Image Versus Information:

Changing Societal Norms and Optimal Privacy

S. Nageeb Ali [1] Roland Bénabou [2]

First version: March 2015

This version: January 2018 [3]

[1] Pennsylvania State University and THRED. Email: nageeb@psu.edu.

[2] Princeton University, NBER, CEPR, IZA, CIFAR, BREAD, THRED and BRIQ. Email: rbenabou@princeton.edu.

[3] 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, Stephen Morris, Casey Mulligan, Justin Rao, Joel Sobel, Pierre-Luc Vautrey, as well as to participants in many seminars and conferences. We thank Edoardo Grillo, Tetsuya Hoshino, Charles Lin, Pellumb Reshidi, and Ben Young for 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.

Abstract

We analyze the costs and benefits of using social image to foster virtuous behavior. A Principal seeks to motivate reputation-conscious agents to supply a public good. Each agent chooses how much to contribute based on his own mix of public-spiritedness, private assessment of the public good, and reputational concern for appearing prosocial. By making individual behavior more visible to the community the Principal can amplify reputational payoffs, thereby reducing free-riding at low cost. Because societal preferences constantly evolve, however, she knows only imperfectly both the social value of the public good (which matters for choosing her own investment, matching rate or legal policy) and the importance attached by agents to social esteem and sanctions. Increasing publicity makes it harder for the Principal to learn from what agents do (the "descriptive norm") what they really value (the "prescriptive norm"), thus presenting her with a tradeoff between incentives and information aggregation. We derive the optimal degree of privacy/publicity (as well as the optimal level of monetary incentives when both policy tools can be combined) and how it depends on the economy's stochastic and informational structure. We show in particular that in a fast-changing society (greater variability in the fundamental or the image-motivated component of average preferences), privacy should generally be greater than in a more static one.

Keywords: social norms, privacy, transparency, incentives, esteem, reputation, shaming punishments, conformity, social learning, societal change.

JEL Classification: D62, D64, D82, H41, K42, Z13.

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, 2009).

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," 1977, pp. 401-405).

1 Introduction

1.1 Why Privacy?

Social visibility is a powerful incentive. When people know that others will learn of their actions, they contribute more to public goods and charities, 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?

This question is 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 charities publicize donors' names on buildings and plaques. On the sanctions side, many U.S. states and towns use updated forms of the pillory: televised "perp walks," internet posting of the identities of people convicted or merely arrested for a host of offences (tax or child support delinquency, spousal abuse, drunk driving, etc.); publishing the licence plates of cars photographed in areas of drug trafficking or prostitution; and sentencing offenders to "advertise" their deeds by means of special clothing, lawn 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.[2]

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

[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, Lacetera and Macis (2010).

[2] In Peru, businesses convicted of tax evasion can be shut down, with a sign plastered in front; conversely, municipalities publish an "honor list" of households who have always paid their property taxes on time (Del Carpio (2014)). Shaming can also be organized by activists, as with the "Occupy Wall Street" movement, the hacking of Ashley Madison's list of members, or a number of "naming" lists posted following the #metoo disclosures.

said is rapidly falling to zero –it is in fact maintaining privacy and anonymity that is becoming increasingly expensive.[3] The trends described above are therefore likely to accentuate, whether impulsed by public authorities, activist groups or individual whistleblowers.[4]

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 (2003), Brennan and Pettit (1990, 2004), Frey (2007)) and shame (Kahan (1996), Kahan and Posner (1999), Reeves (2013), Jacquet (2015)), on grounds of both efficiency and expressive justice. R. Posner (1977, 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.[5]

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.[6] 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 "irresponsible" behaviors. 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 (for instance, prohibiting credit bureaus from reporting past arrests), and a major debate over the *"right to be forgotten"* is ongoing with search-engine and social-media companies.

There is of course a case for protecting individuals' information from the eyes of parties with malicious intent: governments repressing dissenters, firms using data on consumer's habits

[3] A flourishing image-ransoming industry is even developing in the United States. These "shame entrepreneurs" operate by re-posting 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)).

[4] As part of its 2014-2017 Five-Year Plan China is developing a "Social Credit System" in which the behavior of every citizen, firm and public entity will be rated, with the resulting score conditioning access to credit, private and public services (housing, travel, welfare) as well as public praise and shame lists. Besides political activities, it will grade everyone's credit history, business and employment practices, social responsibility, legal and "trust" record, tax compliance, traffic violations, etc., in order to build a nationwide "culture of sincerity." The Alibaba Group (one of the companies involved) already runs its own "Sesame Credit" system, which publicly rates people by the products they buy, the activities they engage in, and the friends they keep. Besides earning priority for certain services, "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, because the higher their Sesame rating, the more prominent their dating profile is" (Botsman (2017)).

[5] 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.

[6] In the United States, "Peeple", a paying app in which individuals can grade others in each of the professional, personal and dating categories, went online in 2016, but has so far met with strong backlash.

to engage in exploitation, hackers intent on identity theft, or rivals seeking trade secrets. We focus here on a very different notion of privacy –how much citizens know about each other's behaviors– and on 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 to publicizing behaviors, and doing so leads agents to provide needed public goods, it is desirable to maintain or protect a certain degree of privacy. This remains a fortiori true under less ideal conditions.[7]

1.2 Our 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 which 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,* and how a (benevolent or selfish) Principal should optimally set its level. Third, we take up the intrinsically novel problem of how *monetary and reputational incentives* should jointly be used, and fully solve for the optimal policy mix.

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:

1. *Variability in the power of social image.* The rewards and sanctions generated by publicizing an individual's actions stem from the reactions that this elicits from his family, peers, or neighbors. As these involve the *emotional responses* of many people and their degree of *coordination,* their severity is hard to predict and fine-tune a priori (Whitman (1998), E. Posner (2000)). Depending on place, time, group and individual contingencies, the response can range from mild ostracism to mob action, be easy or hard to escape, etc.[8] Variability in agents' concerns about social image and sanctions will, in turn, generate inefficient variations in compliance (not reflecting true variations in social value), which become amplified as individuals' actions are made more visible or salient.[9]

2. *Rigid or misguided public policy.* Public stigmatization has long been used to repress non-believers, mixed-race relationships, single-mothers, homosexuals, etc. At the time, such behaviors were widely considered immoral and socially nefarious, and accordingly

[7] 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.

[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.

also punished by the law. Societal preferences *evolve*, however (due to technology, education, migration), requiring legislators and courts to learn from prevailing mores and "community standards" how policies and institutions should be adapted. If people are too worried about social stigma, these shifts will remain obscured, resulting in rigidification and maladaptation not just of *private conduct* but also of *public policy*.

Full reversals of societal values are fairly common in modern societies: overt racism, sexism or domestic violence went from "normal" to deeply scorned within a decade or two, while divorce, cohabitation ("living in sin"), homosexuality or drug use shifted from intensely stigmatized to widely acceptable. It is thus inevitable that *some* conducts seen as abhorrent today will become mundane within a couple of decades, and vice-versa.[10]

Even for behaviors that remain unambiguously bad or good from a social point view (drunk driving, tax compliance, etc.), the learning issue remains. 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. While signals such as polls may be available they are very imperfect, especially on issues where social approval comes into play, so the actual behavior of the population remains a potentially valuable indicator.[11]

Because "minimal-privacy" advocates ignore that what constitutes a public good, a heinous deed, a minor or wide-ranging externality (and hence also a proper signal of prosociality) is subject to unpredictable shifts, they mistakenly equate social conformity with the common good, often quite explicitly.[12] We seek to clarify these issues.

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 derive private values from the prevalence of some behavior in society. In either case their collective information, suitably aggregated, is informative about the social value of the conduct in question. Each one chooses how to act based on his own mix of public-spiritedness, information and concern for appearing prosocial. The Principal can amplify or dampen these reputational payoffs, by making individual behaviors more or less visible to the community. While this entails little cost (none, for simplicity), she faces an inference problem: because societal preferences change, she knows only imperfectly 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. If the Principal suppresses image motivations by making contributions anonymous,

[10] Uncertainty lies only in which ones it will be (e.g., organ sales, prostitution, atheism, consuming animals) and which way the cursor will move. For current trends, see Elias et al. (2016). We will also show that the problem with low privacy lies *not* in increased uniformity of individual choices (it may even have the opposite effect, generating inefficient image-seeking variations) but in the reduced informativeness of *aggregate* behavior.

[11] On polls, see also footnote 23. The problem with low privacy lies *not* in increased uniformity of individual choices (we show that it can have the opposite effect) but in the reduced informativeness of *aggregate* behavior.

[12] For instance, R. Posner (1977), pp. 401-405), criticizing an earlier scholar advocating for a right to privacy, writes: "Bloustein is saying merely that if people were forced to conform their private to their public behavior there would be more uniformity in private behavior across people –that is to say, people would be better behaved if they had less privacy. This result he considers objectionable... for reasons he does not attempt to explain."

she can precisely infer societal preferences from agents' aggregate behavior. People will free-ride substantially, however, leaving her with much of the burden of public-good provision. On the other hand, if she leverages social image to spur compliance, she exacerbates her own signal-extraction problem by making aggregate behavior more sensitive to variations in the importance of social payoffs. The Principal thus faces a tradeoff between using *image as an incentive* and gaining better *information* on societal preferences.[13]

1.3 Applications

1.a. Public good provision, charitable donations. 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 local schools, parks, transportation, or development projects in remote countries. 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)). Publicly recognizing and honoring individuals' or NGO's efforts encourages commitment, but also makes it a less precise signal of true social value. Similar examples in which a government learns from the level of (non)compliance about the public's perceived efficiency or legitimacy of some policy or prohibition include alcohol and drug use, electoral turnout, and even tax evasion or "morale."

1.b. Agency incentives. Representatives in a sales team can often privately observe how well the product fits customer needs. Publicizing individual sales records leads them to exert more effort in promoting it, thus alleviating the moral-hazard problem (Larkin (2011)), but it deprives the firm of valuable feedback: seeing high sales, it may not realize that its product needs further development without which success will be short-lived, or that it involves hidden risks.

2.a. 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 (Obergefell v. Hodges, 2015):

> *"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... "*

[13] 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 tradeoffs 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.

> "*Changed understandings of marriage are characteristic of a Nation where new dimensions of freedom become apparent to new generations... Well into the 20th 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.*"

In this second class of applications, changing mores can reveal aggregate shifts in the distribution of sentiments about whether some behavior constitutes a social ill or a social good: women working, being legally and financially independent; divorce, single-parenthood; gays in the army or as teachers, same-sex couples; physical punishments, ethnic jokes and slurs, cannabis consumption, or, most recently, sexual harassment. Conversely, where behavior is highly constrained by the fear of social stigma, assessing social preferences by what people do –the *"descriptive norm"*– will be a poor indicator of what they really value –the *"prescriptive norm"*; laws and other policies will then lag far behind evolving mentalities.[14] Section 2.1 explains how these issues map into the variant of the model (analyzed in Section 6) where agents derive *private values* from the prevalence of some behavior, rather than a *common value* from a public good of imperfectly known quality.

2.b. Consumer and corporate social responsibility. Firms are increasingly pressured or shamed by activists into behaving "responsibly" on issues of environmental impact, child labor, workplace safety, treatment of animals, etc. To the extent that such reputational incentives make up for deficient regulation or Pigovian taxation they are beneficial, but the strong signaling effects they create make it hard for consumers and investors to know which practices are truly socially valuable and which ones are just "greenwashing". The same applies to "green" and "fair trade" consumer goods, typically heavily advertised and often conspicuously consumed.

1.4 Related Literature

Our study relates to several lines of work examining the impact of transparency on individual and collective decision-making. A first strand focuses on signaling, especially in a public-goods context.[15] 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.

[14] Similar issues arise in the debate over freedom of speech versus "political correctness." Activist groups and media outlets commonly use publicity to curtail acts and words considered offensive (Loury (1994), Morris (2001)). In Brazil, a campaign tracked down the geotagged locations of people posting racist comments on social media, then reprinted them on giant billboards and public buses in the source's neighborhood (names and pictures were blurred). More recently, a French newspaper reprinted such posts with the authors' full identities.

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

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 tradeoff 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) show 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 adjusts continuously how much privacy to accord individuals, faces uncertainty about they will respond to it, and cares not about who does what but about the informational content of the collective behavior.[16]

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 with the views expressed by more "senior" agents thought to be a priori more knowledgeable (Ottaviani and Sørensen (2001)). When competence is a private type, on the other hand, the incentive to signal it generates "anti-conformist" or activist tendencies: agents will overreact to their own signals, reverse precedents, etc.; which of the two forces dominates then depends on the details on game's information and strategic structure (Levy (2005, 2007)), Visser and Swank (2007)).[17] In our framework, agents' incentives to signal their types simultaneously have positive (mean-contribution) and negative (excessive variance and information-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.[18] 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 (2015), Acemoglu and Jackson (2017)).

[16] Daughety and Reinganum (2010) also show that waivable privacy rights do not help reduce wasteful signaling. Bénabou and Tirole (2011) show, on the other hand, that if the value of image (e.g., the "going rate" to have one's name on a university or hospital building) is known to the Principal, tax incentives can be adjusted to offset any reputation-motivated distortions in the level or allocation of contributions. This is another reason why the Principal's not knowing the exact value of image is important here.

[17] 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 (2015)).

[18] Bénabou and Tirole (2006) study signaling agents with heterogenous (privately known) image-concerns, and Fischer and Verrecchia (2000) and Frankel and Kartik (2017) 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), nor a Principal who seeks to incentivize agents through publicity (and payments) and learn from their behavior.

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: (1) There is *no strategic interdependence* in payoffs. (2) Individuals act based *solely on their private information* (which is multidimensional), before observing any public variable. The equilibrium norm ($\bar{a}$) is learnt only afterwards, and what matters is not the "macro" visibility of this or any other statistic, but instead the "micro" visibility of personal, *idiosyncratic* choices. (3) 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. (4) What is essential there is that agents *observe* an *actual* common signal, whereas here it is that each one *think,* rightly or wrongly, that they *might* be *observed* by others.

The remainder of the paper is organized as follows. Section 2 presents the basic model. Section 3 derives agents' equilibrium learning and behavior under any fixed level of publicity, establishing a surprisingly simple *benchmarking* result for social inferences in this complex informational environment. Section 4 examines the Principal's resulting tradeoffs and solves for the optimal publicity level and contributions-matching rate; Section 5 then characterizes their full comparative statics. Section 6 combines monetary and reputational incentives, deriving the optimal policy mix and its comparative statics. Section 7 presents several extensions, and Section 8 concludes. Proofs are in Appendix A, extensions in (supplementary) Appendix B.

2 Model

We study the interaction between a continuum of small agents ($i \in [0,1]$) and a single large Principal (P), each of whom chooses how much to contribute (in time, effort or money) to a public good. Depending on the context, these actors may correspond to: (i) a government and its citizens; (ii) a charitable organization and potential donors; (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.

A. 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 (non-reputational) payoff is

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

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 $N\left(\bar{v}, s_v^2\right)$ 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.

The quality or social value of the public good is *a priori* uncertain, with agents and the Principal starting with common prior belief that θ is distributed as $N\left(\bar{\theta}, \sigma_\theta^2\right)$. Each agent i receives a private noisy signal, $\theta_i \equiv \theta + \varepsilon_i$, in which the error is distributed as $N\left(0, s_\theta^2\right)$, 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 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 importance of a good reputation varies across individuals, communities and periods; it is greater, for instance, for people engaged in long-run relationships based on trust than where exchange occurs through impersonal markets and complete contracts. Social enforcement –punishing or shunning perceived free-riders, rewarding model citizens– also relies on mobilizing emotional reactions and achieving group coordination, both of which are hard to predict. 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); for tractability, we take the three components to be independent of each other.[21] An individual j observing i's contribution 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

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

We assume that a social image $R\left(a, \theta_i, \mu_i\right)$ yields for agent i a (normalized) payoff of $\mu_i x[R\left(a, \theta_i, \mu_i\right) - \bar{v}]$, where μ_i reflects his baseline concern for social esteem and $x \geq 0$ parame-

[19] Alternatively, each could have his own genuine valuation θ_i for it (private-values case): see Section 7.1.

[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] In particular, if μ_i was correlated with v_i or θ_i the inference problems of agents and Principal would no longer have a linear-normal structure.

trizes 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

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

B. Principal. The Principal's final payoff is a convex combination of agents' total utility and her own private benefits and costs from the overall supply of the (quality-adjusted) public good:

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

The first line captures agents' standard costs and benefits from public-goods provision. In the second line, $\alpha \in [0,1]$ captures the extent to which Principal internalizes their intrinsic "joy of giving," relative to these material payoffs, and $\tilde{\alpha}$ that to which she internalizes their image gains and losses. 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; (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 non-trivial we shall assume that $\omega > 0$, so that, on average, the Principal does want to increase private contributions (or norm compliance). To cut down on the number of cases we shall focus the exposition on the case where $b > 0$, which in turn implies that $\varphi > 0$ and $\partial\omega/\partial\bar{\theta} = \lambda\alpha + (1-\lambda)b > 0$, meaning that "higher quality" is indeed something that the Principal values positively. Her preferences over the quality of the public good are thus congruent with those of the agents, even though her preferences over the level and sharing of its supply may be quite different.[22]

Our framework includes as special cases:

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

(b) 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.

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

In order to set her own provision a_P efficiently, the Principal must learn about θ. A key piece of data 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 $\theta \sim N\left(\bar{\theta}, \sigma_\theta^2\right)$ about the quality of the public good and may also obtain an independent signal $\theta_P \equiv \theta + \varepsilon_P$, with error distributed as $N(0, s_{\theta,P}^2)$. 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]

C. Timing. The game unfolds as follows:

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

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

3. The aggregate contribution $\bar{a}$ is publicly observed.

4. The Principal observes her own signal θ_P.

5. 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).

2.1 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:

[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 who are weary of revealing "discreditable" preferences –as recent electoral outcomes in the UK and US have clearly shown. Allowing the Principal to obtain an independent, noisy signal of μ would also not affect our analysis.

1. Agent's contributions become driven in larger part by variations in their reputational concerns, rather than by their signals about the social value of the public good.

2. 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, hence too difficult for her to learn from.

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

Private vs. 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 value θ. This corresponds to a setting with *common values*, such as charitable 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 each agent's ex-post, full-information payoff depends on his own perspective or taste θ_i concerning the public good or other externalities generated by the behavior in question.[24]

This flexibility is important, as the "changing social mores" applications discussed earlier typically correspond to behaviors from which different people experience or perceive very different externalities. These externalities may be material (recreational drug use), socioeconomic (single parenthood, working mothers), psychological (sexual harassment, hostile work atmosphere) or even purely *moral* ("offensive to my values, disgusting," "demeaning to human dignity").[25] In such settings, θ_i reflects agent i's subjective (dis)utility from the externality (whether incurred directly or through the internalized welfare of others, e.g., children), s_θ^2 measures the *dispersion of opinions* (societal disagreement) and the policy-setting Principal cares about $\theta = E[\theta_i]$ as reflecting the *average preference* over these spillovers.[26] Section 7.1 shows that all results and formulas are either identical to, or special cases of, the corresponding ones in the common-values 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.

[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.

[25] Other commonly perceived externalities involve a mix of these different concerns: "undermining the institution of the family," "sacrilegious," "corrupting society," "sowing hatred and division," etc.

[26] More generally, aggregate shifts in the distribution of private values. 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; thus, in (1), θ_i affects both the term in a_i and that in $(\bar{a}+a_P)$. 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 vs. attitudes toward those of others, enjoyment of polluting activities vs. vulnerability to airborne particulates or climate change, etc. Typically they will be positively correlated, due to: (a) the common prosocial-concern component; (b) 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 non-zero correlation, at the cost of additional notation.

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 tradeoff 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 or/and 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 non-compliers 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. A different approach would be to focus more on one specific channel. We do so in Section 7.4, showing how similar results emerge when the Principal controls only the precision which each agent's action is observed by others.

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.[27] 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 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.[28]

To focus squarely on the effects of publicity, we initially abstract from the use of standard material incentives to induce compliance.[29] In Sections 6-7, however, we allow the Principal to combine visibility with either costly *monetary incentives* or *legal regulations* (quantity mandates), and this either ex-ante (before observing equilibrium compliance) or ex-post (one she has observed and learned from it). All key results and comparative statics remain unchanged.

[27] Papers studying an informed-principal problem in related contexts include Bénabou and Tirole (2003), Sliwka (2008), Van der Weele (2013) and Bénabou and Tirole (2011), who study how *laws shape norms,* whereas we focus here on the complementary mechanism of how *norms shape laws.*

[28] 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.

[29] 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 (A.27) of the Appendix.

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

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*

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

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

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

The resulting aggregate contribution (or compliance level) is

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

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 information sets –namely, different signals (θ_i, μ_i) that are predictive of the average θ

and μ, hence also of the θ_j's and μ_j's which observers will have at their disposal to extract v_i from a_i, using (9).[30] 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 again share the *same beliefs* about his motivation: $E[v_i|a, \bar{a}, \theta_j, \mu_j] = E[v_i|a, \bar{a}]$.[31] 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:

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

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

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

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

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

The overjustification effect. 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:*

(1) 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.

(2) The impact of visibility on contributions, $\beta(x) \equiv x\xi(x)$, is strictly increasing in x, with

[30] 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.

[31] 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.

$\lim_{x\to\infty} \beta(x) = +\infty$, *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* $\mu > 0$.[32]

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 in intrinsic motivation much more than in information about θ, which intensifies the signaling game and thereby raises contributions.[33] As $\xi(x) \to 1$ the equilibrium becomes fully revealing, with each agent's social image exactly matching his actual preference: $E[v_i \mid a_i] = v_i$. 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.

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

Exogenous variations in privacy. Inspection of (11) already makes apparent the key trade-offs: a higher x increases aggregate contributions $\bar{a}(x)$ but also causes them to vary inefficiently, and most importantly 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 tradeoff into account, a case we now turn to formally.

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

[33] This somewhat subtle comparative static 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 7.1, 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.

4 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 non-zero, they are trivially small compared to direct spending on public goods, subsidies or law enforcement.[34]

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

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

(b) A case where $\sigma_\mu^2 > 0$ but the Principal, like the agents, observes the realization of μ once x has been set, but prior to choosing a_P.

(c) 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: (a) how the Principal would set publicity if she could fine-tune its impact, $x\xi\mu$, perfectly; (b) the "variance effect" that emerges when she cannot do so but observes μ *ex post*; (c) 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.[35] Agents' social-learning problems, meanwhile, become simpler, with the image return $\xi(x)$ reducing to the constant ξ given by (14). In Section 4.4 we allow for $s_\mu^2 > 0$ and show that all key results remain unchanged.

4.1 Fine-Tuned Publicity: An Image-Based Pigovian Policy 4.1

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

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

[34] This cost advantage is one of the main arguments put forward by proponents of publicity and shame (e.g., Kahan (1996), Brennan and Pettit (2004), Jacquet (2015)). As mentioned earlier, given technological evolutions it may soon be *reducing* x from its laissez-faire level (protecting privacy) that necessitates costly investments.

[35] A further simplification is that agents' reputational gains and losses sum to zero (by Bayes' rule, $\int_0^1 \mu R(a_i, \theta_i, \mu_i) di = \mu\bar{v}$), so the corresponding term (and concern) vanishes from the Principal's objective function (4), as $\tilde{\alpha} = 0$. 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_i's are also private information.

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 $\bar{a}$ she uses in choosing x are thus simply her priors $\bar{\theta}, \bar{\mu}$ and $\tilde{a}(x) = \bar{v} + \bar{\theta} + x\xi\bar{\mu}$. Substituting into the objective function (4) and differentiating leads to an optimal level

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

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

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 which the Principal fine-tunes to exactly offset free-riding, i.e. the gap between the public good's social value w and agents' average willingness to contribute voluntarily, $\bar{v}$. 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.

4.2 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 of 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 (A.7), 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 tradeoff 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*

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

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 (2000), so the Principal naturally wields it more cautiously than under the Pigovian policy: $x^{SI} < x^{FB}$, for all $\lambda > 0$.

4.3 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 also receives a private signal θ_P, allowing her to update her prior beliefs to an *interim* estimate $\bar{\theta}_P$ with mean square error $\sigma_{\theta,P}^2 \equiv E\left[(\theta - \bar{\theta}_P)^2\right]$, where

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

$$\sigma_{\theta,P}^2 = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_{\theta,P}^2}\right)^2 s_{\theta,P}^2 + \left(\frac{s_{\theta,P}^2}{\sigma_\theta^2 + s_{\theta,P}^2}\right)^2 \sigma_\theta^2. \tag{22}$$

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

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

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.[36] After observing $\bar{a}$, the Principal optimally sets $a_P = \varphi\left(w + E\left[\theta|\bar{a}, a_P\right]\right)/(1-\lambda)k_P$; substituting in (20) and (23) yields:

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

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

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

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

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:

$$EV(x) = E\tilde{V}(x) - \frac{\varphi^2\sigma^2_{\theta,P}}{2(1-\lambda)k_P}\left[1-\gamma(x)\right]. \tag{27}$$

The Principal's first-order condition is now

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

The first term, previously explicited 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, and even below the symmetric-information benchmark of Section 4.2.

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*

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

[36] 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.

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.[37] 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 5 for the more general model where agents may differ in how they value reputation.

4.4 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_i's wash out in the aggregate contribution $\bar{a}(x)$, implying:

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* $EV(x) - \tilde{E}V(x)$ *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* $x\xi(x) = \beta(x)$*.*

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

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

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 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

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

which is ensured in particular if either (i) $\tilde{\alpha} \leq 1/2$, or (ii) $s_\mu^2 < \bar{\mu}^2 + \sigma_\mu^2 = E[\mu]^2$, 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.

[37] By (24), it equals $x/(1+Ax^2)^2$, 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.

Proposition 6. *When the Principal is uncertain about the importance of social image, the optimal degree of publicity x^* solves the implicit equation*

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

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

Depending on whether the Principal discounts the value of image by more or less than 1/2, x^* will now be below or above the value characterized in Proposition (5). As before, (32) could have multiple solutions but all stable ones, including the global optimum, share the same comparative-statics properties, to which we now turn.

5 Comparative Statics of the Optimal Policy Bundle

Let us now examine how the Principal's choice of *publicity* x^* and *matching rate* $m^* = \gamma(x^*)/[\rho k_P(1-\lambda)]$ depend on key features of the environment.[38]

A. 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 I below. These properties are quite intuitive. For instance, a principal who faces a higher costs of own funds, or who 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.

		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	$\overline{v}$	Decreasing	Increasing
Principal's relative cost	k_P	Increasing	Decreasing

Table I: Comparative-Static Effects of First-Moment Parameters

We next turn to the dependence of the optimal policies on *second-moment* parameters of cross-sectional heterogeneity and aggregate variability.

B. 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

[38] In analyzing how x varies with some primitive parameter η we establish that the objective function is supermodular in (x, η). By Topkis's Theorem, the set of global maximizers is then increasing in x and η. If the global maximizer is unique, optimal publicity is then increasing in η the usual sense; if there are multiple global maximizers, it is increasing in the strong set order.

that she weighs agents' warm glow positively she appreciates variability, but on the other hand 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 tradeoff: 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 $\lambda(\alpha - 1/2)$.*

C. 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 disregard their signal and *follow the common prior*. To neutralize the second effect and highlight the Principal's tradeoff 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 $\sigma_{\theta,P}^2$; see (10), (21) and (24). Therefore:

Proposition 8. *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 optimal visibility x^* decreases with ex-ante uncertainty σ_θ^2 over θ, while the optimal matching rate γ^* increases with it.*

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, as we show in Section 7.1, these comparative statics always emerge unambiguously in the case of private values, regardless of the value of $s_\theta^2/\sigma_\theta^2$.

D. Variability in the importance of social image or social enforcement.

1. 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

[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 $(-1/2)\int_0^1 a_i^2 d_i$.

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 –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 on the Principal's payoffs and decisions:*

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

(2) 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 $\sigma_\mu > \bar{\sigma}$.

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

2. 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.*

E. Precision of private signals

1. Principal's signal. When the noise $s_{\theta,P}^2$ affecting her independent information increases, the Principal is naturally worse off. To see how she responds, note from (28) that $s_{\theta,P}^2$ appears only in the information-distortion effect, through $\gamma(x)$; thus, from (24), we have $\partial^2 EV(x)/\partial x \partial s_{\theta,P}^2 < 0$. This is again 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 $s_{\theta,P}$ 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 information, $s_{\theta,P}^2$, whereas her optimal matching rate m^* increases with it.*

2. 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 non-monotonic 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 he Principal has "enough" to learn from agents, meaning that their private signals are sufficiently informative: as $s_\theta/\sigma_\theta \to 0$, ρ 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. Therefore, we have:

Proposition 12. *When agents' private signals about the quality of the public good are far more precise than their prior over it ($s_\theta^2/\sigma_\theta^2$ sufficiently small), the Principal's payoff is decreasing in the variance of their signals, s_θ^2. Her optimal choice of publicity is increasing in s_θ^2, and her optimal matching rate is independent of it.*

In the closely parallel setting with *private values*, the above comparative-statics once again hold regardless of the value of $s_\theta^2/\sigma_\theta^2$; see Section 7.1 below. Table II summarizes the results from the preceding propositions.

	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 $k_P \geq \bar{k}_P$
$s_{\theta,P}^2$	Decreasing	Increasing
s_θ^2	Increasing for s_θ/σ_θ small, or private values	Invariant

Table II: Comparative-Statics Effects of Second-Moment Parameters

6 Combining Reputational and Material Incentives

In most real-world settings Principals have access to both monetary and image incentives to induce desirable behaviors by agents. This makes the question how they should optimally be combined a 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 first examine how the social equilibrium among agents changes when they face both incentives together, then solve for the Principal's optimal policy mix. Section 7.3 will analyze the sequential case, in which the Principal initially controls only image incentives (or "informal institutions"), 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 I-II) remains essentially unchanged, and derive new ones. The details are relegated to the Appendices, whereas we highlight here the key insights and results.

In period 1, the Principal now simultaneously chooses the level of visibility x among agents and an incentive rate $y \geq 0$ paid to each of them per unit of contribution, 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 matching rate $m(x)$.

Denote $a_i(x)$ the equilibrium individual strategies of 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's contributions $\xi(x)$ unchanged.[40] The same is therefore true of the aggregate $\bar{a}(x)$, so that 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 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 contributions, incentives raise their marginal costs, and thereby affect 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 $\lambda = 1/2$. There exists $\bar{\kappa} \in (0, +\infty)$ such that:*

1. *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*

$$y^* = \tilde{y} - \tau\bar{\mu}\beta(x^*), \tag{33}$$

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

2. *For any κ, the comparative-statics of x^* and m^* with respect to all parameters in Tables I-II remain unchanged, but 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.*

3. *For all $\kappa \leq \bar{\kappa}$, the comparative statics of y^* with respect to $k_P, s^2_{\theta,P}$ $\bar{v}$, and σ^2_μ, σ^2_θ and s^2_θ are the exact reverse of those of x^*. It is independent of s^2_v, while comparative statics with respect to $w, \bar{\theta}$, α and $\bar{v}$ are generally ambiguous.*

[40] 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.

The first set of results is quite intuitive: when monetary incentives are costless, it is it 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, until the point where monetary incentives have 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^*;[41] we also get new predictions about the behavior of y^*.

7 Extensions and Variants

7.1 Private Values

Our analysis has focused on the case of *common values,* in which each agent ultimately cares about some objective quality of the public good, θ, and uses her information to assess it; accordingly, he also values the information held about θ by others. With *private values*, in contrast, tastes or sentiments towards the public good are dispersed and heterogeneous –with s_θ^2 now measuring the extent of disagreement– and the Principal cares about θ as the average preference. Each agent now knows precisely the reward he derives from contributing to (and consuming) the public good, but remains uncertain of how his contributions will be interpreted by his peers, since he does not know their private values; as to the Principal, she remains unsure of the aggregate social value.[42] Such a setting may be a better fit for privacy issues concerning *social norms* and *political views,* and how the latter should shape formal institutions: these are typically instances where agents ultimately "agree to disagree" about what is socially valuable, or simply benefit differently from some public good or externality.

This variant of the problem turns out to essentially reduce to a special case of the common-values benchmark. The key insight (see Appendix B for details) is that, because each agent is now sure of how she values the public good (θ_i), she contributes as if she were getting a perfect signal about a common value, that is, as if $\rho = 1$.[43] The entire analysis is thus the same as in Sections 2-6, but replacing ρ by 1 in all formulas in the text. An important further implication is that all of the prior comparative-statics results for (x^*, m^*, y^*) carry over, with those relative to s_θ^2 and σ_θ^2 now holding *without any restriction* on the ratio s_θ/σ_θ.

[41] 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, the planner substitutes toward the use of publicity, which does not have such a cost.

[42] In this setting, θ_i is specific to the good at hand (the environment, human rights, etc.), or equivalently varies with situational factors affecting the agent's cost of contributing to that particular cause. By contrast, v_i is a general degree of prosocial orientation or other-regard that carries across contexts and periods. Consequently, reputations are still formed over v_i, rather than $v_i + \theta_i$.

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

7.2 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 and other formal institutions (legislators, Supreme Court, etc.) aim to prescribe behaviors deemed appropriate in light of current "values" and mores. A related motive is that laws that deviate too much from current values are likely to generate significant distortions (black markets, dissimulation), or even become unenforceable.

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 2; 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 Appendix B). We also derive the optimal mandate for any x, whether exogenous or optimally chosen ($x = x^*$):

$$a^* = \frac{\varphi}{\lambda}\left(w + \mathbb{E}\left[\theta|\bar{a}, \theta_P\right]\right) = \frac{\varphi}{\lambda}\left(w + [1-\gamma(x)]\bar{\theta}_P + \gamma(x)\hat{\theta}\right), \tag{34}$$

with $\hat{\theta}$, $\bar{\theta}_P$ and $\gamma(x)$ still defined as in Section 4.3. This makes clear (as for the matching rate in (25)) that *the law responds to the (descriptive) norm* $\bar{a}$ –but the less so, the less privately each individual chose his action.

7.3 Norms Informing Incentives

Consider now the case where the key policy that the Principal wants to "get right" in light of possible shifts in agents' 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 non-trivial resource costs.

To deal with 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; (ii) then, 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 of $(1+\kappa)y$.[44] The paper's core insights apply again, linking in particular the optimal degree of publicity to the Principal's need for information about

[44] One could also combine ex-ante incentives (chosen prior to observing compliance), as in Section 6, with the ex-post incentives studied here.

θ, agents' individual and collective knowledge about it, and the strength of their reputational concerns. Correspondingly, the form of equilibrium solutions and all the comparative statics of x^* remain unchanged. As to the optimal second-period incentive rate, it is

$$y^* = \widehat{y} + \frac{(1+b) - \rho(1+\kappa)}{1+2\kappa} E[\theta|\theta_P, \bar{a}], \tag{35}$$

where $\hat{y}$ is a constant and $E[\theta|\theta_P, \bar{a}] = [1-\gamma(x^*)]\,\bar{\theta}_P + \gamma(x^*)\hat{\theta}$ 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 $\hat{y}$).

7.4 Noisy Observability

An alternative specification of publicity is one in which each agent's action is observed with some noise, which the Principal can affect. Suppose that, when i contributes a_i, others observe $\hat{a}_i = a_i + \varepsilon_i$ and $\varepsilon_i \sim N(0, s_\varepsilon^2/x^2)$, where x may be chosen by the Principal. The return to image (or observers' signal-to-noise ratio) $\xi(x)$ now becomes

$$\xi(x) = \frac{s_v^2}{\xi(x)^2 s_\mu^2 + s_\varepsilon^2/x^2 + s_v^2 + \rho^2 s_\theta^2}, \tag{36}$$

which remains very similar to the earlier expression, (10). Consequently, we show in the Supplementary Appendix that all key implications remain unchanged.

7.5 Simple Dynamics

Our model has highlighted 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 interpreted the results in dynamic terms – a fast or slow-changing society, formal institutions adapting to those changes or remaining rigid, etc. For completeness, we extend here the results to a simple dynamic environment.

Each generation of agents lives for one period, subdivided into two subperiods during which they interact among themselves and with a Principal, just as before: they contribute, signal, consume public goods, etc. At the start of each period $t = 0, 1, 2, \ldots$, Nature chooses the aggregate shocks (θ_t, μ_t) affecting that generation's preferences. The initial θ_0 and μ_0 are drawn from a normal distribution, after which θ_t and μ_t follow AR-1 processes: $\theta_t = \varrho_\theta \theta_{t-1} + \varepsilon_t^\theta$, with $\varepsilon_t^\theta \sim N(0, \sigma_\theta^2)$ and $\varrho_\theta \leq 1$, and $\mu_t = \varrho_\mu \mu_{t-1} + \varepsilon_t^\mu$,with $\varepsilon_{t-1}^\mu \sim N(0, \sigma_\mu^2)$ and $\varrho_\mu \leq 1$. At the beginning of each period $t \geq 1$, all agents and the Principal observe the previous generation's average values $(\theta_{\tau-1}, \mu_{\tau-1})$, and on that basis the game proceeds as before. While agents are short-lived, the Principal may be long-lived, discounting payoffs at rate δ.

Conditional on the current priors $\varrho_\theta \theta_{t-1}$ and $\varrho_\mu \mu_{t-1}$: (i) each agent faces the same problem as in the static analysis; (ii) the optimal policy for the Principal is to set publicity x and choose

a_P using the same decision rules in each period: because θ_t and μ_t will be revealed prior to next period's decisions, her choices today have no impact on her future payoffs.

8 Conclusion

We studied the interaction between social norms and social learning, and the resulting tradeoff between the incentive benefit of publicizing individual behaviors that constitute public-goods (or bads) and the costs which reduced privacy imposes on society (or any other Principal) when the overall distribution of preferences is subject to unpredictable shifts and evolutions.

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, to appropriately adapt policies and institutions. We showed in particular that where societal attitudes (what behaviors agents regard as socially desirable or undesirable) and/or technologies for monitoring and norms enforcement (means of communication and coordination, e.g., social media) are prone to significant change, a higher degree of privacy is optimal: policy-makers can then better learn, by observing overall compliance, how taxes and subsidies, the law or other institutions should be adapted. When preferences over public goods and reputation remain or have become relatively stable, conversely, visibility should be raised. We also derived the optimal mix of monetary and image incentives, and showed that all the 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 (Kreps (1990), 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.[45]

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

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

9 Appendix A: Main Proofs

Proofs of Proposition 1 on p. 14 and Proposition 2 on p. 15

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* $E[v_i|\theta_j, \mu_j, \bar{a}, a_i]$ *is independent of* (θ_j, μ_j) *and equal to:*

$$E[v_i|\theta_j, \mu_j, \bar{a}, a_i] = \bar{v} + \frac{Bs_v^2}{B^2 s_v^2 + C^2 s_\theta^2 + A^2 s_\mu^2}(a_i - \bar{a}). \tag{A.1}$$

Proof. Subtracting $\bar{a}$ from a_i, and re-arranging, we obtain $Bv_i = B\bar{v} + (a_i - \bar{a}) - (C\varepsilon_i^\theta + A\varepsilon_i^\mu)$, where ε_i^θ and let ε_i^μ denote $\theta_i - \theta$ and $\mu_i - \mu$ respectively. Observe that $(Bv_i, a_i - \bar{a}, \bar{a}, \theta_j, \mu_j, C\varepsilon_i^\theta + A\varepsilon_i^\mu)$ 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 $C\varepsilon_i^\theta + A\varepsilon_i^\mu$ and $a_i - \bar{a}$, and these variables are jointly normally distributed, it follows from independence that

$$E[C\varepsilon_i^\theta + A\varepsilon_i^\mu|a_i, \bar{a}, \theta_j, \mu_j] = E[C\varepsilon_i^\theta + A\varepsilon_i^\mu|a_i - \bar{a}].$$

Observe that

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

and therefore, $E[v_i|\theta_j, \mu_j, \bar{a} - a_i]$ equals the expression in (A.1). ■

From Claim 1 it follows that

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

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

$$a_i = v_i + \rho\theta_i + (1-\rho)\bar{\theta} + x\mu_i\left(\frac{Bs_v^2}{A^2 s_\mu^2 + B^2 s_v^2 + C^2 s_\theta^2}\right). \tag{A.3}$$

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

$$\xi(x) = \frac{s_v^2}{x^2\xi(x)^2 s_\mu^2 + s_v^2 + \rho^2 s_\theta^2}. \tag{A.4}$$

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

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

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 $\rho s_\theta = \sigma_\theta^2/[s_\theta + \sigma_\theta^2/s_\theta]$).

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

$$x = \beta[\beta^2(s_\mu^2/s_v^2) + \rho^2 s_\theta^2/s_v^2 + 1], \tag{A.5}$$

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

Proof of Proposition 3 on p. 18

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 an 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

$$\begin{aligned} E\tilde{V}(x) = \lambda &\left[\alpha\left(s_v^2 + \rho\sigma_\theta^2 + (\bar{v}+\bar{\theta})(\bar{\bar{a}})\right) + (w+\bar{\theta})(\bar{\bar{a}} + \bar{a}_P) + \rho\sigma_\theta^2 + \frac{\sigma_\theta^2\varphi}{(1-\lambda)k_P}\right] \\ &+ (1-\lambda)b\left[(w+\bar{\theta})(\bar{\bar{a}} + \bar{a}_P) + \rho\sigma_\theta^2 + \frac{\sigma_\theta^2\varphi}{(1-\lambda)k_P}\right] \qquad \text{(A.6)} \\ &- \frac{\lambda}{2}[\bar{\bar{a}}^2 + \rho^2(\sigma_\theta^2 + s_\theta^2) + s_v^2 + x^2\xi^2\sigma_\mu^2] - \frac{(1-\lambda)k_P}{2}\left[\bar{a}_P^2 + \sigma_\theta^2\left(\frac{\varphi}{(1-\lambda)k_P}\right)^2\right]. \qquad \text{(A.7)} \end{aligned}$$

Differentiating yields:

$$\begin{aligned} \frac{dE\tilde{V}(x)}{dx} &= \{\lambda\left[\alpha(\bar{v}+\bar{\theta}) + (w+\bar{\theta})\right] + (1-\lambda)b(w+\bar{\theta})\}\xi\bar{\mu} - \lambda\left[\xi\bar{\mu}\left(x\xi\bar{\mu} + \bar{v} + \bar{\theta}\right) + x\xi^2\sigma_\mu^2\right] \\ &= \omega\xi\bar{\mu} - \lambda x\xi^2(\bar{\mu}^2 + \sigma_\mu^2). \end{aligned} \tag{A.8}$$

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). ■

Proof of Proposition 4 on p. 20

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

$$\underline{a}_P(x,\theta_P) = \frac{\varphi(w + (1-\gamma(x))E[\theta|\theta_P] - \frac{\gamma(x)}{\rho}(\bar{v} + x\xi\bar{\mu} + (1-\rho)\bar{\theta}))}{(1-\lambda)k_P}, \tag{A.9}$$

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

Proof of Proposition 5 on p. 20

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 $a_P = (w+E[\theta|\bar{a},\theta_P])\varphi)/(1-\lambda)k_P$, which makes clear how the forecast error $\Delta \equiv E[\theta|\bar{a},\theta_P] - \theta$, derived in (26), distorts her contribution from the full-information optimal by $\frac{\varphi\Delta}{(1-\lambda)k_P}$. 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 $\frac{\varphi^2}{2(1-\lambda)k_P}E[\Delta^2]$, where

$$E\left[\Delta^2\right] = (1-\gamma)^2\sigma_{\theta,P}^2 + (\gamma\xi x/\rho)^2\sigma_\mu^2 = \sigma_{\theta,P}^2\left[(1-\gamma)^2 + \gamma^2(1/\gamma - 1)\right] = \sigma_{\theta,P}^2(1-\gamma), \quad \text{(A.10)}$$

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

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

$$\begin{aligned}\frac{\sigma_{\theta,P}^2}{2}\frac{d\gamma}{dx} &= -\frac{\sigma_{\theta,P}^2}{2}\left(\frac{2\rho^2\sigma_{\theta,P}^2\xi^2\sigma_\mu^2}{\left(\rho^2\sigma_{\theta,P}^2 + x^2\xi^2\sigma_\mu^2\right)^2}x\right) = -\frac{\sigma_{\theta,P}^2\gamma(1-\gamma)}{x} \\ &= -\sigma_{\theta,P}^2\left(\frac{\gamma^2\xi^2\sigma_\mu^2}{\rho^2\sigma_{\theta,P}^2}x\right) = -\frac{\sigma_\mu^2\gamma^2\xi^2 x}{\rho^2}.\end{aligned}$$

Therefore

$$\begin{aligned}\frac{\partial EV}{\partial x} &= \frac{\partial E\tilde{V}}{\partial x} - \frac{\varphi^2}{(1-\lambda)k_P}\left(\frac{\sigma_\mu^2\gamma^2\xi^2 x}{\rho^2}\right) \\ &= (\xi\bar{\mu})\left[(w+\bar{\theta})\varphi - (\bar{v}+\bar{\theta})(1-\alpha)\lambda\right] - \lambda x\xi^2\left(\bar{\mu}^2+\sigma_\mu^2\right) - \frac{\varphi^2}{(1-\lambda)k_P}\left(\frac{\sigma_\mu^2\gamma^2\xi^2 x}{\rho^2}\right),\end{aligned} \quad \text{(A.11)}$$

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 $x^* \in (0,\tilde{x})$ where $\partial EV/\partial x = 0$. ■

Proof of Proposition 6 on p. 22

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*

$$\tilde{x}^{SI} = \frac{\bar{\mu}\omega}{\lambda\xi\left(x^{SI}\right)\left(\bar{\mu}^2+\sigma_\mu^2+(1-2\tilde{\alpha})s_\mu^2\right)}. \quad \text{(A.12)}$$

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 (homogeneousμ) 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 (A.7) giving $EV(x)$, and similarly $dEV^0(x)/dx$ by replacing $x\xi$ with $\beta(x)$ and ξ with $\beta'(x)$ in the expression (A.8) for $dEV(x)/dx$:

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

In the Principal's actual loss function (4), however, the heterogeneity in agents' μ_i's generates an additional loss due to inefficient cost variations, equal to $(\lambda/2)E[(a_i)^2 - (a_i^0)^2] = (\lambda/2)\beta(x)^2 s_\mu^2$, but also generates a gain from their image seeking that corresponds 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

$$\frac{dE\tilde{V}}{dx} = \omega\bar{\mu}\beta'(x) - \lambda\beta'(x)\beta(x)\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right) = 0, \tag{A.13}$$

$$\Longrightarrow \beta\left(\tilde{x}^{SI}\right) = \frac{\bar{\mu}\omega}{\lambda\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right)}, \tag{A.14}$$

which is equivalent to (A.12). ■

Notice that $\tilde{x}^{SI} < \infty$ so long as $\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2 > 0$. That condition is automatically satisfied either if (i) $\tilde{\alpha} < \frac{1}{2}$, 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 the mean image concern μ when setting her contribution. Corollary 1 allows us to simply combine (A.13) and (27) to obtain the relevant version of her first-order condition:

$$\frac{dEV}{dx} = \bar{\mu}\beta'(x)\omega - \lambda\beta'(x)\beta(x)\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right) + \frac{\varphi^2\sigma_{\theta,P}^2}{2(1-\lambda)k_P}\gamma'(x) = 0.$$

Using (24) we have $\gamma'(x) = -[2\sigma_\mu^2/\rho^2\sigma_{\theta,P}^2]\beta(x)\beta'(x)\gamma(x)^2$, leading to:

$$\beta(x^*) = \frac{\bar{\mu}\omega}{\lambda(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2) + \frac{1}{(1-\lambda)k_P}\left(\frac{\varphi\sigma_\mu\gamma(x)}{\rho}\right)^2}, \tag{A.15}$$

which is equivalent to (32). ■

Proof of Proposition 7 on p. 23

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

$$\mathcal{V}(z) = s_v^2\left(\lambda\alpha - \frac{\lambda}{2}\right) + z\bar{\mu}\omega - \frac{\lambda}{2}z^2(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2) - \frac{\varphi^2\sigma_{\theta,P}^2}{2(1-\lambda)k_P}\left[1-\tilde{\gamma}(z)\right] + C, \quad \text{(A.16)}$$

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

$$\bar{\mu}\omega - \lambda z(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2) + \frac{\varphi^2\sigma_{\theta,P}^2}{2(1-\lambda)k_P}\tilde{\gamma}'(z) = 0. \quad \text{(A.17)}$$

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 $x^*(s_v)\xi(x^*(s_v), s_v)$, (A.4) implies that a higher s_v strictly increase $\xi(x^*(s_v), s_v)$. Therefore, $x^*(s_v)\xi(x^*(s_v), s_v)$ remains unchanged only if $x^*(s_v)$ is decreasing in s_v. Finally, (A.16) implies that $d[EV(x^*(s_v^2); s_v^2)]/ds_v^2 = \lambda\left(\alpha - 1/2\right)$. ■

Proof of Proposition 8 on p. 23

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 $\gamma(x^*;\sigma_\theta)$ did not increase with σ_θ the right-hand-side of (30) could not remain equal to zero. Thus, at the optimal x^*, $\gamma(x^*;\sigma_\theta)$ must increase with σ_θ. ■

Proof of Proposition 9 on p. 24

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

$$\frac{\partial^2 EV}{\partial x\partial\sigma_\mu^2} = -\lambda\beta'(x)\beta(x) - \frac{\varphi^2\beta'(x)\beta(x)}{\rho^2(1-\lambda)k_P}\left(\gamma^2 + 2\gamma\sigma_\mu^2\frac{d\gamma}{d\sigma_\mu^2}\right), \quad \text{(A.18)}$$

in which

$$\frac{\partial\gamma}{\partial\sigma_\mu^2} = -\frac{\rho^2\sigma_\theta^2\beta(x)^2}{\left(\rho^2\sigma_\theta^2 + \beta(x)^2\sigma_\mu^2\right)^2} = -\frac{\gamma\beta(x)^2}{\rho^2\sigma_\theta^2 + \beta(x)^2\sigma_\mu^2} = -\frac{\gamma\left(1-\gamma\right)}{\sigma_\mu^2}. \quad \text{(A.19)}$$

Thus,

$$\frac{\partial^2 EV}{\partial x\partial\sigma_\mu^2} = -\beta'(x)\beta(x)\left[\lambda + \frac{\varphi^2\gamma^2}{\rho^2(1-\lambda)k_P}\left(2\gamma - 1\right)\right]. \quad \text{(A.20)}$$

This expression is non-positive if and only if

$$\frac{\lambda(1-\lambda)\rho^2 k_P}{\varphi^2} \geq \gamma^2(1-2\gamma). \quad \text{(A.21)}$$

Because $\gamma^2\left(1-2\gamma\right)$ 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:

(a) As σ_μ^2 tends to 0, $\gamma(x^*(\sigma_\mu^2);\sigma_\mu^2)$ approaches 1, because by Proposition 5, $x^*(\sigma_\mu^2)$ remains bounded above: $x^*(\sigma_\mu^2) < \bar{x}$. Therefore, (A.21) holds for σ_μ small enough.

(b) As σ_μ^2 tends to ∞, $x^*(\sigma_\mu^2)$ must tend to 0 fast enough that the product $\sigma_\mu^2 x^*(\sigma_\mu^2)$ 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 $\sigma_\mu^2\left[x^*(\sigma_\mu^2)\right]^2$ tends to 0, and therefore $\gamma(x^*(\sigma_\mu^2);\sigma_\mu^2)$ tends to 1. Thus, for σ_μ^2 large enough (A.21) holds, and $x^*(\sigma_\mu^2)$ decreases to 0. ■

Proof of Proposition 10 on p. 24

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 $EV(x^*(s_\mu^2), s_\mu^2)$ cancel out at the first order, leaving only the direct effect $(-\lambda/2)\beta(x)^2(1-2\tilde{\alpha})$, which is less than 0 if and only $\tilde{\alpha} < \frac{1}{2}$. ■

Proof of Proposition 12 on p. 25

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

$$\xi = \frac{s_v^2}{x^2\xi^2 s_\mu^2 + s_v^2 + s_\theta^2}, \tag{A.22}$$

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 $d\left[EV(x^*(s_\theta);s_\theta)\right]/ds_\theta = -\lambda s_\theta < 0$. Finally, since the Principal keeps $x^*(s_\theta)\xi(x^*(s_\theta), s_\theta)$ constant as s_θ increases, it follows from (A.22) that $\xi(x^*(s_\theta), s_\theta)$ must decrease in s_θ. To compensate, $x^*(s_\theta)$ must then be increasing in s_θ. ■

Preliminary results for Section 6

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 there is whether or not agents derive intrinsic satisfaction from the part of their contribution which 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 $(1+\kappa)(\bar{a}(x)+y)y$, are thus the following:

$$\tilde{V}(x,y,a_P) = V(x,a_P) + (w+\theta)\varphi y - \frac{1}{4}\int \left[(a_i(x)+y)^2 - a_i(x)^2\right] di - \frac{1}{2}\kappa y[\bar{a}(x)+y], \quad \text{(A.23)}$$

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

In the initial period, the Principal optimizes $E[\tilde{V}(x,y,a_P)]$ over (x,y) conditional on her priors, knowing that she will later choose a_P optimally given agents' behavior and what 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 $E\left[2\left(w+\theta\right)\varphi - y - \bar{a}(x) - 2\kappa y - \kappa\bar{a}(x)\right] = 0$ for an interior optimum, or:

$$y = \frac{2\left(w+\bar{\theta}\right)\varphi - (1+\kappa)\bar{a}(x)}{1+2\kappa} \equiv \tilde{y} - \tau\beta(x), \quad \text{where} \quad \text{(A.24)}$$

$$\tilde{y} \equiv \frac{2\omega - \kappa(\bar{v}+\bar{\theta})}{1+2\kappa}; \quad \tau \equiv \bar{\mu}\frac{1+\kappa}{1+2\kappa}. \quad \text{(A.25)}$$

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

a. Incentive and Variance Effects. Consider first the benchmark of Sections 4.1-4.2, 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 (A.23) is

$$\frac{\partial EV(x,a_P)}{\partial x} - \frac{1}{2}(1+\kappa)\bar{\mu}y\beta'(x) = \xi\bar{\mu}\omega - \frac{1}{2}x\xi^2(\bar{\mu}^2+\sigma_\mu^2) - \frac{1}{2}(1+\kappa)\bar{\mu}y^*\xi \leq 0, \quad \text{(A.26)}$$

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 (A.8), this yields

$$x^* = \frac{\bar{\mu}\left[\omega - (1/2)(1+\kappa)y^*\right]}{(1/2)\xi(\bar{\mu}^2+\sigma_\mu^2)} \equiv \frac{\bar{\mu}\tilde{\omega}}{(1/2)\xi(\bar{\mu}^2+\sigma_\mu^2)} \quad \text{(A.27)}$$

when this is nonnegative, otherwise $x^* = 0$. This last case, however, requires that $y^* = \tilde{y}$ by (A.24) and $y \geq 2\omega/(1+\kappa)$ by (A.26) so it only occurs for $\kappa = 0$. Comparing to (19) in the main text, we see that he 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

$$y^* = \frac{2\omega\left(\sigma_\mu^2 - \kappa\bar{\mu}^2\right) - \kappa(\bar{v}+\bar{\theta})(\bar{\mu}^2+\sigma_\mu^2)}{(1+2\kappa)\sigma_\mu^2 - \kappa^2\bar{\mu}^2}, \quad \text{(A.28)}$$

an explicit solution that is non-negative as long as

$$\kappa \leq \bar{\kappa} \equiv \frac{2\omega\sigma_\mu^2}{2\bar{\mu}^2\omega + (\bar{\mu}^2 + \sigma_\mu^2)(\bar{v} + \bar{\theta})}. \tag{A.29}$$

Indeed, the denominator of (A.28) 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 fortiori for all $\kappa < \bar{\kappa} = (\sigma_\mu^2/\bar{\mu}^2)[1 + (\bar{\mu}^2 + \sigma_\mu^2)(\bar{v} + \bar{\theta})/2\omega\bar{\mu}^2]^{-1} < \sigma_\mu^2/\bar{\mu}^2$.

The unique solution to the joint maximization over (x, y) is therefore: (i) for $\kappa \leq \bar{\kappa}$, y^* given by (A.28) and x^* given by (A.27); (ii) for $\kappa \geq \bar{\kappa}$, $y^* = 0$ and $x^* = (2\bar{\mu}\omega/\xi)/(\bar{\mu}^2 + \sigma_\mu^2)$, as in the original model. Substituting y^* into the effective wedge, $\tilde{\omega} = \omega - (1/2)(1 + \kappa)y^*$ and the latter into (A.27), yields an explicit formula for x^* :

$$x^* = \begin{cases} \frac{\kappa\bar{\mu}(\omega + (1/2)(1+\kappa)(\bar{v}+\bar{\theta}))}{(1/2)\xi((1+2\kappa)\sigma_\mu^2 - \kappa^2\bar{\mu}^2)} & \text{for } \kappa < \bar{\kappa} \\ \frac{\bar{\mu}\omega}{(1/2)\xi(\bar{\mu}^2+\sigma_\mu^2)} & \text{for } \kappa \geq \bar{\kappa} \end{cases}. \tag{A.30}$$

The comparative statics of x^*(with associated m^*) and y^* could be obtained directly from (A.30)-(A.28), but will be proven below for the more general model. The only one specific to the present case (i.e., not in Tables I-II) 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 $\bar{\mu}$.

b. Information Distortion and Heterogeneous Image Concerns. For the more general problem with information distortion (and heterogeneous image concern), the first-order condition (A.26) takes the form (as long as $y = \tilde{y} - \tau\beta(x) > 0$):

$$\omega\bar{\mu} - \frac{1}{2}(1 + \kappa)\bar{\mu}\left[\tilde{y} - \tau\beta(x)\right] - \frac{1}{2}\beta(x)\left(\bar{\mu}^2 + \sigma_\mu^2 + (1 - 2\tilde{\alpha})s_\mu^2\right) - \frac{2\varphi^2\sigma_\mu^2}{\rho^2 k_P}\beta(x)\gamma(x)^2 = 0. \tag{A.31}$$

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

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

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 on p. 26.

Denote the Principal's objective function $E\tilde{V}$ (with $\tilde{V}$ given by (A.23)) as $\mathcal{V}(x, y, \Theta)$, where Θ is the vector of all the model's parameters; $\beta(x)$ 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

$$\mathcal{V}_y(x, y, \Theta) = \omega - \frac{1}{2}\kappa(\bar{v} + \bar{\theta}) - \frac{1}{2}\bar{\mu}(1 + \kappa)\beta(x) - \frac{1}{2}(1 + 2\kappa)y \leq 0, \tag{A.33}$$

$$\mathcal{V}_x(x, y, \Theta) = \beta'(x)\left[\omega\bar{\mu} - \frac{1}{2}(1 + \kappa)\bar{\mu}y - \frac{1}{2}\beta(x)[\bar{\mu}^2 + \sigma_\mu^2 + (1 - 2\tilde{\alpha})s_\mu^2] - \frac{2\varphi^2\sigma_\mu^2}{\rho^2 k_P}T(x)\right] \leq 0, \tag{A.34}$$

together with the complementary-slackness conditions $y\mathcal{V}_y(x, y, \Theta) = x\mathcal{V}_x(x, y, \Theta) = 0$.

Note first that, except at $\kappa = 0$, it cannot be that $y^* > 0 = x^*$, otherwise (A.34) yields $y^* = [2\omega - \kappa(\bar{v} + \bar{\theta})]/(1 + 2\kappa) = \tilde{y}$ and (A.33) 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 I-II 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:

$$\begin{pmatrix} \frac{\partial y^*}{\partial \eta} \\ \frac{\partial x^*}{\partial \eta} \end{pmatrix} = H^{-1}(x^*, y^*)\begin{pmatrix} -\mathcal{V}_{y\eta} \\ -\mathcal{V}_{x\eta} \end{pmatrix} = \frac{1}{|H|}\begin{pmatrix} \mathcal{V}_{xx} & -\mathcal{V}_{xy} \\ -\mathcal{V}_{xy} & \mathcal{V}_{yy} \end{pmatrix}\begin{pmatrix} -\mathcal{V}_{y\eta} \\ -\mathcal{V}_{x\eta} \end{pmatrix}, \tag{A.35}$$

where the Hessian matrix of the system (A.33)-(A.34),

$$H = \begin{pmatrix} -\frac{1}{2}(1 + 2\kappa) & -\frac{1}{2}(1 + \kappa)\bar{\mu}\beta'(x) \\ -\frac{1}{2}(1 + \kappa)\bar{\mu}\beta'(x) & -\beta'(x)\left[\frac{1}{2}\beta'(x)(\bar{\mu}^2 + \sigma_\mu^2 + (1 - 2\tilde{\alpha})s_\mu^2) + \frac{2\varphi^2\sigma_\mu^2}{\rho^2 k_P}T'(x)\right] \end{pmatrix}, \tag{A.36}$$

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

Comparative statics with respect to κ. From (A.35)-(A.36), we have

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

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

$$\begin{aligned} |H|\frac{\partial x^*}{\partial \kappa} &= \frac{1}{2}(1 + \kappa)\bar{\mu}\beta'(x)\left(y^* + \frac{1}{2}\bar{\mu}\beta(x^*) + \frac{1}{2}(\bar{v} + \bar{\theta})\right) - \frac{1}{4}(1 + 2\kappa)\bar{\mu}y^*\beta'(x^*) \\ &= \frac{1}{4}\beta'(x^*)\left\{\bar{\mu}y^* + \bar{\mu}(1 + \kappa)\left[\bar{\mu}\beta(x^*) + \bar{v} + \bar{\theta}\right]\right\} > 0. \end{aligned}$$

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 $\bar{\kappa} \in (0,\ 2\omega/(\bar{v} + \bar{\theta}))$ 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.

Comparative statics with respect to ω.

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

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.

Comparative statics with respect to $\bar{\theta}$.

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

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).

Comparative statics with respect to $k_P, s^2_{\theta,P}$ σ^2_μ, σ^2_θ and s^2_θ. For any parameter η that does not appear in (A.33), i.e. that does not directly affect y^*, we have $\mathcal{V}_{y\eta} = 0$, so by (A.35):

$$\begin{pmatrix} \frac{\partial y^*}{\partial \eta} \\ \frac{\partial x^*}{\partial \eta} \end{pmatrix} = \frac{\mathcal{V}_{x\eta}}{|H|} \begin{pmatrix} \mathcal{V}_{xy} \\ -\mathcal{V}_{yy} \end{pmatrix}. \tag{A.37}$$

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 $\eta \in \{k_P, s^2_{\theta,P}\ \sigma^2_\mu\}$, as none of these parameters enters $\beta(x)$. Furthermore, $\mathcal{V}_{xk_P}, \mathcal{V}_{xs^2_{\theta,P}}$ and $\mathcal{V}_{x\sigma^2_\mu}$ are all independent of y, and so have the same signs as in the benchmark model, where $y \equiv 0$: as shown in Tables I-II, this means that x^* is increasing in k_P, decreasing in $s^2_{\theta,P}$ and decreasing in σ^2_μ 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^2_θ and σ^2_θ are generally ambiguous except: (i) when s_θ/σ_θ becomes small enough, so that ρ approaches one; (ii) in the private-values specification, where ρ is simply replaced by 1. In those cases $\mathcal{V}_x$ no longer depends on $\eta \in \{s^2_\theta,\ \sigma^2_\theta\}$ and $\mathcal{V}_{x\eta}$ is independent of y, so again Table II still implies that x^* is increasing in s^2_θ and decreasing in σ^2_θ, while y^* has the opposite variations.

Comparative statics with respect to s^2_v. Simplifying (A.34) by $\beta' > 0$, note that x enters (A.33)-(A.34) 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^2_v does not enter (A.33)-(A.34) other

than through β, this means that the optimal y^* and $\beta^* = x^*\xi(x^*, s_v^2)$ are independent of it; the second property, together with (10), implies as before that x^* is strictly decreasing in s_v^*.

Comparative statics with respect to $\bar{v}$. At an interior solution for (x^*, y^*), we can write

$$\begin{pmatrix} \frac{\partial y^*}{\partial \bar{v}} \\ \frac{\partial x^*}{\partial \bar{v}} \end{pmatrix} = H^{-1}(x^*, y^*) \begin{pmatrix} \frac{1}{2}(1+\kappa) \\ \frac{1}{2}\bar{\mu}\beta'(x^*) \end{pmatrix} = \begin{pmatrix} -\frac{1}{1+2\kappa}\left[(1+\kappa) + \bar{\mu}(1+\kappa)\beta'(x^*)\frac{\partial x^*}{\partial \bar{v}}\right] \\ \frac{1}{4|H|}\kappa^2\bar{\mu}\beta'(x^*) \end{pmatrix},$$

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 zero. 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; (ii) strictly increasing in v as long as $\kappa \leq \bar{\kappa}(v)$, we have:

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

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 I that x^* is strictly decreasing in $\bar{v}$, so $x^*(v_1, \bar{\kappa}(\bar{v}_2)) < x^*(v_2, \bar{\kappa}(\bar{v}_2))$ 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 zero; afterwards, 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 $\bar{v}$. ■

10 Appendix B (for Online Publication Only): Extensions

10.1 Analysis of Private Values in Section 7.1

In the private values environment, each agent's direct (non-reputational) payoff is

$$U_i^{PV}(v_i, \theta_i, w; a_i, \bar{a}, a_P) \equiv (v_i + \theta_i)\, a_i + (w + \theta_i)\,(\bar{a} + a_P) - C\,(a_i)\,. \tag{B.1}$$

The contrast between (1) and (B.1) is that payoffs in the former are determined by θ, which an agent estimates from her signal θ_i, whereas that in the private values setting are determined by θ_i. The reputational payoffs remain unchanged from before.

The Principal cares about θ as the average sentiment towards the public good, but also agent's individual utilities. Her final payoff is

$$V^P \equiv \lambda\left[\alpha\int_0^1 (v_i+\theta_i)\,a_i\,di + \tilde{\alpha}\int_0^1 x\mu_i[R(a_i,\theta_i,\mu_i) - \bar{v}]di + (w+\theta)(\bar{a}+a_P) - \int_0^1 C(a_i)di\right]$$
$$+ (1-\lambda)\left[b(w+\theta)(\bar{a}+a_P) - k_P C(a_P)\right]. \tag{B.2}$$

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

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

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

$$a_i = v_i + \theta_i + x\mu_i \left(\frac{Bs_v^2}{A^2 s_\mu^2 + B^2 s_v^2 + C^2 s_\theta^2}\right). \tag{B.3}$$

Therefore, $B = C = 1$, $D = 0$, and $A = xs_v^2 / \left(A^2 s_\mu^2 + s_v^2 + s_\theta^2\right)$. 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 $\xi(x^2\xi^2 s_\mu^2 + s_v^2 + s_\theta^2) = s_v^2$; the right-hand side is continuous and decreasing in ξ, clearly cutting the diagonal at a unique solution $\xi(x)$. Q.E.D.

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 $\tilde{\xi}(x)$ 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_i's. 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 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.

10.2 Analysis of Norms Shaping Laws in Section 7.2

Agent i's non-reputational payoffs in period 1 and 2 are:

$$U_i^1(v_i, \theta, w, a_i) = (v_i + \theta)a_i + (w + \theta)\bar{a} - \frac{a_i^2}{2}, \tag{B.4}$$

$$U_i^2(v_i, \theta, w, a^*) = (w + \theta)\bar{a} - \frac{(a^*)^2}{2}, \tag{B.5}$$

and he solves: $\max_{a_i} \mathbb{E}\left[U_i^1 + x\,\mu_i\,(R(a_i,\theta_i,\mu_i) - \bar{v}) + \delta\,U_i^2\right]$. Agents thus derive no intrinsic satisfaction from compulsory contributions; the analysis would remain the same if they did, however. The same steps as in Proposition 1 lead again to $a_i = v_i + \rho\,\theta_i + (1-\rho)\bar{\theta} + x\xi(x)\mu$, with $\xi(x)$ unchanged from (19). Turning now to the Principal, her objective function is $E[V^1 + \delta V^2]$, where $\tilde{R} \equiv \int_0^1 \mathbb{E}\left[v_i|a,\bar{a}\right]\,dj$ and

$$\begin{aligned}
V^1 &= \lambda\left(\alpha\int_0^1 (v_i+\theta)a_i\,d_i + (w+\theta)\bar{a} + \tilde{\alpha}\int_0^1 x\,\mu_i\left(\tilde{R}(a_i,\bar{a}) - \bar{v}\right)di - \int_0^1 \frac{a_i^2}{2}\,di\right)\\
&\quad + (1-\lambda)b(w+\theta)\bar{a},\\
V^2 &= \lambda\left((w+\theta)a^* + \tilde{\alpha}\int_0^1 x\,\mu_i\left(\tilde{R}(a_i,\bar{a}) - \bar{v}\right)di - \int_0^1 \frac{(a^*)^2}{2}\,di\right) + (1-\lambda)b(w+\theta)a^*.
\end{aligned}$$

Maximizing $\mathbb{E}\left[V^2|\bar{a},\theta_P\right]$ over a^* leads to

$$\begin{aligned}
0 &= \lambda\left((w + \mathbb{E}\left[\theta|\bar{a},\theta_P\right]) - a^*\right) + (1-\lambda)b(w + \mathbb{E}\left[\theta|\bar{a},\theta_P\right]),\ \text{or}\\
a^* &= \frac{w\varphi}{\lambda} + \frac{\varphi}{\lambda}\mathbb{E}\left[\theta|\bar{a},\theta_P\right].
\end{aligned} \tag{B.6}$$

If, after choosing x, will learn the realized value of θ or μ (allowing her to invert $\bar{a}$ and learn θ perfectly), this reduces to $a^* = \left[w\varphi + \varphi\,\theta\right]/\lambda$ and substituting into the objective function yields

$$\begin{aligned}
E\tilde{V}(x) =& \lambda\left[\alpha\left((\bar{v}+\bar{\theta})\bar{\bar{a}} + s_v^2 + \rho\sigma_\theta^2 + \delta\frac{\varphi}{\lambda}\sigma_\theta^2\right) + \left((w+\bar{\theta})(\bar{\bar{a}} + \delta\bar{a}^*) + \rho\sigma_\theta^2 + \delta\frac{\varphi}{\lambda}\sigma_\theta^2\right)\right.\\
&+ (1+\delta)\tilde{\alpha}\frac{x^2\xi(x)^2}{(1+\delta)}s_\mu^2 - \frac{1}{2}[\bar{\bar{a}}^2 + s_v^2 + \rho^2(\sigma_\theta^2 + s_\theta^2) + x^2\xi(x)^2(\sigma_\mu^2 + s_\mu^2)]\\
&\left.-\frac{\delta}{2}\left((\bar{a}^*)^2 + \left(\frac{\varphi+\lambda\alpha}{\lambda}\right)^2\sigma_\theta^2\right)\right] + (1-\lambda)\left[b(w+\bar{\theta})(\bar{\bar{a}} + \delta\bar{a}^*) + \rho\sigma_\theta^2 + \delta\frac{\varphi}{\lambda}\sigma_\theta^2\right],
\end{aligned} \tag{B.7}$$

where: $\bar{a}^* \equiv \left(w\varphi + \varphi\bar{\theta}\right)/\lambda$. The first order condition is

$$\begin{aligned}
0 =& \lambda\left[\alpha(\bar{v}+\bar{\theta})\bar{\mu}\beta'(x) + (w+\bar{\theta})\bar{\mu}\beta'(x) + 2\tilde{\alpha}s_\mu^2 x\xi(x)\beta'(x) - (\bar{v}+\bar{\theta} + x\xi(x)\bar{\mu})\bar{\mu}\beta'(x)\right.\\
&\left.- x\xi(x)(\sigma_\mu^2 + s_\mu^2)\beta'(x)\right] + (1-\lambda)\left[b(w+\bar{\theta})\bar{\mu}\beta'(x)\right],
\end{aligned}$$

which leads to:

$$x^* = \frac{\bar{\mu}\omega}{\xi(x^*)\,\lambda\,\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right)}. \tag{B.8}$$

When the Principal does not observe θ (or μ), finally, Proposition 14 shows that the expectation in (B.6) remains unchanged: $\mathbb{E}\left[\theta|\theta_P,\bar{a}\right] = [1-\gamma(x)]\bar{\theta}_P + \gamma(x)\hat{\theta}$, with $\bar{\theta}_P$ still given by (21) and $\gamma(x)$ by (24). Hence, similarly to (27):

$$EV(x) = E\tilde{V}(x) - \frac{\delta}{2}\frac{\varphi^2}{\lambda}\sigma_{\theta,P}^2[1-\gamma(x)].$$

Noting, as in the Proof of Proposition 6, that $\gamma'(x) = -(2\sigma_\mu^2/\rho^2\sigma_{\theta,P}^2)\beta(x)\beta'(x)\gamma(x)^2$, and sub-

stituting into the first-order condition for $EV(x)$ yields

$$x^* = \frac{\omega\,\bar{\mu}}{\xi(x^*)\left(\lambda\left(\bar{\mu}^2+\sigma_\mu^2+(1-2\tilde{\alpha})s_\mu^2\right)+\frac{\delta}{\lambda}\left(\frac{\varphi\sigma_\mu\,\gamma(x^*)}{\rho}\right)^2\right)}. \tag{B.9}$$

Given the similarity with the benchmark expressions, the same comparative statics follow.

10.3 Analysis of Norms Shaping Incentives in Section 7.3

The Principals' second-period policy is now to set an incentive rate y', under which agents contribute a second time, rather than constraining them to a legal mandate a^*. For simplicity we assume here that there is no reputational payoff in the second period (no period 3 in which agents would play some continuation game were reputation was valuable). As to the Principal, she again has intertemporal objective function $V^1+\delta V^2$, with components now given by:

$$\begin{aligned} V^1 &= \lambda\left(\alpha\int_0^1 (v_i+\theta)a_i\,d_i+(w+\theta)\bar{a}+\tilde{\alpha}\int_0^1 x\,\mu_i\left(\tilde{R}(a_i,\bar{a})-\bar{v}\right)di-\int_0^1\frac{a_i^2}{2}\,di\right)\\ &\quad+(1-\lambda)b(w+\theta)\bar{a}, \qquad\qquad (B.10)\\ V^2 &= \lambda\left(\alpha\int_0^1 (v_i+\theta)(a_i'-y')\,d_i+(w+\theta+y')\bar{a}'-\int_0^1\frac{(a_i')^2}{2}\,di\right)\\ &\quad+(1-\lambda)\left[b(w+\theta)-(1+\kappa)y'\right]\bar{a}', \qquad\qquad (B.11)\end{aligned}$$

where "primes" denote second-period actions and, as in the case of first-period incentives: (i) the Principal faces a shadow cost $(1+\kappa)$ per unit of funds; (ii) agents derive intrinsic satisfaction only from the portion of their contributions a_i' that is not directly driven by the incentive y'.

Using the notation $a_i(x)$ to denote equilibrium contributions in the baseline model, given by (9), it is clear, given our assumptions, that:

(a) In the first period, agents contribute again the very same $a_i(x)$, for every realization of their (v_i,θ_i,μ_i). Thus both the informativeness $\xi(x)$ of actions about individual types and the informativeness $\gamma(x)$ of aggregate compliance $\bar{a}(x)$ about θ remain unchanged.

(b) In the second period, since agents no longer have any reputational concerns (equivalently, $x'\equiv 0$) but now face material incentives y, each of them contributes $a_i'(y)\equiv a_i(0)+y'$.

Let us again focus (for simplicity only) on the case where $\lambda=1/2$. The problem of the Principal in period 2 is to choose y' to maximize $E[V^2|\theta_P,\bar{a}]$:

$$\max_{y'} E\left[\alpha\int_0^1 (v_i+\theta)a_i(0)di+(w+\theta)(1+b)(\bar{a}(0)+y')-\int_0^1\frac{(a_i(0)+y')^2}{2}di-\kappa y(\bar{a}(0)+y')|\theta_P,\bar{a}\right]$$

The first order condition yields the optimal level of incentives

$$y' = \frac{w(1+b) - (1+\kappa)(\bar{v} + (1-\rho)\bar{\theta})}{1+2\kappa} + \frac{(1+b) - \rho(1+\kappa)}{1+2\kappa} E[\theta|\theta_P, \bar{a}]. \tag{B.12}$$

Quite intuitively, it is increasing in her posterior $E[\theta|\theta_P, \bar{a}]$, but with a slope that declines with the shadow cost of funds κ.

Consider now period 1. As observed above, since reputation is based only on actions and that period, $a_i(x), \xi(x)$ and $\gamma(x)$ all remain unchanged from the benchmark model, so there only remains to solve for the optimal x. As usual, consider first the case in which θ (or μ) is observed by the principal at the beginning of period 2. Then, (B.12) becomes:

$$y' = \frac{w(1+b) - (1+\kappa)[\bar{v} + (1-\rho)\bar{\theta}]}{1+2\kappa} + \frac{[(1+b) - \rho(1+\kappa)]}{1+2\kappa}\theta. \tag{B.13}$$

The Principal's objective function in period 2 is thus independent of x, implying that the optimal x maximizes $E[V^1]$ and is therefore given (A.12), in which we set $\lambda = 1/2$:

$$\tilde{x} = \frac{2\bar{\mu}\omega}{\xi(\tilde{x})[\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2]}. \tag{B.14}$$

Suppose, finally, that the Principal does not observe either θ or μ, and thus uses $\bar{a}$ and θ_P to update her prior. The optimal incentive rate in period 2 is given by (B.12), in which $E[\theta|\theta_P] = \bar{\theta}_P$, $\gamma(x)$, $E[\theta|\theta_P, \bar{a}] = (1-\gamma(x))\bar{\theta}_P + \gamma(x)\hat{\theta}$ and $V(\Delta) = \sigma^2_{\theta,P}(1-\gamma(x))$ all remain unchanged from the baseline model. Note that, as a result, y' rises with the observed $\bar{a}$, but with a slope that decreases in κ. Consequently, with $\lambda = 1/2$ we have

$$EV(x) = \tilde{EV}(x) - \frac{\delta}{4}\left[\frac{(1+b) - \rho(1+\kappa)}{1+2\kappa}\right]^2 \sigma^2_{\theta,P}(1-\gamma(x)), \tag{B.15}$$

which leads to

$$\frac{\partial EV(x)}{\partial x} = \frac{\partial \tilde{EV}(x)}{\partial x} - \frac{\delta}{2}\left(\frac{[(1+b) - \rho(1+\kappa)]\sigma_\mu \gamma(x)}{\rho(1+2\kappa)}\right)^2 x\xi(x)\beta'(x)$$

and the equation defining the optimal x^*

$$x^* = \frac{2\omega\bar{\mu}}{\xi(x^*)\left[\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2 + \delta\left(\frac{[(1+b)-\rho(1+\kappa)]\sigma_\mu\gamma(x)}{\rho(1+2\kappa)}\right)^2\right]}. \blacksquare \tag{B.16}$$

References

Acemoglu, Daron and Matthew O. Jackson. 2017. "Social Norms and the Enforcement of Laws." *Journal of European Economic Association* 15 (2):245–295.

Acquisti, Alessandro, Curis Taylor, and Liad Wagman. 2016. "The Economics of Privacy." *Journal of Economic Literature* 54 (2):442–492.

Algan, Yann, Yochai Benkler, Mayo Fuster Morell, and Jérôme Hergueux. 2013. "Cooperation in a Peer Production Economy: Experimental Evidence from Wikipedia." In *Workshop on Information Systems and Economics*. 1–31.

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):1607–1636.

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):544–555.

Ashraf, Nava, Oriana Bandiera, and 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):419–445.

Bar-Isaac, Heski. 2012. "Transparency, Career Concerns, and Incentives for Acquiring Expertise." *BE Journal of Theoretical Economics* 12 (1):1–15.

Bénabou, Roland. 2012. "Groupthink: Collective Delusions in Organizations and Markets." *Review of Economic Studies* 80 (2):429–462.

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

———. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5):1652–1678.

———. 2011. "Laws and Norms." NBER Working Paper 17579.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy* 102 (5):841–877.

Besley, Tomothy, Anders Jensen, and Torsten Persson. 2015. "Norms, Enforcement, and Tax Evasion." IIES Working Paper.

Bolton, Patrick, Markus K Brunnermeier, and Laura Veldkamp. 2013. "Leadership, Coordination, and Corporate Culture." *Review of Economic Studies* 80 (2):512–537.

Botsman, Rachel. 2017. "Big data meets Big Brother as China moves to rate its citizens." *Wired* .

Brennan, Geoffrey and Philip Pettit. 1990. "Unveiling the Vote." *British Journal of Political Science* 20 (3):311–333.

———. 2004. *The Economy of Esteem.* Oxford University Press New York.

Carlsson, Hans and Eric Van Damme. 1993. "Global games and equilibrium selection." *Econometrica* :989–1018.

Cooter, Robert D. 2003. "The Donation Registry." *Fordham Law Review* 72 (5):1981–1989.

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. 2014. "Are The Neighbors Cheating? Evidence from a Social Norm Experiment on Property Taxes in Peru." INSEAD.

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

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. 2015. "How Transparency Kills Information Aggregation." IZA Discussion Paper No. 9027.

Fischer, Paul E. and Robert E. Verrecchia. 2000. "Reporting Bias." *Accounting Review* 75 (2):229–245.

Fox, Justin and Richard Van Weelden. 2012. "Costly Transparency." *Journal of Public Economics* 96 (1):142–150.

Frankel, Alex and Navin Kartik. 2017. "Muddled Information." *Journal of Political Economy* (Forthcoming).

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):269–284.

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):1622–1625.

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):209–239.

Holmström, B. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *Review of Economic Studies* :169–182.

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.* Pantheon Books, Random House.

Jia, Ruexie and Torsten Persson. 2017. "Individual vs. Social Motives in Identity Choice: Theory and Evidence from China." Working Paper.

Kahan, Dan M. 1996. "Between Economics and Sociology: The New Path of Deterrence." *Michigan Law Review* 95 (5):2477–2497.

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 (1):365–392.

Kreps, David M. 1990. "Corporate Culture and Economic Theory." In *Perspectives on Positive Political Economy*, edited by James E. Alt and Kenneth A. Shelpsle. Cambridge Univ Press, 90–143.

Kuran, Timur. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification.* 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 & Organization* 76 (2):225–237.

Landier, Augustin, David Sraer, and David Thesmar. 2009. "Optimal Dissent in Organizations." *The Review of Economic Studies* 76 (2):761–794.

Larkin, Ian. 2011. "Paying 30K for a Gold Star: An Empirical Investigation Into the Value of Peer Recognition to Software Salespeople." Harvard Business School.

Levy, G. 2005. "Careerist Judges and the Appeals Process." *RAND Journal of Economics* 36 (2):275–297.

———. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting rules." *American Economic Review* 97 (1):150–168.

Linardi, Sera and Margaret A McConnell. 2011. "No excuses for good behavior: Volunteering and the social environment." *Journal of Public Economics* 95 (5):445–454.

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):428–461.

Morgan, John and Phillip C. Stocken. 2008. "Information Aggregation in Polls." *American Economic Review* 98 (3):864–896.

Morris, Stephen. 2001. "Political Correctness." *Journal of Political Economy* 109 (2):231–265.

Morris, Stephen and Hyun Shin. 2002. "Social Value of Public Information." *The American Economic Review* 92 (5):1521–1534.

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 Hansen, and Stephen Turnovsky. Cambridge University Press, 56–114.

Ottaviani, Marco and Peter Sørensen. 2001. "Information Aggregation in Debate: Who Should Speak First?" *Journal of Public Economics* 81 (3):393–421.

Posner, Eric A. 1998. "Symbols, Signals, and Social Norms in Politics and the Law." *Journal of Legal Studies* 27 (2):765–797.

———. 2000. *Law and Social Norms*. Harvard University Press.

Posner, Richard A. 1977. "The Right of Privacy." *Georgia Law Review* 12:393–422.

———. 1979. "Privacy, Secrecy, and Reputation." *Buffalo Law Review* 28:1–55.

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

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* .

Ronson, Jon. 2015. "How One Stupid Tweet Blew Up Justine Sacco's Life." *New York Times* .

Segal, David. 2013. "Mugged by a Mug Shot Online." *New York Times* .

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

Van der Weele, Joel. 2013. "The Signalling Power of Sanctions in Social Dilemmas." *Journal of Law, Economics and Organization* 28 (1):103–25.

Vesterlund, Lise. 2003. "The informational Value of Sequential Fundraising." *Journal of Public Economics* 87 (3):627–657.

Visser, Bauke and Otto H. Swank. 2007. "On Committees of Experts." *Quarterly Journal of Economics* 122 (1):337–372.

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

Whitman, James Q. 1998. "What Is Wrong with Inflicting Shame Sanctions?" *The Yale Law Journal* 107 (4):1055–1092.