriol and Gary-Bobo 2012 on the optimal sample size or number of representatives) and also invites strategic responses from agents who would like to influence the Principal's policy (Morgan and Stocken 2008; Hummel, Morgan, and Stocken 2013) or are wary of revealing "discreditable" preferences—as recent electoral outcomes in the United Kingdom and United States have clearly shown. Allowing the Principal to obtain an independent, noisy signal of μ would also not affect our analysis.

(iv) The Principal observes her own signal θ_P.

(v) The Principal chooses her contribution a_P, and the total supply $\bar{a} + a_P$ is enjoyed by all.

We shall focus, for tractability, on Perfect Bayesian Equilibria in which an agent's contribution is linear in his type, (v_i, θ_i, μ_i).

A. *Discussion of the Model*

At the core of our model are two related tensions between the benefits of publicity—on average, it improves the provision of public goods and economizes on costly incentives—and the distortions it generates in agents' and the Principal's decisions:

(i) Agents' contributions become driven in larger part by variations in their image concerns rather than by their signals about the social value of the public good (Section II).

(ii) A Principal who does not precisely know the extent to which agents care about social payoffs must use publicity carefully, lest it make their behavior excessively image driven—that is, too uncorrelated with the true quality of the public good and hence too difficult for her to learn from (Section III).

To identify these forces as cleanly as possible, we made a number of specific assumptions.

Private versus Common Values.—In the benchmark specification, agents' ex post payoffs from contributing to and consuming the public good reflect some objective, universally agreed-upon quality or social impact θ. This corresponds to a setting with *common values*, such as monetary contributions or productive efforts over which people's preferences are aligned but their signals differ. The model equally applies, however, to the case of *private values*, in which agents fundamentally disagree over some externality-generating behavior, deriving heterogeneous payoffs from it even under full information.[24]

This flexibility is essential to cover the full range of applications discussed earlier, from traditional public goods to pure "social mores." The latter indeed involve conducts from which people typically experience very different externalities—whether physical (smoking, drug use), socioeconomic (working women, single parenthood), psychological (sexist or racist comments), or even purely moral: "offensive to my values, sacrilegious, sowing hatred and division, demeaning to human dignity," etc. In such settings, θ_i reflects agent i's *subjective* (dis)utility from

[24] This distinction is similar to the one discussed within the context of global games (Carlsson and van Damme 1993, Morris and Shin 2006), but here reputation also comes into play. Thus, in contrast to θ_i, v_i is a general degree of prosociality or other-regard, and therefore still the trait over which reputations are formed.

the externality, s_θ^2 measures the *dispersion of attitudes* (societal disagreement), and the policy-setting Principal cares about the *average preference* $\theta = E[\theta_i]$, as reflecting aggregate shifts in the distribution of private tastes. Section IIIE shows that the private-values case maps into a simple subcase of the common-values model, so our main analysis centers on that more encompassing specification.

Separability in Intrinsic Motivation and Quality.—The model features multidimensional signaling with a single-dimensional action space, which leads to pooling between types with high intrinsic motivation v_i, favorable information or private value θ_i, and strong image concerns μ_i. Moreover, each agent lacks information about others' signals and so cannot perfectly anticipate how they will interpret his actions. Social incentives thus involve both multidimensional heterogeneity and higher-order uncertainty, making the problem a complex one. Specifying agents' preferences as separable in intrinsic motivation and public good quality allows us to keep it tractable and derive simple, closed-form solutions. The basic trade-off between incentives and information would, however, apply even with complementarity between these dimensions.

Formalizing Publicity.—A Principal's influence on the visibility of agents' actions can operate through many channels: the probability and/or precision with which these are observed, their moral salience, the number of people who observe them, the time they remain "on the record," and even the social payoffs attached to image—e.g., how much "popular justice" or discrimination against noncompliers is tolerated or encouraged. The specification $x\mu_i$ allows for maximal flexibility as to the channels involved (in particular, the effect is potentially unbounded), so that limits on x will emerge solely from the Principal's optimal choice.

Alternatively, one may conceptualize "privacy" as the noise with which each agent's action is observed by others. Specifically, suppose that when i contributes a_i, others observe $\hat{a}_i = a_i + \epsilon_i$, where $\epsilon_i \sim N\left(0, s_\varepsilon^2/x^2\right)$ and x is chosen by the Principal. We analyze this alternative case in the online Appendix and show that all results remain unchanged.

Timing of Information and Publicity.—Having the Principal first set the degree of publicity and then observe her signal θ_P allows us to abstract from an "Informed Principal" problem. Were the timing reversed, her choice of x would convey information about the quality of the public good, which is a different strategic force from that of interest here.[25] The choice of publicity/privacy would then also commingle the Principal's motive to learn from agents with her incentive to signal to them.

Principal's Other Policies.—The Principal chooses her level of provision a_P after agents make their decisions, but all results are identical when she commits in

[25] Papers studying an informed-Principal problem in related contexts include Bénabou and Tirole (2003), Sliwka (2007), Van der Weele (2012), and Bénabou and Tirole (2011), who study how *laws shape norms*, whereas we focus here on the complementary mechanism of how *norms shape laws*.

advance to a *matching rate* on private contributions. This invariance reflects the fact that each a_i is negligible in the aggregate, together with the assumption (implicit in how a_P enters (1)) that agents derive intrinsic utility only from their own contribution and not from the induced matching.[26]

To focus squarely on the effects of publicity, we initially abstract from the use of standard material incentives to induce compliance.[27] In Sections V and VI, we then allow the Principal to combine visibility with either *legal regulations* (quantity mandates) or costly *monetary incentives*, and to do so either ex ante (before observing equilibrium compliance) or ex post (once she has observed and learned from it). All key results and comparative statics remain unchanged.

Simple Dynamics.—Our model highlights the effects of aggregate variability and idiosyncratic differences in societal preferences on agents' behavior and the optimal decisions of a Principal. While the model is a static one, we occasionally interpret the results in dynamic terms—a fast- or slow-changing society, formal institutions adapting to those changes or remaining rigid, etc. Our results extend easily to a simple dynamic OLG environment, but for conciseness we refer the interested reader to our working paper.

II. Equilibrium Behavior: Social Norms and Social Learning

We analyze here the social equilibrium between agents that obtains for any given level of publicity. This is an interesting question in its own right, especially with each individual facing both first-order uncertainty about the population means θ and μ and higher-order uncertainty about the beliefs of others, which in turn determine how his actions are likely to be judged.

Maximizing his utility (3), each agent chooses his contribution level a_i to satisfy

$$C'(a) = v_i + E[\theta|\theta_i] + x\mu_i \frac{\partial R(a, \theta_i, \mu_i)}{\partial a}. \tag{7}$$

This equation embodies the agent's three basic motivations: his baseline intrinsic utility from contributing, his posterior belief about the quality of the public good, and the impact of contributions on his expected image. To form his optimal estimate of θ, he combines his private signal and prior expectation according to

$$E[\theta|\theta_i] = \rho\theta_i + (1-\rho)\bar{\theta}, \tag{8}$$

where $\rho = \sigma_\theta^2 / (\sigma_\theta^2 + s_\theta^2)$ is the *signal-to-noise ratio* in his inference. We next show that when agents use linear strategies, $\partial R(a, \theta_i, \mu)/\partial a$ is constant, leading to a unique outcome.

[26] There is no "right answer" on what these preferences should be: the limited evidence on this question (Harbaugh, Mayr, and Burghart 2007) suggests that while induced contributions from some outside source do generate some intrinsic satisfaction, it is markedly less than that associated to own contributions.

[27] We can thus interpret ω as the wedge left after the Principal has already used any standard incentives at her disposal. This is precisely formalized in equation (A35) of the Appendix.

PROPOSITION 1 (Equilibrium Behavior and Benchmarking): *Fix* $x \geq 0$. *There is a unique linear equilibrium, in which an agent of type* (v_i, θ_i, μ_i) *chooses*

$$a_i(v_i, \theta_i, \mu_i) = v_i + \rho\theta_i + (1-\rho)\bar{\theta} + \mu_i x\xi(x), \tag{9}$$

where $\rho \equiv \sigma_\theta^2/(\sigma_\theta^2 + s_\theta^2)$ *and* $\xi(x)$ *is the unique solution to*

$$\xi(x) = \frac{s_v^2}{x^2\xi(x)^2 s_\mu^2 + s_v^2 + \rho^2 s_\theta^2}. \tag{10}$$

The resulting aggregate contribution (*or compliance level*) *is*

$$\bar{a}(x, \theta, \mu) = \bar{v} + \rho\theta + (1-\rho)\bar{\theta} + \mu x\xi(x). \tag{11}$$

A Sufficient Statistic Result.—Greater intrinsic motivation v_i, better perceived quality θ_i, and a stronger image concern μ_i naturally lead an agent to contribute more. Most remarkable, however, is the *simplicity* of the social image computations that emerge from this complex setting, as reflected by the common marginal impact $\xi(x)$ that an additional unit of contribution has on one's expected image. This is also, intuitively, the *signal-to-noise ratio* faced by an *observer* when trying to infer someone's type v_i from their action, knowing that behavior reflects private preferences, private signals, and image concerns according to (9).

Strikingly, the expected image return is the *same for all agents*, even though they have *different signals* (θ_i, μ_i) that are predictive of the average θ and μ, hence also of the θ_js and μ_js that observers will have at their disposal to extract v_i from a_i, using (9).[28]

The reason for this result is a form of *benchmarking*: an observer j does not need to separately estimate and filter out the contributions of θ_i and μ_i to a_i but only that of the linear combination $\rho\theta_i + \mu_i x\xi(x)$, and for this purpose $\rho\theta + \mu x\xi(x)$, hence also $\bar{a}$, is a *sufficient statistic*. Thus, whereas $E[\theta_i | a, \bar{a}, \theta_j, \mu_j]$ and $E[\mu_i | a, \bar{a}, \theta_j, \mu_j]$ both depend on j's private type, $E[a_i - \rho\theta_i - \mu_i x\xi(x) | a_i, \bar{a}, \theta_j, \mu_j]$ does not, so all observers of agent i will share the *same beliefs* about his motivation: $E[v_i | a, \bar{a}, \theta_j, \mu_j] = E[v_i | a, \bar{a}]$.[29] Agent i's ex post reputation will thus only depend on $a_i - \bar{a}$, implying in turn that his own (θ_i, μ_i), while critical to forecast $\bar{a}$ itself ex ante, will not affect the marginal return: $\partial R(a, \theta_i, \mu_i)/\partial a = \xi(x)$.

Put differently, when a_i is *judged against the benchmark* $\bar{a}$, contributions above average (say) must reflect a higher than average preference, signal, or image concern:

$$a_i - \bar{a} = v_i - \bar{v} + \rho(\theta_i - \theta) + x\xi(x)(\mu_i - \mu). \tag{12}$$

[28] In standard models of signaling or career concerns (e.g., Holmström 1999), agents have no uncertainty about the beliefs of those who will be judging them, resulting mechanically in a common (and deterministic) image return. Here no one knows how the audience will interpret their actions, and everyone has different signals about the state variable critical to this question. This is what makes the result notable, and indeed the fact that higher-order beliefs "drop out" arises only as an *equilibrium* property.

[29] To see that this is far from obvious a priori, note that it would no longer be the case if $\bar{a}$ itself was observed with noise or subject to small-sample variations from a finite number of agents. These represent potentially interesting extensions of the current model.

Observers assign to each source of variation a weight proportional to its relative variance, *conditional on* $\bar{a}$, so that

$$E\left[v_i|a_i,\bar{a}\right] = \left(1-\xi\right)\bar{v} + \xi\left(\bar{v}+a_i-\bar{a}\right) = \bar{v} + \xi\left(a_i-\bar{a}\right), \tag{13}$$

where $\xi(x)$ is given as a fixed point by (10). When there is no idiosyncratic variance in the value of image, $s_\mu^2 = 0$, the problem simplifies further, leading to a value

$$\xi = \frac{s_v^2}{s_v^2 + \rho^2 s_\theta^2}. \tag{14}$$

When image concerns differ, $s_\mu^2 > 0$, the informativeness of individual behavior is further reduced by the possibility that it might have been motived by image seeking (a high μ_i); thus $\xi(x) < \xi(0) \equiv \xi$, with a slight abuse of notation. This "overjustification effect" is amplified as actions become more visible, resulting in a *partial crowding out*: $\beta(x) \equiv x\xi(x)$, and thus also $\bar{a}(x)$, increase *less than one for one* with x.

PROPOSITION 2 (Comparative Statics of Social Interactions): *In equilibrium:*

(*i*) *The social image return* $\xi(x)$ *is strictly increasing in the dispersion of agents' preferences* s_v^2, *decreasing in their aggregate variability* σ_θ^2, *in the level of publicity x and in the dispersion of agents' image concerns* s_μ^2, *and U-shaped in the quality of their signals*, s_θ^2.

(*ii*) *The impact of visibility on contributions*, $\beta(x) \equiv x\xi(x)$, *is strictly increasing in x, with* $\lim_{x\to\infty}\beta(x) = \infty$, *and it shares the properties of* $\xi(x)$ *with respect to all variance parameters. The same is true of the aggregate contribution* $\bar{a}(x)$, *as long as* $\mu > 0$.[30]

The first two properties are quite intuitive. First, signaling motives are amplified by a greater cross-sectional dispersion s_v^2 in the preferences v_i that observers are trying to infer. Second, decreasing the variance σ_θ^2 of the aggregate shock means that each agent is less responsive to his private information θ_i (as it is more likely to be noise), so individual variations in contribution are again more indicative of differences in intrinsic motivation. The attribution-garbling role of differences in image concerns, s_μ^2, was explained earlier.

The last comparative static is more novel and subtle: the U-shape of ξ and β in s_θ^2 reflects the idea that reputational effects are strongest when agents have the same *interim* belief about the quality of the public good. This occurs when their private signals are either very precise ($s_\theta \to 0$), and hence all close to the true θ, or on the contrary very imprecise ($s_\theta \to \infty$), leading them to put a weight close to 1 on the common prior $\bar{\theta}$. In both cases, differences in contributions reflect differences

[30] This restriction means that agents want to be perceived as prosocial rather than antisocial; since $\bar{\mu}$ is taken to be large, this case occurs with probability close to 1.

in intrinsic motivation much more than in information about θ, which intensifies the signaling game and thereby raises contributions. This nonmonotonicity emerges from the common-value environment, in which each agent estimates the quality of the "true" public good based on his signal. By contrast, in the private-values environment studied in Section IIIE, agents effectively behave as if $\rho = 1$; in that case, the social image return $\xi(x)$ is strictly decreasing in s_θ^2, further simplifying the result.

As $\xi(x) \to 1$, the equilibrium becomes fully revealing, with each agent's social image exactly matching his actual preference: $E[v_i|a_i] = v_i$. Yet his contribution exceeds by $x\mu_i$ that which he would make were his type directly observable: the contest for status traps everyone in an expectations game where they cannot afford to contribute less than the equilibrium level.

Two more general lessons also emerge from Propositions 1–2:

Visibility and Conformity.—The model shows that the link between these two notions is more subtle than usually thought: a higher x shifts all contributions in the same direction (typically positive, given $\bar{\mu} > 0$), but simultaneously increases their *dispersion* between agents whenever $s_\mu^2 > 0$.

Exogenous Variations in Privacy.—Inspection of (11) already makes apparent some key trade-offs: a higher x increases aggregate contributions $\bar{a}(x)$ but also causes them to vary inefficiently, and most importantly it reduces their reliability as an indicator of what actually constitutes the social good (θ). This last point applies to any observer, and in a dynamic setting we can think of each generation trying to learn what is "the right thing to do" from the behavior of its elders. Propositions 1–2 imply that in low-privacy environments such as small villages, early societies, and other close-knit groups, social norms and formal institutions will be *slow to adapt* and often *remain inefficient* for a long time. Principals who can influence the general level of privacy will naturally take the above trade-off into account, a case we now turn to formally.

III. Optimal Publicity and Matching Policies

We model the degree of public visibility and memorability of agents' actions as a parameter $x \in \mathbb{R}_+$ that scales reputational payoffs up or down to $x\mu_i R(a, \theta_i, \mu_i)$. To focus on how a *social value of privacy* arises endogenously, we assume that the Principal can vary x costlessly. While the costs of honorific ceremonies, medals, public shame lists, etc., are nonzero, they are trivially small compared to direct spending on public goods, subsidies, or law enforcement. This cost advantage is one of the main arguments put forward by proponents of publicity and shame.

We uncover three distinct motivations for the Principal to grant agents some degree of privacy, and in order to isolate each one, we consider in turn

(i) a simple benchmark without any variability in the average image motive, $\sigma_\mu^2 = 0$;

(ii) a case where $\sigma_\mu^2 > 0$, but the Principal observes the realization of μ once x has been set but prior to choosing a_P; and

(iii) the main setting of interest, in which the Principal is uncertain about the realizations of both aggregate shocks, θ and μ.

These three nested cases provide insights into, respectively, (i) how the Principal would set publicity if she could fine-tune its impact, $x\xi\mu$, perfectly; (ii) the "variance effect" that emerges when she cannot do so but observes μ ex post; and (iii) the "information-distortion effect" that arises when publicizing behavior generates a signal-extraction problem.

To further simplify the exposition, we shall initially focus on the case in which *all agents share the same value for social image*: $\mu_i = \mu$ for every i, or equivalently $s_\mu^2 = 0$. This assumption (almost universal in the literature on signaling) will most clearly highlight the role of *aggregate* variability in reputational concerns, which is key to the Principal's learning problem.[31] Agents' social-learning problems, meanwhile, become simpler, with the image return $\xi(x)$ reducing to the constant ξ given by (14). In Section IIID we allow for $s_\mu^2 > 0$ and show that all key results remain unchanged.

A. *Fine-Tuned Publicity: An Image-Based Pigovian Policy*

Consider first the simple case where agents' image motive is invariant: both they and the Principal believe with probability 1 that $\mu = \bar{\mu}$, so $\sigma_\mu^2 = 0$. Upon observing the aggregate contribution $\bar{a}$, the Principal perfectly infers θ by inverting (11), allowing her to optimally set

$$a_p = \frac{(w+\theta)\big[\lambda + (1-\lambda)b\big]}{k_P(1-\lambda)} = \frac{(w+\theta)\varphi}{k_P(1-\lambda)}, \tag{15}$$

where φ was defined in (5). This full revelation of θ also makes the Principal's own signal θ_P, received at the interim stage, redundant. Anticipating this at the ex ante stage, the expectations of θ, μ, and $\bar{a}$ that she uses in choosing x are thus simply her priors $\bar{\theta}$, $\bar{\mu}$, and $\tilde{a}(x) = \bar{\nu} + \bar{\theta} + x\xi\bar{\mu}$. Substituting into the objective function (4) and differentiating leads to an optimal level

$$x^{FB} = \frac{\big(w+\bar{\theta}\big)\varphi - \big(\bar{\nu}+\theta\big)\lambda(1-\alpha)}{\lambda\xi\bar{\mu}} = \frac{\omega}{\lambda\xi\bar{\mu}} > 0, \tag{16}$$

where the superscript stands for "First Best" and the wedge $\omega > 0$ was defined in (6).

[31] A further simplification is that agents' reputational gains and losses sum to 0 (by Bayes' rule, $\int_0^1 \mu R\big(a_i,\theta_i,\mu_i\big)\,di = \mu\bar{\nu}$), so the corresponding term (and concern) vanishes from the Principal's objective function (4), as if $\tilde{\alpha} = 0$. Of course, if agents really have a common μ, it should not be too hard for the Principal to find out its value. The initial focus on this case is thus only a simplifying expository device on the way to the more realistic, full-fledged model, where the μ_is are also private information.

Image-Based Pigovian Policy.—Consider in particular a Principal who values the public good exactly like the agents but puts no weight on their "warm glow" utilities from contributing: $\tilde{\alpha} = \alpha = 0$, $b = 0$, and $\lambda = 1/2$. The optimal level of visibility is then

$$x^{FB} = \frac{w - \bar{v}}{\xi \bar{\mu}}. \tag{17}$$

This corresponds to a "Pigovian" image subsidy that the Principal fine-tunes to exactly offset free riding, namely the gap between equilibrium private compliance $\bar{a}$ and its socially optimal level, $w + \theta$. More generally, by using *publicity as an incentive* according to (16), the Principal is able to achieve her preferred overall level of public good provision, fully offsetting the wedge ω, just as she would with monetary subsidies.[32]

B. *The Variance Effect*

When there are variations in the average importance of social image, $\sigma_\mu^2 > 0$, the Principal can no longer finely adjust publicity ex ante to precisely control agents' compliance and achieve her first-best through (15)–(16). If she learns the realization of μ ex post (once x has been set), she is again able, upon observing $\bar{a}$, to infer the true θ by inverting (11). As before, she will thus ignore her signal θ_P and set a_P without error according to (15). For any choice of publicity x, however, the aggregate contribution $\bar{a}(x) = \bar{v} + \theta + x\xi\mu$ will now reflect not only the realized quality of the public good θ but also unrelated variations in μ. Using the distribution of $\bar{a}(x)$, we can derive the Principal's expected payoff from x, denoted $E\tilde{V}(x)$. Relegating that derivation to the Appendix (A9), we focus here on the corresponding optimality condition

$$\frac{dE\tilde{V}(x)}{dx} = \underbrace{(\xi\bar{\mu})\omega}_{\text{Incentive Effect}} - \underbrace{\lambda x \xi^2 \left(\bar{\mu}^2 + \sigma_\mu^2\right)}_{\text{Variance Effect}}. \tag{18}$$

The two opposing terms clearly show the trade-off between leveraging social pressure to promote compliance and the inefficient, image-driven variations in aggregate contributions that arise from greater publicity. To the extent (λ) that the Principal internalizes the costs thus borne by the agents, she also loses from this *Variance Effect* and thus wants to moderate it.

PROPOSITION 3 (Incentive and Variance Effects): *When the Principal faces no ex post uncertainty about* μ (*symmetric information*), *she sets publicity level*

$$x^{SI} = \frac{\bar{\mu}\omega}{\lambda\xi\left(\bar{\mu}^2 + \sigma_\mu^2\right)} = \frac{x^{FB}}{1 + \sigma_\mu^2/\bar{\mu}^2}, \tag{19}$$

[32] Optimal compliance and thus publicity are bounded, because at higher levels the *private* costs of contributing incurred by agents exceed the corresponding social value. This is similar to analyses of optimal penalties in law and economics (Polinsky and Shavell 1979, Kaplow 1992), where nonmaximal deterrence is generally optimal.

where x^{FB} was defined in (16). *This optimal x^{SI} is increasing in w, $\bar{\theta}$, α, b, and σ_θ^2; decreasing in $\bar{v}$, s_v^2, and σ_μ^2; and U-shaped in s_θ^2 and in $1/\bar{\mu}$.*

The variance effect makes publicity a blunt instrument of social control, as emphasized by Whitman (1998) and E. Posner (2002), so the Principal naturally wields it more cautiously than under the Pigovian policy: $x^{SI} < x^{FB}$ for all $\lambda > 0$.

C. *Publicity and Information Distortion*

We now turn to the main setting of interest, in which the Principal does not observe the realization of μ and therefore faces an attribution problem: the overall contribution or compliance rate $\bar{a}$ reflects both public-good quality θ and social-enforcement concerns, μ. Using her *expected* value of μ to invert (11), she now obtains a noisy (but still unbiased) signal of θ:

$$\hat{\theta} \equiv \frac{1}{\rho}\left[\bar{a} - \bar{v} - x\xi\bar{\mu} - (1-\rho)\bar{\theta}\right] = \theta + \left(\frac{x\xi}{\rho}\right)(\mu - \bar{\mu}) \sim \mathcal{N}\left(\theta, \frac{x^2\xi^2\sigma_\mu^2}{\rho^2}\right). \tag{20}$$

Greater publicity makes the aggregate contribution less informative (in the Blackwell sense), as it magnifies its sensitivity to variations in image concerns, μ. This *Information-Distortion Effect* will cause the Principal to makes mistakes in setting her contribution a_P—or any other second-stage decision such as tax incentives, laws, etc. Moderating this informational loss is the fact that she receives a private signal $\theta_P \sim N\left(\theta, s_P^2\right)$, as described in Section I. This allows her to update her prior beliefs to an *interim* estimate $\bar{\theta}_P$ with mean square error σ_P^2, where

$$\bar{\theta}_P \equiv \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_P^2}\right)\theta_P + \left(\frac{s_P^2}{\sigma_\theta^2 + s_P^2}\right)\bar{\theta}, \tag{21}$$

$$\sigma_P^2 \equiv \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_P^2}\right)^2 s_P^2 + \left(\frac{s_P^2}{\sigma_\theta^2 + s_P^2}\right)^2 \sigma_\theta^2. \tag{22}$$

Combining this information with the signal $\hat{\theta}$ inferred from $\bar{a}$, the Principal's posterior expectation of θ is

$$E\left[\theta|\bar{a}, \theta_P\right] = \left[1 - \gamma(x)\right]\bar{\theta}_P + \gamma(x)\hat{\theta}, \tag{23}$$

where the weight

$$\gamma(x) \equiv \frac{\rho^2\sigma_P^2}{\rho^2\sigma_P^2 + x^2\xi^2\sigma_\mu^2}, \tag{24}$$

measures the relative precision of $\hat{\theta}$, or equivalently the *informational content* of compliance $\bar{a}$. The signal-garbling effect of publicity for the Principal is clearly apparent from the fact that γ decreases with x, which matters to the Principal

when her own signal, θ_P, is noisy.[33] After observing $\bar{a}$, the Principal optimally sets $a_P = \varphi\big(w + E\big[\theta|\bar{a},\theta_P\big]\big)/\big(1-\lambda\big)k_P$; substituting in (20) and (23) yields the following result.

PROPOSITION 4 (Optimal Matching): *The Principal's contribution policy is equivalent to setting a baseline investment* $\underline{a}_P\big(x,\theta_P\big)$ (*given in the Appendix*) *and a matching rate*

$$m\big(x\big) \equiv \frac{\gamma\big(x\big)\varphi}{\rho k_P\big(1-\lambda\big)} \tag{25}$$

on private contributions $\bar{a}$. *The less informative is* $\bar{a}$ (*in particular, the higher is publicity* x); *the lower is the matching rate.*

Conditioning on the true realizations of θ and μ, (11), (20), and (23) imply that the Principal's forecast error is equal to

$$\Delta \equiv E\big[\theta|\bar{a},\theta_P\big] - \theta = \big[1-\gamma(x)\big]\big(\bar{\theta}_P - \theta\big) + \frac{\gamma\big(x\big)x\xi}{\rho}\big(\mu - \bar{\mu}\big). \tag{26}$$

Her ex ante expected payoff is reduced, relative to the symmetric-information benchmark, by a term proportional to the variance of these forecasting mistakes, which simple derivations in the Appendix show to be proportional to her loss of information:

$$EV\big(x\big) = E\tilde{V}\big(x\big) - \frac{\varphi^2\sigma_P^2}{2\big(1-\lambda\big)k_P}\big[1-\gamma\big(x\big)\big]. \tag{27}$$

The Principal's first-order condition is now

$$\frac{dEV\big(x\big)}{dx} = \underbrace{\frac{d\tilde{EV}\big(x\big)}{dx}}_{\text{Incentive and Variance Effects}} - \underbrace{\frac{\varphi^2\sigma_\mu^2\xi^2}{\rho^2\big(1-\lambda\big)k_P}\gamma\big(x\big)^2 x}_{\text{Information-Distortion Effect}}. \tag{28}$$

The first term, previously exposited in (18), embodies the beneficial incentive effect of visibility and its variability cost. The new term is the (marginal) loss from distorting information, which naturally leads to a lower choice of publicity than the optimal Pigovian policy, even below the symmetric-information benchmark of Section IIIB.

[33] In a more general context (departing from linear strategies), if agents' behavior involves discrete bunching, increases in x could sometimes make $\bar{a}$ more informative by "breaking down" atoms of pooling. In this (somewhat less interesting) case, the Principal's cost of using publicity naturally declines.

PROPOSITION 5 (Optimal Privacy): *When the Principal is uncertain about the importance of social image, the optimal degree of publicity* $x^* \in \left(0, x^{SI}\right)$ *solves the implicit equation*

$$x = \frac{\bar{\mu}\omega}{\xi\left(\lambda\left(\bar{\mu}^2 + \sigma_\mu^2\right) + \frac{1}{(1-\lambda)k_P}\left(\frac{\varphi\sigma_\mu\gamma(x)}{\rho}\right)^2\right)}. \tag{29}$$

In general, (29) could have multiple solutions, because the cost of information distortion is not globally convex: the marginal loss, proportional to $\gamma(x)^2 x$, is hump shaped in x.[34] While there may thus be multiple local optima, *all are below* x^{SI} (the optimum absent information-distortion issues), and therefore so is the *global optimum* x^*. All also share the same comparative-statics properties, which we shall analyze in Section IV for the more general model where agents may differ in how they value reputation.

D. *Allowing for Heterogeneous Image Concerns*

When people differ in how image driven they are, $s_\mu^2 > 0$, agents' inference and decision problems become more complex (though still fully tractable, as shown in Proposition 1) due to the overjustification effect. This heterogeneity, on the other hand, has *no impact on the Principal's learning problem*: as seen from (11), idiosyncratic differences in μ_is wash out in the aggregate contribution $\bar{a}(x)$, implying the following result.

COROLLARY 1: *At any given level of* x, *the informational content* $\gamma(x)$ *of aggregate compliance* $\bar{a}(x)$, *the Principal's optimal matching rate* $m(x)$ *and her informational loss* $EV(x) - E\tilde{V}(x)$ *from not observing the aggregate realization* μ *remain the same as in* (24), (25), *and* (27) *respectively, except that* $x\xi$ *is replaced everywhere by* $x\xi(x) = \beta(x)$, *defined from* (10).

Relegating derivations to the Appendix, the marginal effect of publicity on the Principal's payoff now takes the form

$$\frac{dEV(x)}{dx} = \beta'(x)\left[\omega\bar{\mu} - \lambda\beta(x)\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2 - \frac{\varphi^2\sigma_\mu^2\gamma(x)^2}{\rho^2(1-\lambda)k_P}\right)\right], \tag{30}$$

showing how s_μ^2 worsens the variance effect by creating inefficient individual differences in behavior, but also benefits a Principal who values agents' pure image utility, with weight $\tilde{\alpha}$ (a standard "convex surplus" effect). To rule out the

[34] By (24), it equals $x/\left(1 + Ax^2\right)^2$, where $A \equiv \xi^2\sigma_\mu^2/\rho^2\sigma_{\theta,P}^2$. Simple derivations show this function to be increasing up to $x = 1/\sqrt{3A}$, then decreasing.

uninteresting and implausible case where she cares so much about agents' image satisfaction that this dominates all other concerns and makes the optimal x infinite, we shall assume in what follows that

$$\bar{\mu}^2 + \sigma_\mu^2 + \left(1 - 2\tilde{\alpha}\right) s_\mu^2 \; > \; 0, \tag{31}$$

which is ensured in particular if either (i) $\tilde{\alpha} \leq 1/2$ or (ii) $s_\mu^2 < \bar{\mu}^2 + \sigma_\mu^2 = E\left[\mu\right]^2$, meaning that idiosyncratic variations are not too large compared to the prior mean and/or aggregate variations.

Most importantly, we see from (30) that keeping fixed agents' inferences about each other, i.e., β, the Principal's informational loss concerning θ remains unchanged. Setting dEV/dx to 0 then yields the following results.

PROPOSITION 6 (Optimal Privacy: General Case): *When the Principal is uncertain about the importance of social image, the optimal degree of publicity x^* solves the implicit equation*

$$x^* = \frac{\bar{\mu}}{\xi\left(x^*\right)}\left(\frac{\omega}{\lambda\left(\bar{\mu}^2 + \sigma_\mu^2 + \left(1 - 2\tilde{\alpha}\right) s_\mu^2\right) + \frac{1}{\left(1-\lambda\right) k_P}\left(\frac{\varphi \sigma_\mu \gamma\left(x^*\right)}{\rho}\right)^2}\right), \tag{32}$$

where $\xi\left(x\right)$ is given by (10) *and $\gamma\left(x\right)$ remains given by* (24). *The solution is thus identical to that in Proposition* 5, *except that σ_μ^2 is replaced by $\sigma_\mu^2 + s_\mu^2\left(1 - 2\tilde{\alpha}\right)$ and ξ by $\xi\left(x\right)$ everywhere.*

As long as the Principal places a weight $\tilde{\alpha} \leq 1/2$ on agents' social image, she will set x^* below the level of Proposition 5, which itself is below the x^{SI} of Proposition 3. If $\tilde{\alpha}$ is very high, however, the optimal x^* could be above either or both of these levels. As before, (32) could have multiple solutions, but all stable ones, including the global optimum, share the same comparative statics properties, analyzed in Section IV below.

E. *Disagreeing over the Norm: Private Values*

We have so far focused on the case of *common values*, in which each agent cares about some objective quality θ of the public good (e.g., tax or environmental compliance) and uses his own signal θ_i to assess it. Accordingly, he also "respects" the information held about θ by others. For other social norms applications (minority rights, religion, abortion, etc.), a more appropriate setting may be one of *private values*, in which people ultimately "agree to disagree" about what is socially valuable or moral, or they just benefit differently from the public good or externality. Agent's direct payoffs thus change from (1) to

$$U_i^{PV}\left(v_i, \theta_i, w; a_i, \bar{a}, a_P\right) \equiv \left(v_i + \theta_i\right) a_i + \left(w + \theta_i\right)\left(\bar{a} + a_P\right) - C\left(a_i\right), \tag{33}$$

in which the θ_is now represent intrinsically different preferences, with s_θ^2 measuring the extent of disagreement in the population.[35] Reputational payoffs are still given by (2): each agent knows his own tastes exactly but remains uncertain of how his actions will be interpreted by others, since he does not know their private values θ_j.[36] The Principal's problem, finally, is also largely unchanged: she now cares about the average preference θ, through agents' total welfare, and more generally needs to learn this aggregate shift in the distribution to set her preferred policy; see (A28) for details.

This private-values case turns out to map exactly into a special case of the common values benchmark. The key insight (see Section A.A2 for details) is that because each agent is now *sure* of how he values the public good (θ_i), he contributes *as if* he were getting a perfect signal about a common value, that is, as if $\rho = 1$.[37] The entire analysis is thus unchanged from Section I to Section V, but for *simply replacing* ρ *by* 1 in all formulas.

IV. Comparative Statics of the Optimal Policy Bundle

Let us now examine how the Principal's choice of *publicity* x^* and *matching rate* $m^* = \gamma(x^*)/\left[\rho k_P(1-\lambda)\right]$ depend on key features of the environment. We cover common and private values together: due to the "$\rho \equiv 1$" correspondence identified above, the results are either identical across the two cases or even stronger in the latter.

Basic Results.—From (30) it is clear that $\partial^2 EV/\partial x \partial \omega > 0$ and $\partial^2 EV/\partial x \partial k_P > 0$, leading to the results summarized in Table 1 below.[38] These properties are quite intuitive. For instance, a Principal who faces a higher cost of own funds, or internalizes agents' warm glow utility, wants to encourage private contributions. She therefore makes behavior more observable and, as it becomes less informative, also reduces her matching rate.

We next turn to the dependence of the optimal policies on *second-moment* parameters of cross-sectional heterogeneity and aggregate variability, summarized below in Table 2.

[35] For simplicity, we let the same θ_i parametrize agent i's intrinsic preference for engaging in the activity and his utility from its overall level. Such is the case when the motivation for doing a_i is purely prosocial, but in general the two values could differ: own sexual preferences versus attitudes toward those of others, enjoyment of polluting activities versus vulnerability to airborne particulates or climate change, etc. Typically they will be positively correlated due to (i) the common prosocial concern component and (ii) psychological mechanisms such as projection bias and people's reluctance to recognize that they may be harming others. The model and all results could be extended to any nonzero correlation, at the cost of additional notation.

[36] In this setting, the θ_is are tastes (or costs) specific to the particular good or behavior in question. By contrast, v_i is agent i's general degree of prosocial orientation or other-regard, which carries across contexts and periods. Consequently, reputations are still formed over v_i, rather than $v_i + \theta_i$.

[37] In making inferences about each other's v_i, agents still use the true signal-to-noise ratio $\sigma_\theta^2/\left(\sigma_\theta^2 + s_\theta^2\right) < 1$, but because of the sufficient-statistic result discussed earlier, this ends up not affecting the equilibrium outcome.

[38] For any parameter η, supermodularity of $V(x,\eta)$ implies that $\operatorname{argmax}_x V(x,\eta)$ increases with η in the strong set-order sense, and in the usual sense when the maximizer is unique.

Table 1—Comparative Static Effects of First-Moment Parameters

		Optimal publicity x^*	Optimal matching rate m^*
Baseline externality	w	Increasing	Decreasing
Ex ante expected quality	$\bar{\theta}$	Increasing	Decreasing
Weight on agents' warm glow	α	Increasing	Decreasing
Average intrinsic motivation	$\bar{v}$	Decreasing	Increasing
Principal's relative cost	k_P	Increasing	Decreasing

Table 2—Comparative Static Effects of Second-Moment Parameters

	Optimal publicity x^*	Optimal matching rate m^*
s_v^2	Decreasing	Invariant
σ_θ^2	Decreasing for s_θ/σ_θ small or private values	Increasing for s_θ/σ_θ small or private values
σ_μ^2	Decreasing outside $[\underline{\sigma}, \bar{\sigma}]$ or if $k_P \geq \bar{k}_P$	Increasing outside $[\underline{\sigma}, \bar{\sigma}]$, or if $k_P \geq \bar{k}_P$
s_P^2	Decreasing	Increasing
s_θ^2	Increasing for s_θ/σ_θ small or private values	Invariant

Heterogeneity in Intrinsic Motivation.—An increase in s_v^2 directly raises the variability of individual contributions, and this has both costs and benefits for the Principal. To the extent that she weighs agents' warm glow positively, she appreciates variability, but on the other hand, she suffers from internalizing its effect on their total contribution cost.[39]

In addition to these direct effects, a rise in s_v^2 also increases the marginal impact of contributions on image $\xi(x)$ and thus the incentive to contribute, $\beta(x) = x\xi(x)$. For fixed x, this affects all three components of the Principal's trade-off: it raises average contributions but further increases their sensitivity to μ, and consequently also worsens the information loss (γ declines). When publicity is optimally chosen, however, these three effects balance out exactly: because $\xi(x)$ and x enter *EV* only through the product $x\xi(x)$, we can think of the Principal as *directly optimizing over* the value of β. Changing s_v^2 therefore only has a direct effect on her payoff. For the same reason, the Principal responds at the margin only to the direct (variance) effect of an increase in s_v^2: she reduces x to partially offset it, so as to keep $\beta(x)$ constant. Since s_v^2 influences γ and m only through the value of $\beta(x)$, both remain unchanged.

PROPOSITION 7: *The optimal publicity x^* choice is decreasing in s_v^2, the variance of intrinsic motivation in the population, while the optimal matching rate m^* is independent of it. The Principal's expected payoff (at the optimal x^*) changes with s_v^2 proportionately to $\lambda(\alpha - 1/2)$.*

Variability in Societal Preferences.—Comparative statics with respect to σ_θ^2 are less straightforward, as it matters through two very different channels: it represents the Principal's *ex ante uncertainty* about θ, but also the extent to which agents

[39] Since in equilibrium each a_i is increasing in v_i, a mean-preserving spread in v_i increases the benefit term $\alpha \int_0^1 v_i a_i d_i$ in (4), but it also magnifies the cost term $(-1/2)\int_0^1 a_i^2 d_i$.

disregard their signal and *follow the common prior*. To neutralize the second effect and highlight the Principal's trade-off between raising $\bar{a}$ and learning about θ, let us focus here on the limiting case in which agents' private signals are far more informative than their prior, so that $s_\theta/\sigma_\theta \approx 0$ or, equivalently, $\rho \approx 1$. In this case, $\xi(x)$ becomes independent of σ_θ^2, which then enters (30) only by raising $\gamma(x)$, through its effect on σ_P^2; see (10), (21), and (24). Therefore, the following result holds.

PROPOSITION 8: *The optimal visibility x^* decreases with ex ante uncertainty σ_θ^2 over θ, while the optimal matching rate γ^* increases with it (i) in the private-values case and (ii) in the common-values case, when agents' private signals about the quality of the public good are sufficiently more precise than their prior over it $(s_\theta^2/\sigma_\theta^2$ small enough$)$.*

The assumption of a small $s_\theta^2/\sigma_\theta^2$ is somewhat restrictive, but it captures the most relevant settings for the question we ask, namely those in which the Principal has a lot to learn from agents relative to the prior. Moreover, the same comparative statics emerge unambiguously under private values, regardless of the value of $s_\theta^2/\sigma_\theta^2$.

Variability in the Importance of Social Image or Social Enforcement.—

(**i**) **Average Social Image Concern.**—An increase in σ_μ^2 does not affect ρ or $\xi(x)$, so it leaves the incentive effect of visibility unchanged. For fixed publicity x, it naturally makes $\bar{a}$ less informative about θ, so $\gamma(x)$ declines. It also leads to a higher variance effect, so for both reasons the Principal is worse off. The effects of σ_μ^2 on the optimal publicity and matching rate, on the other hand, are generally ambiguous: by (28), the marginal information cost is proportional to $\sigma_\mu^2 \beta(x)\gamma^2(x)$, which can be seen from (24) to be hump shaped in σ_μ^2, given x.

Somewhat surprisingly, the Principal may thus use *more publicity* when the source of noise in her learning problem increases. Such a "paradoxical" possibility (confirmed by simulations) only arises for intermediate values of σ_μ^2 (where the marginal information cost is near its minimum), however. When σ_μ^2 is sufficiently low or high, on the contrary, the information effect goes in the same direction as the variance effect, leading the Principal to *reduce publicity*—the more unpredictable is agents' sensitivity to it the more so, as one would expect.

Another (more straightforward) case in which the result is unambiguous is when k_P is large enough: since the Principal will not contribute much anyway, information is not very valuable to her, so as σ_μ^2 rises, her main concern is the variance effect. In what follows we shall denote $\bar{k}_P \equiv \varphi^2/\left[27\lambda(1-\lambda)\rho^2\right]$.

PROPOSITION 9: *Variability in the importance of social image, σ_μ^2, has the following effects:*

(i) *The Principal's payoff is decreasing in σ_μ^2.*

(*ii*) *If* $k_P \geq \bar{k}_P$, *the optimal level of publicity* x^* *also decreases with* σ_μ^2. *Otherwise, there exist* $\underline{\sigma}$ *and* $\bar{\sigma}$ *such that* x^* *is decreasing in* σ_μ^2 *if either* $\sigma_\mu < \underline{\sigma}$ *or* $\sigma_\mu > \bar{\sigma}$.

(*iii*) *As* $\sigma_\mu \to \infty$, $x^* \to 0$ (*full privacy*), *while as* $\sigma_\mu \to 0$, x^* *approaches the symmetric-information level* x^{SI} *that solves* $x\gamma(x) = \bar{\mu}\omega / \left[\lambda\left(\bar{\mu}^2 + (1 - 2\tilde{\alpha})s_\mu^2\right)\right]$.

(**ii**) **Heterogeneity in Image Concerns.**—For given x, an increase in s_μ^2 influences the image incentive $\beta(x)$ in complex ways (see (10)), so the resulting comparative statics of optimal privacy are generally ambiguous. For the Principal's payoff, on the other hand, the impact of s_μ^2 depends very simply on whether or not she internalizes agents' image utility gains enough to compensate for the economic costs arising through the greater variance effect.

PROPOSITION 10: *The Principal's expected payoff is strictly decreasing in* s_μ^2 *if* $\tilde{\alpha} < 1/2$ *and increasing otherwise.*

Precision of Private Signals.—

(**i**) **Principal's Signal.**—When the noise s_P^2 affecting her independent information increases, the Principal is naturally worse off. To see how she responds, note from (28) that s_P^2 appears only in the information-distortion effect, through $\gamma(x)$; thus, from (24), $\partial^2 EV(x)/\partial x \partial s_P^2 < 0$. This is intuitive: as the Principal becomes less well informed about agents' preferences, she reduces publicity so as to learn more from their behavior. Since γ increases with s_P and decreases with x, it follows that so does the optimal matching rate: a Principal with access to less independent information relies more on agents' behavior as a guide for her own actions.

PROPOSITION 11: *The Principal's payoff and optimal publicity choice* x^* *decrease with the variance of her signal,* s_P^2, *whereas her optimal matching rate* m^* *increases with it.*

(**ii**) **Agents' Signals.**—The quality of agents' private information has more ambiguous effects. At a given level of x, greater idiosyncratic noise s_θ^2 reduces everyone's responsiveness to their private signal, and thereby also the informativeness of aggregate contributions. At the same time, recall from Proposition 2 that the reputational return ξ is U-shaped in s_θ^2: the level, variance, and informativeness of agents' contributions are thus nonmonotonic in s_θ^2, and therefore so are the Principal's optimal level of publicity and matching rate.

We can again say more when the Principal has "enough" to learn from agents, meaning that their private signals are sufficiently informative: as $s_\theta/\sigma_\theta \to 0, \rho$ approaches 1 and the Principal's optimality condition (30) involves x and ξ

only through their product $\beta(x) = x\xi(x)$, while s_θ^2 enters it only through $\xi(x)$. It follows that the optimal value of β is independent of s_θ, while the associated x must rise with it so as to maintain that constancy.

PROPOSITION 12: *Optimal publicity x^* increases with the variance of agents' signal s_θ^2, while the matching rate m^* remains invariant and the Principal's payoff falls* (i) *in the private-values case and* (ii) *in the common-values case, when agents' private signals about the quality of the public good are sufficiently more precise than their prior over it* ($s_\theta^2/\sigma_\theta^2$ *small enough*).

V. Combining Reputational and Material Incentives

In most real-world settings, Principals have access to both monetary and image incentives to induce desirable behaviors. This makes the question of how these should optimally be combined a very natural one, but so far it has not been examined in the literature.

Material incentives come at a cost, such as the deadweight loss from taxation; publicity involves (almost) no direct cost but has indirect ones—in our case inefficient variations in compliance and/or reduced information about changing preferences. In this section we examine how the social equilibrium among agents changes when they face both incentives together, then solve for the Principal's optimal policy mix. The next section will analyze the sequential case, in which the Principal initially controls only image incentives (or "informal institutions") and then learns from the social outcome how material incentives (or "formal institutions") should be set up. In both cases we show that all the comparative statics of the original model (Tables 1–2) remain essentially unchanged, and we derive new ones. The details are relegated to the Appendices, whereas we highlight here the key insights and results.

In period 1, let the Principal now simultaneously choose the level of visibility x among agents and an incentive rate $y \geq 0$, paid to each of them per unit of contribution[40] at a resource cost of $(1+\kappa)y$, where $\kappa \geq 0$ represents a deadweight loss or other opportunity cost of funds. Everything else remains unchanged, with the second-period policy characterized by a baseline level $\underline{a}_P(x, \theta_P)$ and a matching rate $m(x)$.

Denote $a_i(x)$ the equilibrium strategies of individual agents in the baseline model, namely (9). It is is clear that in the presence of a common incentive rate, the (linear) equilibrium in the augmented model is simply given by $\tilde{a}_i(x,y) \equiv a_i(x) + y$, with the informational content of individual contributions $\xi(x)$ unchanged.[41] The same is therefore true of the aggregate $\bar{a}(x)$, so the Principal's signal-extraction problem is also unchanged, with the relative informational content $\gamma(x)$ of average compliance still given by (24). Thus, the

[40] In this extended analysis, it must be that the Principal observes not only the aggregate contribution $\bar{a}$ but also individual ones, or at least some noisy version of each a_i. Since both the Principal's and agents' payoffs are quasilinear in money, ya_i is then agent i's expected incentive payment when his contribution is a_i.

[41] This contrasts with Bénabou and Tirole (2006), in which each agent's marginal value for money may be a private type, so that introducing incentives generates an additional signal-extraction problem.

key trade-off between publicity and learning remains. The incentive and variance effects of publicity are somewhat different, however, and this will affect the optimal x^*. First, to the extent that monetary incentives can be used to close some of the wedge ω, publicity has less of a role to play. Second, by increasing agents' levels of contribution, incentives raise their marginal costs and thereby worsen the variance effect. The following results are derived, without much loss of generality, for the case where $\lambda = 1/2$, which corresponds to a social planner who cares about aggregate social welfare.

PROPOSITION 13 (Optimal Mix of Material and Image Incentives): *Let* $\lambda = 1/2$. *In both the public and private value cases, there exists* $\kappa \in (0, \infty)$ *such that:*

(*i*) *As* κ *increases from* 0 *to* $\bar{\kappa}$, y^* *decreases from* 2ω *to* 0, *while* x^* *increases from* 0 *to the benchmark-model solution given by* (32). *The two policy tools are linked by*

$$y^* = \tilde{y} - \tau\bar{\mu}\beta(x^*), \tag{34}$$

reflecting their substitutability, where $\tilde{y}$ *and* τ *are positive constants. For* $\kappa \geq \bar{\kappa}$, $y^* = 0$ *and* x^* *remains invariant.*

(*ii*) *For any* κ, *the comparative statics of* x^* *and* m^* *with respect to all parameters in Tables* 1–2 *remain unchanged but for one*: x^* *is now inverse U-shaped in* $\bar{v}$ (*increasing where* $y^* > 0$, *decreasing as before where* $y^* = 0$), *while* m^* *again has the opposite variations.*

(*iii*) *For all* $\kappa \leq \bar{\kappa}$, *the comparative statics of* y^* *with respect to* k_P, s_P^2, $\bar{v}$, σ_μ^2, σ_θ^2, *and* s_θ^2 *are the exact reverse of those of* x^*. *It is independent of* s_v^2, *while its comparative statics with respect to* w, $\bar{\theta}$, α, *and* $\bar{v}$ *are generally ambiguous.*

The first set of results is quite intuitive: when monetary incentives are costless, it is optimal to fully close the wedge $\omega/\lambda = 2\omega$ using this tool without resorting to publicity, which always entails distortions. As the opportunity cost of funds rises, the Principal increasingly substitutes publicity, up to the point where monetary incentives become too costly and using only image is optimal; we are then back to the benchmark model. Next, nearly all the comparative-statics properties of x^* (and m^*) remain unchanged when it is used alongside with monetary payments y^*;[42] we also get new predictions about the behavior of y^*.

[42] The exception is $\bar{v}$. Intuitively, a greater willingness to pay not only reduces the need for monetary incentives (the wedge ω) but increases their cost $(1+\kappa)y\bar{a}(x,y)$ to the Planner, by raising $\bar{a}$; as a result, she substitutes toward the use of publicity, which does not have such a cost.

VI. From Informal to Formal Institutions

A. *Norms Informing the Law*

In our motivating examples, we mentioned that laws often codify or reflect preexisting social norms, and that Principals who shape these laws (legislators, Supreme Court, etc.) aim to prescribe behaviors deemed appropriate in light of current "values" and mores. Laws and regulations that deviate too much from current preferences also generate significant distortions in terms of compliance burdens, enforcement costs, and evasion (black markets, corruption); see, e.g., Polinsky and Shavell (2007).

To formalize these ideas, we extend the model to have agents contribute twice, with a Principal who, instead of providing her own contribution a_P to the public good, *mandates a level of compliance* a^* (e.g., an emissions standard) that every agent must adhere to in the second period (which is a proxy for all subsequent interactions). In the first period, nothing is changed: the Principal sets publicity level $x \geq 0$, then each agent i chooses a_i with the same utility function as in Section I; note that because of the mandate, there will be no updating of reputations after period 1. As before, in setting x the Principal takes into account not only the costs and benefits of first-period public goods provision, but also what she will learn from the aggregate $\bar{a}$ about *how the law or mandate should be set.*

All the results, including the Principal's choice of publicity x^* and its comparative statics properties, remain closely analogous to the earlier ones (see Supplementary Appendix B). We also derive the optimal mandate for any x, whether exogenous or optimally chosen $(x = x^*)$:

$$a^* = \frac{\varphi}{\lambda}\big(w + E\big[\theta|\bar{a},\theta_P\big]\big) = \frac{\varphi}{\lambda}\Big(w + \big[1-\lambda(x)\big]\,\bar{\theta}_P + \gamma(x)\hat{\theta}\Big), \tag{35}$$

with $\hat{\theta}$, $\bar{\theta}_P$, and $\gamma(x)$ still defined as in Section IIIC. This makes clear (as for the matching rate in (25)) that *the law responds to the* (*descriptive*) *norm* $\bar{a}$—but less so, the less privately each individual chooses his action.

B. *Norms Informing Incentives*

Consider now the case where the key policy that the Principal wants to "get right" in light of possible shifts in societal preferences is an incentive rate. The law or mandate examined above was a limiting case where these ex post incentives take a drastic form—e.g., prohibitively high fines or prison sentences, which the Principal is somewhat able to deliver at very low cost. In practice, enforcement is costly, and many policies also take the form of subsidies, bonuses, taxes, etc., which entail nontrivial resource costs.

To deal with the question of *learning how incentives should be set*, we now consider the case of a Principal who (i) in period 1, sets publicity x to incentivize a first round of contributions and (ii) in period 2, instead of investing a_P herself, makes agents face an incentive rate $y \geq 0$ per unit of contribution, at an opportunity cost to

herself of $(1+\kappa)y$.[43] The paper's core insights apply again, linking in particular the optimal degree of publicity to the Principal's need for information about θ, agents' individual and collective knowledge about it, and the strength of their reputational concerns. Correspondingly, the form of the equilibrium solutions and all comparative statics of x^* remain unchanged. As to the optimal second-period incentive rate, it is

$$y^* = \hat{y} + \frac{(1+b) - \rho(1+\kappa)}{1+2\kappa} E\left[\theta|\theta_P, \bar{a}\right], \tag{36}$$

where $\hat{y}$ is a constant and $E\left[\theta|\theta_P, \bar{a}\right] = \left[1-\gamma(x^*)\right]\bar{\theta}_P + \gamma(x^*)\hat{\theta}$ is again the solution to the Principal's optimal-learning problem, given by (23)–(24) but for the new value of x^* (given in Appendix B, together with $\hat{y}$).

VII. Conclusion

The interaction between social norms and social learning faces institution designers with a tradeoff. Publicizing individual behaviors that constitute public goods (or bads) leads to more prosocial compliance on average, but low privacy also impedes the proper adaptation of societal standards and policies to emergent shifts in the overall distribution of preferences.

First, such imperfect knowledge renders publicity hard to fine-tune, generating inefficient variations in both individual and aggregate behavior. Second, leveraging social image concerns makes it even harder to infer from prevailing norms the true social value of the public good or conduct in question, in order to appropriately adapt policies and institutions. We showed in particular that where societal attitudes and/or technologies for monitoring and norm enforcement (e.g., social media) are prone to significant change, *a higher degree of privacy* is optimal. When average preferences over public goods and reputation remain relatively stable, conversely, the visibility of individual actions should be raised. We also derived the optimal mix of monetary (e.g., tax) and social image incentives and showed that all results extend to this more realistic and complex setting.

The framework is quite flexible, allowing for many extensions. For instance, in the literature on corporate culture, a key role of leaders is to coordinate expectations and efforts toward common objectives (Hermalin and Katz 2006; Bolton, Brunnermeier, and Veldkamp 2013). Our analysis adds a new dimension, namely the balancing act of exploiting peer pressure to align agents' incentives with the values and goals of the organization, while also allowing enough "quiet" contrarian behavior to learn how these should adapt over time.[44]

Another important extension would be to incorporate what social psychologists term *pluralistic ignorance*, namely the fact that agents themselves must often parse out how much of the prevailing mode of behavior around them is

[43] One could also combine ex ante incentives (chosen prior to observing compliance), as in Section V, with the ex post incentives studied here.

[44] Other roles for dissent in organizations arise in Landier, Sraer, and Thesmar (2009) and Bénabou (2012).

driven by deep preferences versus image motivations. In the current setup, society as a whole (benevolent Principal, next generation) faced this problem, but in equilibrium individuals ended up not having to, thanks to the benchmarking result. In a richer dynamic context than those we have considered, each agent will combine his idiosyncratic signals with past noisy observations of the prevailing norm in order to determine how to act next. This will better capture pluralistic ignorance and allow us to examine conditions under which it will persist.

APPENDIX A

A. *Main Proofs*

PROOFS OF PROPOSITIONS 1 AND 2:

Consider linear strategies of the form $a_i = A\mu_i + Bv_i + C\theta_i + D$, implying that $\bar{a} = A\mu + B\bar{v} + C\theta + D$. We first establish the following result.

CLAIM 1 (Benchmarking): *The expectation* $E\left[v_i|\theta_j, \mu_j, \bar{a}, a_i\right]$ *is independent of* $\left(\theta_j, \mu_j\right)$ *and equal to*

$$E\left[v_i|\theta_j, \mu_j, \bar{a}, a_i\right] = \bar{v} + \frac{Bs_v^2}{B^2 s_v^2 + C^2 s_\theta^2 + A^2 s_\mu^2}\left(a_i - \bar{a}\right). \tag{A1}$$

PROOF:

Subtracting $\bar{a}$ from a_i and rearranging, we obtain $Bv_i = B\bar{v} + \left(a_i - \bar{a}\right) - \left(C\epsilon_i^\theta + A\epsilon_i^\mu\right)$, where ϵ_i^θ and ϵ_i^μ denote $\theta_i - \theta$ and $\mu_i - \mu$ respectively. Observe that $\left(Bv_i, a_i - \bar{a}, \bar{a}, \theta_j, \mu_j, C\epsilon_i^\theta + A\epsilon_i^\mu\right)$ is jointly normally distributed: every linear combination of these components is a linear combination of a set of independent normal random variables and therefore has a univariate normal distribution. Because $\bar{a}$, θ_j, and μ_j are uncorrelated to both $C\epsilon_i^\theta + A\epsilon_i^\mu$ and $a_i - \bar{a}$ and these variables are jointly normally distributed, it follows from independence that

$$E\left[C\epsilon_i^\theta + A\epsilon_i^\mu|a_i, \bar{a}, \theta_j, \mu_j\right] = E\left[C\epsilon_i^\theta + A\epsilon_i^\mu|a_i - \bar{a}\right]. \tag{A2}$$

Observe that

$$\begin{pmatrix} v_i \\ a_i - \bar{a} \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{v} \\ 0 \end{pmatrix}, \begin{pmatrix} s_v^2 & Bs_v^2 \\ Bs_v^2 & B^2 s_v^2 + C^2 s_\theta^2 + A^2 s_\mu^2 \end{pmatrix}\right), \tag{A3}$$

and therefore, $E\left[v_i|\theta_j, \mu_j, \bar{a} - a_i\right]$ equals the expression in (A1). ∎

From Claim 1 it follows that

$$
\begin{aligned}
(\text{A4}) \quad & R(a_i, \theta_i, \mu_i) \\
& = E\left[E\left[v_i | a_i, \bar{a}\right] | \, \theta_i, \mu_i\right] \\
& = E\left[\left(\bar{v} + \frac{Bs_v^2}{A^2 s_\mu^2 + B^2 s_v^2 + C^2 s_\theta^2}\left(a_i - \bar{a}\right)\right) \middle| \, \theta_i, \mu_i\right] \\
& = \bar{v} + \frac{Bs_v^2\left(a_i - A\left\{\nu\mu_i + (1-\nu)\bar{\mu}\right\} - B\bar{v} - C\left\{\rho\theta_i + (1-\rho)\bar{\theta}\right\} - D\right)}{A^2 s_\mu^2 + B^2 s_v^2 + C^2 s_\theta^2},
\end{aligned}
$$

where $\nu \equiv \sigma_\mu^2 / \left(\sigma_\mu^2 + s_\mu^2\right)$. Utility maximization then yields the first-order condition

$$
(\text{A5}) \qquad a_i = v_i + \rho\theta_i + (1-\rho)\bar{\theta} + x\mu_i\left(\frac{Bs_v^2}{A^2 s_\mu^2 + B^2 s_v^2 + C^2 s_\theta^2}\right).
$$

Therefore, $B = 1$, $C = \rho$, $D = (1-\rho)\bar{\theta}$, and $A = xs_v^2 / \left(A^2 s_\mu^2 + s_v^2 + \rho^2 s_\theta^2\right)$. Substituting $A = x\xi(x)$ yields

$$
(\text{A6}) \qquad \xi(x) = \frac{s_v^2}{x^2 \xi(x)^2 s_\mu^2 + s_v^2 + \rho^2 s_\theta^2}.
$$

It remains to show that for each choice of x, $\xi(x)$ is unique. Given x, $\xi(x)$ solves the equation

$$
(\text{A7}) \qquad \xi = \frac{s_v^2}{x^2 \xi^2 s_\mu^2 + s_v^2 + \rho^2 s_\theta^2}.
$$

The right-hand side is continuous and decreasing in ξ, clearly cutting the diagonal at a unique solution $\xi(x)$. Furthermore, $\xi(x)$ must be strictly decreasing in x, strictly increasing in s_v^2, strictly decreasing in s_μ^2 and in σ_θ^2, and U-shaped in s_θ (noting that $\rho s_\theta = \sigma_\theta^2 / \left[s_\theta + \sigma_\theta^2 / s_\theta\right]$).

To derive comparative statics, note that $\beta(x) = x\xi(x)$ solves the implicit equation

$$
(\text{A8}) \qquad x = \beta\left[\beta^2\left(s_\mu^2 / s_v^2\right) + \rho^2 s_\theta^2 / s_v^2 + 1\right],
$$

which makes clear that $\beta(x)$ is strictly increasing in x, with $\lim_{x\to\infty} \beta(x) = \infty$. ∎

PROOF OF PROPOSITION 3:

For each agent i, $a_i = x\xi\mu + v_i + \rho\theta_i + (1-\rho)\bar{\theta}$, and therefore $\bar{a}(\theta, \mu) \equiv x\xi\mu + \bar{v} + \bar{\theta} + \rho\left(\theta - \bar{\theta}\right)$. Let $\bar{\bar{a}} \equiv x\xi\bar{\mu} + \bar{v} + \bar{\theta}$ represent the expected aggregate contribution.

Since the Principal observes μ, she can infer θ perfectly from $\bar{a}$ and so will set $a_P = (w+\theta)\varphi/(1-\lambda)k_P$ independently of x (recall that $\varphi \equiv \lambda + b(1-\lambda)$). Let us define $\bar{a}_P \equiv (w+\bar{\theta})\varphi/(1-\lambda)k_P$ as the expected Principal's contribution. Integrating (4) over θ and μ, we have

$$
\begin{aligned}
(\text{A9}) \qquad E\tilde{V}(x) \;=\; & \lambda\Big[\alpha\big(s_v^2 + \rho\sigma_\theta^2 + (\bar{v}+\bar{\theta})(\bar{\bar{a}})\big) + (w+\bar{\theta})(\bar{\bar{a}} + \bar{a}_P)\Big] \\
& + \lambda\left[\rho\sigma_\theta^2 + \frac{\lambda\sigma_\theta^2\varphi}{(1-\lambda)k_P}\right] \\
& + (1-\lambda)b\left[(w+\bar{\theta})(\bar{\bar{a}} + \bar{a}_P) + \rho\sigma_\theta^2 + \frac{\sigma_\theta^2\varphi}{(1-\lambda)k_P}\right] \\
& - \frac{\lambda}{2}\Big[\bar{\bar{a}}^2 + \rho^2\big(\sigma_\theta^2 + s_\theta^2\big) + s_v^2 + x^2\xi^2\sigma_\mu^2\Big] \\
& - \frac{(1-\lambda)k_P}{2}\left[\bar{a}_P^2 + \sigma_\theta^2\left(\frac{\varphi}{(1-\lambda)k_P}\right)^2\right].
\end{aligned}
$$

Differentiating yields

$$
(\text{A10}) \qquad \frac{dE\tilde{V}(x)}{dx} \;=\; \omega\xi\bar{\mu} - \lambda x\xi^2\big(\bar{\mu}^2 + \sigma_\mu^2\big),
$$

where we recall that $\omega \equiv (w+\bar{\theta})\varphi - \lambda(1-\alpha)(\bar{v}+\bar{\theta})$ as defined in (6). For all $\lambda > 0$, the expression is strictly concave in x, so the first-order condition yields the unique optimum, given by (19); when $\sigma_\mu^2 = 0$, it simplifies to (16).

The formula for $m(x)$ was shown in the text. For the baseline investment, it is

$$
(\text{A11}) \quad \underline{a}_P(x,\theta_P) \;=\; \frac{\varphi\Big(w + \big(1-\gamma(x)\big)E\big[\theta|\theta_P\big] - \frac{\gamma(x)}{\rho}\big(\bar{v} + x\xi\bar{\mu} + (1-\rho)\bar{\theta}\big)\Big)}{(1-\lambda)k_P},
$$

which follows from the same equations, together with (21). ∎

PROOF OF PROPOSITION 5:

For every θ, were the Principal to observe θ or the realization of μ, she would choose $a_P = (w+\theta)\varphi/(1-\lambda)k_P$. When she is unable to observe θ or μ, she sets $a_P = \big(w + E\big[\theta|\bar{a},\theta_P\big]\big)\varphi/(1-\lambda)k_P$, which makes clear how the forecast error $\Delta \equiv E\big[\theta|\bar{a},\theta_P\big] - \theta$, derived in (26), distorts her contribution from the full-information optimal by $\varphi\Delta/\big((1-\lambda)k_P\big)$. Given the quadratic loss from setting the right level of contributions, it is straightforward to show that the loss induced in

her payoffs from the full-information benchmark is then $\left(\varphi^2/\left(2(1-\lambda)k_P\right)\right)E\left[\Delta^2\right]$, where

$$E\left[\Delta^2\right] = \left(1-\gamma\right)^2\sigma_P^2 + \left(\gamma\xi x/\rho\right)^2\sigma_\mu^2 = \sigma_P^2\left[\left(1-\gamma\right)^2 + \gamma^2\left(1/\gamma - 1\right)\right] = \sigma_P^2\left(1-\gamma\right), \tag{A12}$$

where we abbreviated $\gamma(x)$ as γ and used the fact that $x^2\xi^2\sigma_\mu^2/\rho^2 = \sigma_P^2(1-\gamma)/\gamma$.

Therefore, it follows that (27) characterizes the change in payoffs from information distortion. Note also that

$$\frac{\sigma_P^2}{2}\frac{d\gamma}{dx} = -\frac{\sigma_P^2}{2}\left(\frac{2\rho^2\sigma_P^2\xi^2\sigma_\mu^2}{\left(\rho^2\sigma_P^2 + x^2\xi^2\sigma_\mu^2\right)^2}x\right) = -\frac{\sigma_P^2\gamma\left(1-\gamma\right)}{x} = -\sigma_P^2\left(\frac{\gamma^2\xi^2\sigma_\mu^2}{\rho^2\sigma_P^2}x\right) = -\frac{\sigma_\mu^2\gamma^2\xi^2 x}{\rho^2}. \tag{A13}$$

Therefore

$$\begin{aligned}\frac{\partial EV}{\partial x} &= \frac{\partial E\tilde{V}}{\partial x} - \frac{\varphi^2}{\left(1-\lambda\right)k_P}\left(\frac{\sigma_\mu^2\gamma^2\xi^2 x}{\rho^2}\right)\\ &= \left(\xi\bar{\mu}\right)\left[\left(w+\bar{\theta}\right)\varphi - \left(\bar{v}+\bar{\theta}\right)\left(1-\alpha\right)\lambda\right]\\ &\quad -\lambda x\xi^2\left(\bar{\mu}^2+\sigma_\mu^2\right) - \frac{\varphi^2}{\left(1-\lambda\right)k_P}\left(\frac{\sigma_\mu^2\gamma^2\xi^2 x}{\rho^2}\right),\end{aligned} \tag{A14}$$

which corresponds to (29). Recall now that $E\tilde{V}(x)$ is strictly concave in x and maximized at $\tilde{x} > 0$. Therefore, $\partial EV/\partial x < \partial E\tilde{V}/\partial x \leq 0$ for all $x \geq \tilde{x}$, and at $x = 0$, $\partial EV/\partial x = \partial E\tilde{V}/\partial x > 0$. Consequently, the global maximum of EV on $\mathbb{R}$ is reached at some $x^* \in \left(0,\tilde{x}\right)$ where $\partial EV/\partial x = 0$. ∎

PROOF OF PROPOSITION 6:

We proceed again in two stages, starting with the benchmark of "symmetric uncertainty" where the Principal learns the realization of (the average) μ after x has been set. Then, we incorporate the information-distortion effect.

CLAIM 2: *When the Principal faces no ex post uncertainty about μ and observes it perfectly, she sets a publicity level $\tilde{x}^{SI}$ given by the unique solution to*

$$\tilde{x}^{SI} = \frac{\bar{\mu}\omega}{\lambda\xi\left(\tilde{x}^{SI}\right)\left(\bar{\mu}^2 + \sigma_\mu^2 + \left(1-2\tilde{\alpha}\right)s_\mu^2\right)}. \tag{A15}$$

PROOF:

Proposition 1 shows that given any x, the equilibrium among agents is the same as in the case where $s_\mu^2 = 0$, except that ξ is replaced everywhere by $\xi(x)$, or equivalently $x\xi$ by $\beta(x) = x\xi(x)$ in all type-independent expressions (first and second moments), while at the individual level $\mu x\xi$ is replaced by $\mu_i\beta(x)$.

Let us denote by $a_i^0 \equiv v_i + \rho\theta_i + (1-\rho)\bar{\theta} + \mu x\xi(x)$ the value of a_i corresponding to the mean value of $\mu_i = \mu$, or equivalently the value of a_i in the original (homogeneous μ) model where we simply replace ξ by $\xi(x)$. Similarly, let $\tilde{V}^0(x)$ (respectively, $V^0(x)$) be the utility level the Principal would achieve if agents behaved according to a_i^0 and she observes (respectively, does not observe) the realization of the average μ.

We can obtain $E\tilde{V}^0(x)$ directly by replacing ξ with $\xi(x)$ in the expression (A9) giving $EV(x)$, and similarly $dEV^0(x)/dx$ by replacing $x\xi$ with $\beta(x)$ and ξ with $\beta'(x)$ in the expression (A10) for $dEV(x)/dx$:

$$\frac{dE\tilde{V}^0(x)}{dx} = \omega\bar{\mu}\beta'(x) - \lambda\beta'(x)\beta(x)\left[\bar{\mu}^2 + \sigma_\mu^2\right] = 0. \tag{A16}$$

In the Principal's actual loss function (4), the heterogeneity in agents' μ_is generates an additional loss of $(\lambda/2)E\left[(a_i)^2 - (a_i^0)^2\right] = (\lambda/2)\beta(x)^2 s_\mu^2$ due to inefficient cost variations, but also a gain from their image seeking equal to $\lambda\tilde{\alpha}\beta(x)^2 s_\mu^2$. Therefore, when the Principal observes the realization of μ, the optimal (symmetric information) value of x is given by the first-order condition

$$\frac{dE\tilde{V}}{dx} = \omega\bar{\mu}\beta'(x) - \lambda\beta'(x)\beta(x)\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right) = 0 \tag{A17}$$

$$\Rightarrow \beta(\tilde{x}^{SI}) = \frac{\bar{\mu}\omega}{\lambda\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right)}, \tag{A18}$$

which is equivalent to (A15). Note that $\tilde{x}^{SI} < \infty$ as long as $\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2 > 0$, which is is automatically satisfied if (i) $\tilde{\alpha} < 1/2$ or (ii) $s_\mu^2 < \bar{\mu}^2 + \sigma_\mu^2$. ∎

We now extend the results to the case where the Principal does not know μ when setting her contribution. Corollary 1 in Section IIID allows us to simply combine (A17) and (27) to obtain the relevant version of her first-order condition:

$$\begin{aligned}\frac{dEV}{dx} &= \beta'(x)\left[\bar{\mu}\omega - \lambda\beta(x)\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right)\right] + \frac{\varphi^2\sigma_P^2}{2(1-\lambda)k_P}\gamma'(x) \\ &= 0.\end{aligned} \tag{A19}$$

Using (24) we have $\gamma'(x) = -\left[2\sigma_\mu^2/\rho^2\sigma_P^2\right]\beta(x)\beta'(x)\gamma(x)^2$, leading to

$$\beta(x^*) = \frac{\bar{\mu}\omega}{\lambda\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right) + \frac{1}{(1-\lambda)k_P}\left(\frac{\varphi\sigma_\mu\gamma(x)}{\rho}\right)^2}, \tag{A20}$$

which is equivalent to (32). ∎

PROOF OF PROPOSITION 7:

Denote $x\xi(x)$ by z, and note that $EV(x)$ can be reformulated as

$$\mathcal{V}(z) = s_v^2\Big(\lambda\alpha - \frac{\lambda}{2}\Big) + z\bar{\mu}\omega - \frac{\lambda}{2}z^2\big(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})\,s_\mu^2\big) - \frac{\varphi^2\sigma_P^2}{2(1-\lambda)k_P}\big[1-\tilde{\gamma}(z)\big] + C, \tag{A21}$$

in which $\tilde{\gamma}(z) \equiv \rho^2\sigma_P^2\big/\big[\rho^2\sigma_P^2 + z^2\sigma_\mu^2\big]$ and C is a constant that is independent of s_v^2 and z. The optimal z solves the first-order condition:

$$\bar{\mu}\omega - \lambda z\big(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\big) + \frac{\varphi^2\sigma_P^2}{2(1-\lambda)k_P}\tilde{\gamma}'(z) = 0. \tag{A22}$$

Notice that none of these terms depend on s_v^2, and so the optimal z is independent of s_v^2. Therefore, for each s_v, the optimal $x^*(s_v)\xi(x^*(s_v),s_v)$ is constant. This fact automatically implies that in equilibrium, changes in s_v^2 do not influence γ or the matching rate.

Because the Principal maintains constancy of $x^*(s_v)\xi(x^*(s_v),s_v)$, (A6) implies that a higher s_v strictly increases $\xi(x^*(s_v),s_v)$. Therefore, $x^*(s_v)\xi(x^*(s_v),s_v)$ remains unchanged only if $x^*(s_v)$ is decreasing in s_v. Finally, (A21) implies that $d\Big[EV\big(x^*\big(s_v^2\big);s_v^2\big)\Big]\big/ds_v^2 = \lambda\big(\alpha - 1/2\big)$. ∎

PROOF OF PROPOSITION 8:

Setting $\rho = 1$ in (30), $\partial^2 EV/\partial x\partial\sigma_\theta < 0$ implies that x^* is decreasing in σ_θ^2. Decreasing x decreases $\beta(x)$ (recall that in this limiting case, $\beta(x)$ is independent of σ_θ^2), so if $\gamma\big(x^*;\sigma_\theta\big)$ did not increase with σ_θ, the right-hand side of (30) could not remain equal to 0. Thus, at the optimal x^*, $\gamma\big(x^*;\sigma_\theta\big)$ must increase with σ_θ. ∎

PROOF OF PROPOSITION 9:

The negative impact of increasing σ_μ^2 on payoffs is clear: for every θ and x, changes in σ_μ^2 have no effect on $\bar{a}$, but increase the variance of aggregate contributions and the information cost. To consider their impact on optimal publicity, observe from (30) that

$$\frac{\partial^2 EV}{\partial x\partial\sigma_\mu^2} = -\lambda\beta'(x)\beta(x) - \frac{\varphi^2\beta'(x)\beta(x)}{\rho^2(1-\lambda)k_P}\Big(\gamma^2 + 2\gamma\sigma_\mu^2\frac{d\gamma}{d\sigma_\mu^2}\Big), \tag{A23}$$

in which

$$\frac{\partial\gamma}{\partial\sigma_\mu^2} = -\frac{\rho^2\sigma_\theta^2\beta(x)^2}{\big(\rho^2\sigma_\theta^2 + \beta(x)^2\sigma_\mu^2\big)^2} = -\frac{\gamma\beta(x)^2}{\rho^2\sigma_\theta^2 + \beta(x)^2\sigma_\mu^2} = -\frac{\gamma(1-\gamma)}{\sigma_\mu^2}. \tag{A24}$$

Thus,

$$\frac{\partial^2 EV}{\partial x \partial \sigma_\mu^2} = -\beta'(x)\beta(x)\left[\lambda + \frac{\varphi^2\gamma^2}{\rho^2(1-\lambda)k_P}(2\gamma - 1)\right]. \tag{A25}$$

This expression is nonpositive if and only if

$$\frac{\lambda(1-\lambda)\rho^2 k_P}{\varphi^2} \geq \gamma^2(1-2\gamma). \tag{A26}$$

Because $\gamma^2(1-2\gamma)$ takes on a maximum value of $1/27$, a sufficient condition is that the left-hand side of the equation above exceeds $1/27$. In this case, $\partial x/\partial\sigma_\mu^2 < 0$ for all values of σ_μ. Intuitively, when k_P is large enough, the value of information for the Principal is small (she does not have much of a decision to make), so whether a higher σ_μ^2 improves or worsens the information effect, it is dominated by its worsening of the variance effect.

If the condition is not satisfied, then monotonicity generally does not hold everywhere, but:

(i) As σ_μ^2 tends to 0, $\gamma\left(x^*\left(\sigma_\mu^2\right);\sigma_\mu^2\right)$ approaches 1, because by Proposition 5, $x^*\left(\sigma_\mu^2\right)$ remains bounded above: $x^*\left(\sigma_\mu^2\right) < \bar{x}$. Therefore, (A26) holds for σ_μ small enough.

(ii) As σ_μ^2 tends to ∞, $x^*\left(\sigma_\mu^2\right)$ must tend to 0 fast enough that the product $\sigma_\mu^2 x^*\left(\sigma_\mu^2\right)$ remains bounded above. Otherwise, equation (28) shows that the first-order condition $\partial EV/\partial x = 0$ cannot hold, as the marginal variance effect and the marginal information-distortion effects both become arbitrarily large. It then follows that that $\sigma_\mu^2\left[x^*\left(\sigma_\mu^2\right)\right]^2$ tends to 0, and therefore $\gamma\left(x^*\left(\sigma_\mu^2\right);\sigma_\mu^2\right)$ tends to 1. Thus, for σ_μ^2 large enough, (A26) holds, and $x^*\left(\sigma_\mu^2\right)$ decreases to 0. ∎

PROOF OF PROPOSITION 10:

Since x enters EV only through $\beta(x) = x\xi(x)$, the Principal's problem is again equivalent to optimizing over the value of β, so the indirect effects of s_μ^2 on the optimized objective function $EV\left(x^*\left(s_\mu^2\right), s_\mu^2\right)$ cancel out at the first order, leaving only the direct effect $\left(-\lambda/2\right)\beta(x)^2(1-2\tilde{\alpha})$, which is less than 0 if and only if $\tilde{\alpha} < 1/2$. ∎

PROOF OF PROPOSITION 12:

As $\sigma_\theta \to \infty$, ρ converges to 1 and, therefore, $\xi(x, s_\theta)$ converges to a solution to the equation

$$\xi = \frac{s_v^2}{x^2\xi^2 s_\mu^2 + s_v^2 + s_\theta^2} \tag{A27}$$

for each x. Note that this must be strictly decreasing in s_θ^2. By inspection, x and $\xi(x)$ enter all terms in (30) only through their product $\beta(x)$. Therefore, to study how the optimal $x^*(s_\theta)$ and the Principal's welfare depend on s_θ^2, we can follow the same steps as in the proof of Proposition 7, leading to $d\left[EV\left(x^*(s_\theta);s_\theta\right)\right]/ds_\theta = -\lambda s_\theta < 0$. Finally, since the Principal keeps $x^*(s_\theta)\xi\left(x^*(s_\theta),s_\theta\right)$ constant as s_θ increases, it follows from (A27) that $\xi\left(x^*(s_\theta),s_\theta\right)$ must decrease in s_θ. To compensate, $x^*(s_\theta)$ must then be increasing in s_θ. ∎

B. *Analysis of Private Values in Section IIIE*

The Principal cares about θ as the average sentiment toward the public good, but also about agents' individual utilities and how these are matched to their actions. Her final payoff is

$$\begin{aligned} V^P \equiv \lambda\Big[&(w+\theta)(\bar{a}+a_P) - \int_0^1 C(a_i)\,di \\ &+ \alpha\int_0^1 (v_i+\theta_i)a_i\,di + \tilde{\alpha}\int_0^1 x\mu_i\left[R(a_i,\theta_i,\mu_i) - \bar{v}\right]di\Big] \\ &+ (1-\lambda)\left[b(w+\theta)(\bar{a}+a_P) - k_P C(a_P)\right]. \end{aligned} \tag{A28}$$

We first describe how agents behave under a given value of x, then characterize the optimal degree of publicity and its comparative statics.

PROPOSITION 14 (Equilibrium Behavior and Benchmarking): *Fix $x \geq 0$. All properties are identical to Proposition* 1, *except that ρ is replaced everywhere by the number* 1.

PROOF:

Consider linear strategies of the form $a_i = A\mu_i + Bv_i + C\theta_i + D$, implying that $\bar{a} = A\mu + B\bar{v} + C\theta + D$. From Claim 1 it follows that

$$\begin{aligned} &R(a_i,\theta_i,\mu_i) \\ &\quad = \bar{v} + \frac{Bs_v^2\left[a_i - A\left\{\nu\mu_i + (1-\nu)\bar{\mu}\right\} - B\bar{v} - C\left(\rho\theta_i + (1-\rho)\bar{\theta}\right) - D\right]}{A^2s_\mu^2 + B^2s_v^2 + C^2s_\theta^2}, \end{aligned} \tag{A29}$$

where $\nu \equiv \sigma_\mu^2/\left(\sigma_\mu^2 + s_\mu^2\right)$. Utility maximization then yields the first-order condition:

$$a_i = v_i + \theta_i + x\mu_i\left(\frac{Bs_v^2}{A^2s_\mu^2 + B^2s_v^2 + C^2s_\theta^2}\right). \tag{A30}$$

Therefore, $B = C = 1$, $D = 0$, and $A = xs_v^2/\left(A^2s_\mu^2 + s_v^2 + s_\theta^2\right)$. Substituting $A = x\tilde{\xi}(x)$ yields the expression in (10) but with ρ replaced by 1. It remains to

show that for each choice of x, $\tilde{\xi}(x)$ is unique. Given x, $\tilde{\xi}(x)$ solves the equation $\xi\left(x^2\xi^2 s_\mu^2 + s_v^2 + s_\theta^2\right) = s_v^2$; the right-hand side is continuous and decreasing in ξ, clearly cutting the diagonal at a unique solution $\xi(x)$. ∎

Setting $\rho = 1$ into the expression for $\xi(x)$ yields the relevant reputational reward payoff, $\tilde{\xi}(x)$, and generates the following comparative statics.

PROPOSITION 15 (Comparative Statics of Social Interactions): *All comparative statics are identical to Proposition* 2, *except that* $\tilde{\xi}(x)$ *is decreasing in* s_θ^2.

The only difference with the common-values case is that $\tilde{\xi}$ is now monotonically decreasing in s_θ^2, since this variance now corresponds to a motive for contributing that is orthogonal to the v_is. Observe, finally, that $\tilde{\xi}(x)$ is the same as in (10) with ρ simply replaced by 1.

Principal's Problem.—Her problem is unchanged but for relevant substitutions: setting $\rho = 1$ and adding to her payoff the term $\lambda\alpha s_\theta^2$, which arises from internalizing the gain resulting (by convexity) from the dispersion of contributions motivated by heterogeneous private values. Since this constant is independent of x and a_P, however, it plays no role in the analysis, and the solution to the Principal's problem is simply the same as in the common-value environment but with ρ set to 1 in all the results.

C. *Results for Section V*

Consider how the Principal's objective function (4) is affected by the presence of monetary incentives. First, the aggregate reputational gains term (multiplying $\tilde{\alpha}$) is unchanged, since the presence of a known y does not affect anyone's image. Second, for the aggregate intrinsic motivation term (multiplying α), the question is whether or not agents derive intrinsic satisfaction from the part of their contribution that they know is simply a response to monetary incentives. There is no correct "in principle" answer to that question, nor much available evidence. For simplicity (and without affecting any important results), we will therefore abstract from this term in what follows, assuming that agents do not get additional intrinsic utility of the form $v_i y$ (or, equivalently, that $\alpha = 0$).

The only new terms that appear when agents are paid y and change their behaviors from $a_i(x)$ to $a_i(x) + y$ and the Principal incurs cost $(1+\kappa)\left(\bar{a}(x) + y\right)y$ are thus the following:

$$\text{(A31)}\quad \tilde{V}(x,y,a_P) - V(x,a_P) = (w+\theta)\varphi y - \frac{1}{4}\int\left[\left(a_i(x)+y\right)^2 - a_i(x)^2\right]di - \frac{1}{2}\kappa y\left[\bar{a}(x)+y\right],$$

where we focus on the benchmark case of a social planner ($\lambda = 1/2$).

In the initial period, the Principal optimizes $E\left[\tilde{V}(x,y,a_P)\right]$ over (x,y) conditional on her priors, knowing that she will later choose a_P optimally given agents' behavior

and what she will have learned from it. We can therefore use the Envelope Theorem and neglect at this stage the dependence of a_P on (x,y). Maximizing over y yields $E\big[2(w+\theta)\varphi - y - \bar{a}(x) - 2\kappa y - \kappa\bar{a}(x)\big] = 0$ for an interior optimum, or

$$y = \frac{2\big(w+\bar{\theta}\big)\varphi - (1+\kappa)\bar{a}(x)}{1+2\kappa} \equiv \tilde{y} - \tau\beta(x), \quad \text{where} \tag{A32}$$

$$\tilde{y} \equiv \frac{2\omega - \kappa\big(\bar{v}+\bar{\theta}\big)}{1+2\kappa}; \quad \tau \equiv \bar{\mu}\left(\frac{1+\kappa}{1+2\kappa}\right). \tag{A33}$$

If $\tilde{y} - \tau\beta(x) < 0$, on the other hand, the corner solution $y(x) = 0$ is optimal.

(i) *Incentive and Variance Effects.*—Consider first the benchmark of Sections IIIA–IIIB, in which the Principal will learn μ before choosing a_P (no information-distortion effect) and all agents share the same value for social image ($s_\mu^2 = 0$). In this case, the optimal policy mix (x^*, y^*) of publicity and material incentives can be solved for explicitly.

Since x enters each $a_i(x)$ only through $\mu_i\beta(x)$, the first-order condition for x in (A31) is

$$\begin{aligned}\frac{\partial EV(x,a_P)}{\partial x} - \frac{1}{2}(1+\kappa)\bar{\mu}y\beta'(x) \\ = \xi\bar{\mu}\omega - \frac{1}{2}x\xi^2\big(\bar{\mu}^2+\sigma_\mu^2\big) - \frac{1}{2}(1+\kappa)\bar{\mu}y^*\xi \le 0,\end{aligned} \tag{A34}$$

with equality if $x^* > 0$. Recall now that with $s_\mu^2 = 0$, $\xi(x)$ reduces to the constant ξ given by (14), so $\beta(x) = x\xi$. Together with (A10), this yields

$$x^* = \frac{\bar{\mu}\big[\omega - (1/2)(1+\kappa)y^*\big]}{(1/2)\xi\big(\bar{\mu}^2+\sigma_\mu^2\big)} \equiv \frac{\bar{\mu}\tilde{\omega}}{(1/2)\xi\big(\bar{\mu}^2+\sigma_\mu^2\big)} \tag{A35}$$

when this is nonnegative, otherwise $x^* = 0$. This last case, however, requires that $y^* = \tilde{y}$ by (A32) and $y \ge 2\omega/(1+\kappa)$ by (A34), so it only occurs for $\kappa = 0$. Comparing to (19) in the main text, we see that the wedge has been reduced from ω to $\tilde{\omega} \equiv \omega - (1+\kappa)y^*/2$. To solve for y^*, finally, substitute x^* into $y^* = \tilde{y} - \tau\xi x^*$. Straightforward but tedious derivations yield

$$y^* = \frac{2\omega\big(\sigma_\mu^2 - \kappa\bar{\mu}^2\big) - \kappa\big(\bar{v}+\bar{\theta}\big)\big(\bar{\mu}^2+\sigma_\mu^2\big)}{(1+2\kappa)\sigma_\mu^2 - \kappa^2\bar{\mu}^2}, \tag{A36}$$

an explicit solution that is nonnegative as long as

$$\kappa \le \bar{\kappa} \equiv \frac{2\omega\sigma_\mu^2}{2\bar{\mu}^2\omega + \big(\bar{\mu}^2+\sigma_\mu^2\big)\big(\bar{v}+\bar{\theta}\big)}. \tag{A37}$$

Indeed, the denominator of (A36) is a negative quadratic in κ that is maximized at $\kappa^* = \sigma_\mu^2/\bar{\mu}^2$ and strictly positive at $\kappa = 0$, hence strictly positive for all $\kappa < \sigma_\mu^2/\bar{\mu}^2$, and thus a fortiori for all $\kappa < \bar{\kappa} = \left(\sigma_\mu^2/\bar{\mu}^2\right)\left[1 + (\bar{\mu}^2 + \sigma_\mu^2)\left(\bar{\nu} + \bar{\theta}\right)/2\omega\bar{\mu}^2\right]^{-1} < \sigma_\mu^2/\bar{\mu}^2$.

The unique solution to the joint maximization over (x, y) is therefore (i) for $\kappa \leq \bar{\kappa}$, y^* given by (A36) and x^* given by (A35), and (ii) for $\kappa \geq \bar{\kappa}$, $y^* = 0$ and $x^* = \left(2\bar{\mu}\omega/\xi\right)/\left(\bar{\mu}^2 + \sigma_\mu^2\right)$, as in the original model. Substituting y^* into the effective wedge, $\tilde{\omega} = \omega - \left(1/2\right)\left(1 + \kappa\right)y^*$, and the latter into (A35) yields an explicit formula for x^*:

$$x^* = \begin{cases} \dfrac{\kappa\bar{\mu}\left(\omega + (1/2)(1+\kappa)\left(\bar{\nu} + \bar{\theta}\right)\right)}{(1/2)\xi\left((1+2\kappa)\,\sigma_\mu^2 - \kappa^2\bar{\mu}^2\right)} & \text{for } \kappa < \bar{\kappa} \\ \dfrac{\bar{\mu}\omega}{(1/2)\xi\left(\bar{\mu}^2 + \sigma_\mu^2\right)} & \text{for } \kappa \geq \bar{\kappa} \end{cases}. \tag{A38}$$

The comparative statics of x^* (with associated m^*) and y^* could be obtained directly from (A38)–(A36) but will be proven below for the more general model. The only one specific to the present case (i.e., not in Tables 1–2) is that for $\bar{\mu}$. As long as $\kappa < \bar{\kappa}$, x^* is clearly increasing in $\bar{\mu}$. Note, however, that $\bar{\kappa}$ is decreasing in $\bar{\mu}$; thus, beyond some threshold $\bar{\mu}^*$ we will have $\bar{\kappa} < \kappa$, and from there on, x^* will be hill shaped in $\bar{\mu}$. Thus, as in the original model, x^* continues to exhibit an inverse U-shaped relationship with $\bar{\mu}$.

(*ii*) *Information Distortion and Heterogeneous Image Concerns*.—For the more general problem with information distortion (and heterogeneous image concern), the first-order condition (A34) takes the form (as long as $y = \tilde{y} - \tau\beta(x) > 0$)

$$0 = \omega\bar{\mu} - \frac{1}{2}(1+\kappa)\bar{\mu}\left[\tilde{y} - \tau\beta(x)\right] - \frac{1}{2}\beta(x)\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})\,s_\mu^2\right) - \frac{2\varphi^2\sigma_\mu^2}{\rho^2 k_P}\beta(x)\gamma(x)^2. \tag{A39}$$

Relative to the original model, we see that the wedge ω is again reduced to $\tilde{\omega} \equiv \omega - \left(1/2\right)\left(1 + \kappa\right)y^*$. Given that $\beta(x)$ now equals $x\xi^*(x)$, the analogue of (32) is thus

$$x^* = \frac{\mu}{\xi(x^*)}\left(\frac{\tilde{\omega}}{\frac{1}{2}\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})\,s_\mu^2\right) + \frac{2\left(\varphi\sigma_\mu\gamma(x^*)/\rho\right)^2}{k_P}}\right). \tag{A40}$$

As κ grows large enough, it will again be the case that $y^* = 0$ and x^* reduces to the benchmark case. Moreover, y^* strictly decreases with κ, while x^* strictly increases. We now formally prove these properties and the claimed comparative statics of the

optimal first-period policy mix, (x^*, y^*). Those of the second-period m^* readily follow as in the main case.

PROOF OF PROPOSITION 13:

Denote the Principal's objective function $E\tilde{V}$ (with $\tilde{V}$ given by (A31)) as $V(x, y, \Theta)$, where Θ is the vector of all the model's parameters; $\beta(x)$, $\gamma(x)$, and $T(x) \equiv \beta(x)\gamma(x)^2$ also depend on some components of Θ, but we shall not make this explicit to lighten the notation. Denoting partial derivatives by subscripts, the system that implicitly defines the optimum policy (x^*, y^*) is

$$\text{(A41)} \quad \mathcal{V}_y(x, y, \Theta) = \omega - \frac{1}{2}\kappa\left(\bar{v} + \bar{\theta}\right) - \frac{1}{2}\bar{\mu}(1+\kappa)\beta(x) - \frac{1}{2}(1+2\kappa)y \leq 0,$$

$$\text{(A42)} \quad \mathcal{V}_x(x, y, \Theta) = \beta'(x)\left[\bar{\mu}\omega - \frac{1}{2}(1+\kappa)\bar{\mu}y\right] - \beta'(x)\beta(x)\left[\frac{\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2}{2} + \frac{2\varphi^2\sigma_\mu^2\gamma^2(x)}{\rho^2 k_P}\right] \leq 0,$$

with the complementary slackness conditions $y\,V_y(x, y, \Theta) = x\,V_x(x, y, \Theta) = 0$.

Note first that except at $\kappa = 0$, it cannot be that $y^* > 0 = x^*$; otherwise (A42) yields $y^* = \left[2\omega - \kappa\left(\bar{v} + \bar{\theta}\right)\right] / (1+2\kappa) = \tilde{y}$ and (A41) requires $2\omega / (1+\kappa) \leq y$, a contradiction for any $\kappa > 0$. Next, where $y^* = 0$, the model reduces to the basic one, for which Tables 1–2 provide all the comparative statics on x^* (and m^*). Focusing now on the region where $x^*, y^* > 0$, the comparative statics for an arbitrary parameter η using the Implicit Function Theorem:

$$\text{(A43)} \quad \begin{pmatrix} \dfrac{\partial y^*}{\partial \eta} \\ \dfrac{\partial x^*}{\partial \eta} \end{pmatrix} = H^{-1}(x^*, y^*)\begin{pmatrix} -\mathcal{V}_{y\eta} \\ -\mathcal{V}_{x\eta} \end{pmatrix} = \frac{1}{|H|}\begin{pmatrix} \mathcal{V}_{xx} & -\mathcal{V}_{xy} \\ -\mathcal{V}_{xy} & \mathcal{V}_{yy} \end{pmatrix}\begin{pmatrix} -\mathcal{V}_{y\eta} \\ -\mathcal{V}_{x\eta} \end{pmatrix},$$

where the Hessian matrix of the system (A41)–(A42),

(A44)

$$H = -\begin{pmatrix} \frac{1}{2}(1+2\kappa) & \frac{1}{2}(1+\kappa)\bar{\mu}\beta'(x) \\ \frac{1}{2}(1+\kappa)\bar{\mu}\beta'(x) & \beta'(x)\left[\dfrac{\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})\,s_\mu^2\right)\beta'(x)}{2} + \dfrac{2\varphi^2\sigma_\mu^2 T'(x)}{\rho^2 k_P}\right] \end{pmatrix},$$

must be negative definite, since (y^*, x^*) is a strict local maximum. Therefore $V_{yy} < 0$, $V_{xx} > 0$, and the determinant of H must be positive, $|H| > 0$. Note also that $V_{xy} < 0$, reflecting the strategic substitutability of the two policy instruments.

(*i*) *Comparative Statics with Respect to* κ.—From (A43)–(A44), we have

$$\text{(A45)}\qquad \begin{pmatrix} \frac{\partial y^*}{\partial \kappa} \\ \frac{\partial x^*}{\partial \kappa} \end{pmatrix} = H^{-1}(x^*, y^*) \begin{pmatrix} y^* + \frac{1}{2}\bar{\mu}\beta(x^*) + \frac{1}{2}(\bar{\nu} + \bar{\theta}) \\ \frac{1}{2}\bar{\mu} y^* \beta'(x) \end{pmatrix}.$$

Solving for $\partial x^* / \partial \kappa$ gives

$$\text{(A46)}\quad |H| \frac{\partial x^*}{\partial \kappa} = \frac{(1+\kappa)\bar{\mu}\beta'(x)}{2}\left(y^* + \frac{\bar{\mu}\beta(x^*)}{2} + \frac{\bar{\nu} + \bar{\theta}}{2}\right) - \frac{(1+2\kappa)\bar{\mu} y^* \beta'(x^*)}{4}$$

$$= \frac{1}{4}\beta'(x^*)\left\{ \bar{\mu} y^* + \bar{\mu}(1+\kappa)\left[\bar{\mu}\beta(x^*) + \bar{\nu} + \bar{\theta}\right] \right\} > 0.$$

Therefore, over any range where $y^* > 0$, x^* is strictly increasing in κ; together with the fact that $\Phi(x, y, \kappa)$ is decreasing in κ for all (x, y), this implies that y^* must decrease in κ wherever $y^* > 0$. Therefore, there exists a $\bar{\kappa} \in \left(0, 2\omega / (\bar{\nu} + \bar{\theta})\right)$ such that $y^* > 0$ on $[0, \bar{\kappa})$ and $y^* = 0$ on $[\bar{\kappa}, \infty)$. Over the first interval, x^* rises and y^* declines with κ; over the latter, x^* is given by equation (29) from the main text.

(*ii*) *Comparative Statics with Respect to* ω.—We have

$$\text{(A47)}\quad \begin{pmatrix} \frac{\partial y^*}{\partial \omega} \\ \frac{\partial x^*}{\partial \omega} \end{pmatrix} = H^{-1}(x^*, y^*) \begin{pmatrix} -1 \\ -\beta'(x^*)\bar{\mu} \end{pmatrix} = \frac{1}{|H|} \begin{pmatrix} -\mathcal{V}_{xx} - \frac{1}{2}(1+\kappa)\bar{\mu}^2 \beta'(x^*)^2 \\ \frac{1}{2}\bar{\mu}\kappa\beta'(x^*) \end{pmatrix},$$

so that x^* is strictly increasing in ω. Decomposing the wedge ω, it follows that x^* is strictly increasing in the baseline externality, w, and in the Principal's private benefit, b. The comparative statics of y^*, on the other hand, are generally ambiguous.

(*iii*) *Comparative Statics with Respect to* $\bar{\theta}$.—Since $(\lambda, \alpha) = (1/2, 0)$, we have $\partial\omega / \partial\bar{\theta} = b/2$, so:

$$\text{(A48)}\qquad \begin{pmatrix} \frac{\partial y^*}{\partial \bar{\theta}} \\ \frac{\partial x^*}{\partial \bar{\theta}} \end{pmatrix} = H^{-1}(x^*, y^*) \begin{pmatrix} \frac{1}{2}(\kappa - b) \\ -\frac{1}{2}\beta'(x^*)\bar{\mu} b \end{pmatrix}$$

$$= \frac{1}{|H|} \begin{pmatrix} \frac{1}{2}(\kappa - b)\mathcal{V}_{xx} - \frac{1}{4}(1+\kappa)\bar{\mu}^2 \beta'(x^*)^2 b \\ \frac{1}{4}\bar{\mu}\beta'(x^*)\left(\kappa(1+\kappa) + \kappa b\right) \end{pmatrix},$$

so that x^* is increasing in $\bar{\theta}$, as in the baseline model. The effect of $\bar{\theta}$ on y^* is more ambiguous: if $\kappa > b$, then clearly y^* is decreasing in $\bar{\theta}$; for $\kappa \approx 0$ (and therefore $x^* \approx 0$), on the other hand, one can show that y^* is increasing in $\bar{\theta}$ for all $b > 0$ (details available upon request).

(*iv*) *Comparative Statics with Respect to $k_P, s_P^2\, \sigma_\mu^2, \sigma_\theta^2$, and s_θ^2.*—For any parameter η that does not appear in (A41), i.e., that does not directly affect y^*, we have $\mathcal{V}_{y\eta} = 0$, so by (A43),

$$\begin{pmatrix} \frac{\partial y^*}{\partial \eta} \\ \frac{\partial x^*}{\partial \eta} \end{pmatrix} = \frac{\mathcal{V}_{x\eta}}{|H|} \begin{pmatrix} \mathcal{V}_{xy} \\ -\mathcal{V}_{yy} \end{pmatrix}. \tag{A49}$$

Since $\mathcal{V}_{xy}$ and $\mathcal{V}_{yy}$ are both negative, x^* and y^* have opposite comparative statics with respect to η. Such properties hold for $\eta \in \left\{k_P, s_P^2 \sigma_\mu^2\right\}$, as none of these parameters enter $\beta(x)$. Furthermore, $\mathcal{V}_{xk_P}, \mathcal{V}_{xs_P^2}$, and $\mathcal{V}_{x\sigma_\mu^2}$ are all independent of y and so have the same signs as in the benchmark model, where $y \equiv 0$: as shown in Tables 1–2, this means that x^* is increasing in k_P, decreasing in s_P^2, and decreasing in σ_μ^2 outside some interval $[\underline{\sigma}, \bar{\sigma}]$, or everywhere if $k_P \geq \bar{k}_P$; y^*, meanwhile, has the opposite variations.

As in the baseline model, the comparative statics with respect to s_θ^2 and σ_θ^2 are generally ambiguous except (i) when s_θ/σ_θ becomes small enough, so that ρ approaches 1, and (ii) in the private-values specification, where ρ is simply replaced by 1. In those cases, $\mathcal{V}_x$ no longer depends on $\eta \in \left\{s_\theta^2, \sigma_\theta^2\right\}$ and $\mathcal{V}_{x\eta}$ is independent of y, so again Table 2 still implies that x^* is increasing in s_θ^2 and decreasing in σ_θ^2, while y^* has the opposite variations.

Comparative Statics with Respect to s_v^2.—Simplifying (A42) by $\beta' > 0$, note that x enters (A41)–(A42) only through $\beta(x)$. Therefore, as in the benchmark model, we can think of the Principal directly optimizing on β, together with y. Since s_v^2 does not enter (A41)–(A42) other than through β, this means that the optimal y^* and $\beta^* = x^* \xi\left(x^*, s_v^2\right)$ are independent of it; the second property, together with (10), implies as before that x^* is strictly decreasing in s_v^*.

(*v*) *Comparative Statics with Respect to $\bar{v}$.*—At an interior solution for $\left(x^*, y^*\right)$, we can write, using $\partial\omega/\partial\bar{v} = -\lambda(1-\alpha) = -1/2$,

$$\begin{pmatrix} \frac{\partial y^*}{\partial \bar{v}} \\ \frac{\partial x^*}{\partial \bar{v}} \end{pmatrix} = H^{-1}\left(x^*, y^*\right) \begin{pmatrix} \frac{1}{2}(1+\kappa) \\ \frac{1}{2}\bar{\mu}\beta'\left(x^*\right) \end{pmatrix} \tag{A50}$$

$$= \frac{1}{|H|} \begin{pmatrix} \left(\frac{1+\kappa}{2}\right)\mathcal{V}_{xx} + \bar{\mu}^2 \beta'\left(x^*\right)^2 \\ \frac{1}{4}\kappa^2 \bar{\mu}\beta'\left(x^*\right) \end{pmatrix}$$

so that x^* is increasing in $\bar{v}$, which implies that y^* must be decreasing in it. The overall comparative statics of x^* and y^* also reflect the fact that the threshold $\bar{\kappa}$ varies with $\bar{v}$, however. Let us show that $\bar{\kappa}$ is strictly decreasing in $\bar{v}$, up to a point where it reaches 0. To see this, take $\bar{v}_1$, $\bar{v}_2$ with $\bar{v}_1 < \bar{v}_2$ and suppose that $0 < \bar{\kappa}(\bar{v}_1) \leq \bar{\kappa}(\bar{v}_2)$. Since $x^*(\kappa, \bar{v})$ is (i) strictly increasing in κ up to $\bar{\kappa}(\bar{v})$ and then constant and (ii) strictly increasing in v as long as $\kappa \leq \bar{\kappa}(v)$, we have

$$(A51) \quad x^*\big(v_1, \bar{\kappa}(\bar{v}_2)\big) = x^*\big(v_1, \bar{\kappa}(\bar{v}_1)\big) < x^*\big(v_2, \bar{\kappa}(\bar{v}_1)\big) \leq x^*\big(v_2, \bar{\kappa}(\bar{v}_2)\big).$$

For $\kappa \geq \bar{\kappa}(\bar{v}_2)$, however, $y^*(\kappa, \bar{v}_1) = y^*(\kappa, \bar{v}_2) = 0$, and in that range we know from Table 1 that x^* is strictly decreasing in $\bar{v}$, so $x^*\big(v_1, \bar{\kappa}(\bar{v}_2)\big) < x^*\big(v_2, \bar{\kappa}(\bar{v}_2)\big)$ is a contradiction. Hence, $\bar{\kappa}$ must be strictly decreasing in $\bar{v}$ until it has reached 0; this happens for finite $\bar{v}$, as $\tilde{y}$ reaches 0 when $\kappa(\bar{v} + \bar{\theta}) = 2\omega$, implying that y^* must equal 0, hence $\bar{\kappa} = 0$. This concludes the proof that where $y^* > 0$, x^* is strictly increasing in $\bar{v}$, and y^* decreasing in it, until the point where $\bar{\kappa}(\bar{v})$ has declined to 0; afterward, x^* is decreasing in $\bar{v}$. Thus, overall, x^* is inverse U-shaped in $\bar{v}$; since (25) is independent of $\bar{v}$ and decreasing in x^*, finally, m^* is U-shaped in $\bar{v}$. ∎

REFERENCES

Acemoglu, Daron, and Matthew O. Jackson. 2017. "Social Norms and the Enforcement of Laws." *Journal of European Economic Association* 15 (2): 245–95.

Acquisti, Alessandro, Curis Taylor, and Liad Wagman. 2016. "The Economics of Privacy." *Journal of Economic Literature* 54 (2): 442–92.

Al-Najjar, Nabil I. 2004. "Aggregation and the Law of Large Numbers in Large Economies." *Games and Economic Behavior* 47 (1): 1–35.

Algan, Yann, Yochai Benkler, Mayo Fuster Morell, and Jérôme Hergueux. 2013. "Cooperation in a Peer Production Economy: Experimental Evidence from Wikipedia." https://www.parisschoolofeconomics.eu/IMG/pdf/hergueux_paper-2.pdf.

Andreoni, James. 2006. "Leadership Giving in Charitable Fund-Raising." *Journal of Public Economic Theory* 8 (1): 1–22.

Andreoni, James, and B. Douglas Bernheim. 2009. "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica* 77 (5): 1607–36.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 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." *Journal of Public Economics* 120: 1–17.

Auriol, Emmanuelle, and Robert J. Gary-Bobo. 2012. "On the Optimal Number of Representatives." *Public Choice* 153 (3–4): 419–45.

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

Bénabou, Roland. 2013. "Groupthink: Collective Delusions in Organizations and Markets." *Review of Economic Studies* 80 (2): 429–62.

Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3): 489–520.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

Bénabou, Roland, and Jean Tirole. 2011. "Laws and Norms." NBER Working Paper 17579.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy* 102 (5): 841–77.

Besley, Timothy, Anders Jensen, and Torsten Persson. 2014. "Norms, Enforcement, and Tax Evasion." http://perseus.iies.su.se/~tpers/papers/Draft_140302.pdf.

Bolton, Patrick, Markus K. Brunnermeier, and Laura Veldkamp. 2013. "Leadership, Coordination, and Corporate Culture." *Review of Economic Studies* 80 (2): 512–37.

Botsman, Rachel. 2017. "Big Data Meets Big Brother as China Moves to Rate Its Citizens." *Wired.co.uk*, October 21, 2017. https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion.

Brennan, Geoffrey, and Philip Pettit. 1990. "Unveiling the Vote." *British Journal of Political Science* 20 (3): 311–33.

Brennan, Geoffrey, and Philip Pettit. 2004. *The Economy of Esteem: An Essay on Civil and Political Society*. Oxford: Oxford University Press.

Carlsson, Hans, and Eric van Damme. 1993. "Global Games and Equilibrium Selection." *Econometrica* 61 (5): 989–1018.

Cooter, Robert D. 2004. "The Donation Registry." *Fordham Law Review* 72 (5): 1981–89.

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

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

Del Carpio, Lucia. 2013. "Are the Neighbors Cheating? Evidence from a Social Norm Experiment on Property Taxes in Peru." http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=39366C18CF0E603C2BBD590FF847B3F1?doi=10.1.1.642.7115&rep=rep1&type=pdf.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127 (1): 1–56.

Elías, Julio J., Nicola Lacetera, Mario Macis, and Paola Salardi. 2017. "Economic Development and the Regulation of Morally Contentious Activities." *American Economic Review* 107 (5): 76–80.

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

Fehrler, Sebastian, and Niall Hughes. 2018. "How Transparency Kills Information Aggregation: Theory and Experiment." *American Economic Journal: Microeconomics* 10 (1): 181–209.

Fischer, Paul E., and Robert E. Verrecchia. 2000. "Reporting Bias." *Accounting Review* 75 (2): 229–45.

Fox, Justin, and Richard Van Weelden. 2012. "Costly Transparency." *Journal of Public Economics* 96 (1–2): 142–50.

Frankel, Alex, and Navin Kartik. 2019. "Muddled Information." *Journal of Political Economy* 127 (4): 1739–76.

Frey, Bruno S. 2007. "Awards as Compensation." *European Management Review* 4 (1): 6–14.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102 (1): 33–48.

Harbaugh, William T. 1998. "What Do Donations Buy? A Model of Philanthropy Based on Prestige and Warm Glow." *Journal of Public Economics* 67 (2): 269–84.

Harbaugh, William T., Ulrich Mayr, and Daniel R. Burghart. 2007. "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations." *Science* 316 (5831): 1622–25.

Hermalin, Benjamin E., and Michael L. Katz. 2006. "Privacy, Property Rights and Efficiency: The Economics of Privacy as Secrecy." *Quantitative Marketing and Economics* 4 (3): 209–39.

Holmström, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *Review of Economic Studies* 66 (1): 169–82.

Hummel, Patrick, John Morgan, and Phillip C. Stocken. 2013. "A Model of Flops." *RAND Journal of Economics* 44 (4): 585–609.

Jacquet, Jennifer. 2015. *Is Shame Necessary? New Uses for an Old Tool*. New York: Penguin Random House.

Jia, Ruixue, and Torsten Persson. 2017. "Individual vs. Social Motives in Identity Choice: Theory and Evidence from China." http://perseus.iies.su.se/~tpers/papers/ChildDraft170125.pdf.

Judd, Kenneth L. 1985. "The Law of Large Numbers with a Continuum of IID Random Variables." *Journal of Economic Theory* 35 (1): 19–25.

Kahan, Dan M. 1997. "Between Economics and Sociology: The New Path of Deterrence." *Michigan Law Review* 95 (8): 2477–97.

Kahan, Dan M., and Eric A. Posner. 1999. "Shaming White-Collar Criminals: A Proposal for Reform of the Federal Sentencing Guidelines." *Journal of Law and Economics* 42 (S1): 365–92.
Kaplow, Louis. 1992. "The Optimal Probability and Magnitude of Fines for Acts that Definitely Are Undesirable." *International Review of Law and Economics* 12 (1): 3–11.
Kuran, Timur. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.
Lacetera, Nicola, and Mario Macis. 2010. "Social Image Concerns and Prosocial Behavior: Field Evidence from a Nonlinear Incentive Scheme." *Journal of Economic Behavior and Organization* 76 (2): 225–37.
Landier, Augustin, David Sraer, and David Thesmar. 2009. "Optimal Dissent in Organizations." *Review of Economic Studies* 76 (2): 761–94.
Levy, Gilat. 2005. "Careerist Judges and the Appeals Process." *RAND Journal of Economics* 36 (2): 275–97.
Levy, Gilat. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules." *American Economic Review* 97 (1): 150–68.
Linardi, Sera, and Margaret A. McConnell. 2011. "No Excuses for Good Behavior: Volunteering and the Social Environment." *Journal of Public Economics* 95 (5–6): 445–54.
Lohmann, Susanne. 1994. "Information Aggregation through Costly Political Action." *American Economic Review* 84 (3): 518–30.
Loury, Glenn C. 1994. "Self-Censorship in Public Discourse: A Theory of 'Political Correctness' and Related Phenomena." *Rationality and Society* 6 (4): 428–61.
Morgan, John, and Phillip C. Stocken. 2008. "Information Aggregation in Polls." *American Economic Review* 98 (3): 864–96.
Morris, Stephen. 2001. "Political Correctness." *Journal of Political Economy* 109 (2): 231–65.
Morris, Stephen, and Hyun Song Shin. 2002. "Social Value of Public Information." *American Economic Review* 92 (5): 1521–34.
Morris, Stephen, and Hyun Song Shin. 2006. "Global Games: Theory and Applications." In *Advances in Economics and Econometrics*, Vol. 8, edited by Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, 56–114. Cambridge, UK: Cambridge University Press.
Ottaviani, Marco, and Peter Sørensen. 2001. "Information Aggregation in Debate: Who Should Speak First?" *Journal of Public Economics* 81 (3): 393–421.
Polinsky, A. Mitchell, and Steven Shavell. 1979. "The Optimal Trade-off between the Probability and Magnitude of Fines." *American Economic Review* 69 (5): 880–91.
Polinsky, A. Mitchell, and Steven Shavell. 2007. "The Theory of Public Enforcement of Law." In *Handbook of Law and Economics*, Vol. 1, edited by A.M. Polinsky and S. Shavell, 403–54. Amsterdam: Elsevier.
Posner, Eric A. 1998. "Symbols, Signals, and Social Norms in Politics and the Law." *Journal of Legal Studies* 27 (S2): 765–97.
Posner, Eric A. 2002. *Law and Social Norms*. Cambridge, MA: Harvard University Press.
Posner, Richard A. 1978. "The Right of Privacy." *Georgia Law Review* 12 (2): 393–422.
Posner, Richard A. 1979. "Privacy, Secrecy, and Reputation." *Buffalo Law Review* 28 (1): 1–55.
Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95 (3): 862–77.
Prendergast, Canice. 1993. "A Theory of 'Yes Men.'" *American Economic Review* 83 (4): 757–70.
Reeves, Richard V. 2013. "Shame Is Not a Four-Letter Word." *New York Times*, March 15. https://www.nytimes.com/2013/03/16/opinion/a-case-for-shaming-teenage-pregnancy.html.
Ronson, Jon. 2015. "How One Stupid Tweet Blew Up Justine Sacco's Life." *New York Times*, February 12. https://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html.
Segal, David. 2013. "Mugged by a Mug Shot Online." *New York Times*, October 5. https://www.nytimes.com/2013/10/06/business/mugged-by-a-mug-shot-online.html.
Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentives Schemes." *American Economic Review* 97 (3): 999–1012.
Sun, Yeneng. 2006. "The Exact Law of Large Numbers via Fubini Extension and Characterization of Insurable Risks." *Journal of Economic Theory* 126 (1): 31–69.
Supreme Court of the United States. 2015. *Obergefell v. Hodges*, 576 U.S. 644.
Swank, Otto H., and Bauke Visser. 2013. "Is Transparency to No Avail?" *Scandinavian Journal of Economics* 115 (4): 967–94.

Swank, Otto H., and Bauke Visser. 2015. "Learning from Others? Decision Rights, Strategic Communication, and Reputational Concerns." *American Economic Journal: Microeconomics* 7 (4): 109–49.

Uhlig, Harald. 1996. "A Law of Large Numbers for Large Economies." *Economic Theory* 8 (1): 41–50.

van der Weele, Joël. 2012. "The Signaling Power of Sanctions in Social Dilemmas." *Journal of Law, Economics, and Organization* 28 (1): 103–26.

Vesterlund, Lise. 2003. "The Informational Value of Sequential Fundraising." *Journal of Public Economics* 87 (3–4): 627–57.

Visser, Bauke, and Otto H. Swank. 2007. "On Committees of Experts." *Quarterly Journal of Economics* 122 (1): 337–72.

Warren, Samuel D., and Louis D. Brandeis. 1890. "The Right to Privacy." *Harvard Law Review* 4 (5): 193–220.

Whitman, James Q. 1998. "What Is Wrong with Inflicting Shame Sanctions?" *Yale Law Journal* 107 (4): 1055–92.