Image Versus Information:
Changing Societal Norms and Optimal Privacy

S. Nageeb Ali[1] Roland Bénabou[2]

This version: September 2016 [3]

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

[2]Princeton University, NBER, CEPR, IZA, CIFAR, BREAD and THRED. 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, Justin Rao, Joel Sobel, and Pierre-Luc Vautrey. We thank Edoardo Grillo, Tetsuya Hoshino, Charles Lin, Pellumb Reshidi, and Ben Young for superb research assistance. 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 signal about the value 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 and matching rate, then analyze how they depend 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, 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... The law should in general accord private business information greater protection than it accords personal information. 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?

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 before him, 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 important for institutional design, and of growing policy relevance. Many public and private entities already use esteem as a motivator, such as the military, which offers medals for valor; businesses, which recognize the "employee-of-the-month"; and charities publicizing 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 and pictures of people convicted or even just arrested for a host of offences (tax evasion, 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, signs in front of their houses or paid ads in the newspaper. 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

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

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 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 budget-constrained public authorities, activist groups or individual whistleblowers and "concerned citizens." 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 (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 the argument to its extreme, arguing that there should be essentially zero privacy protection for facts concerning individuals whatever their nature (e.g., sexual behaviors, religious or political opinions, decades-old offenses or medical conditions), including no attorney or spousal exemptions from testimony and no right not to self-incriminate. In this view, market forces will always ensure that any "irrational" discrimination based on revealed attributes and actions is quickly eliminated, leaving only efficient uses of the information that create reputational incentives for socially beneficial behavior.

Yet there remains substantial unease at the idea of shaming as a policy tool, and more generally a widespread view that a society with zero privacy would be "unlivable." 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 across places, times and judicial interpretations. Besides the general attachment to anonymous voting as indispensable to democracy (Brennan and Pettit (1990), however, argue that publicizing votes would better align then with the general interest, versus private ones), there are also many instances where social institutions preserve privacy, even though publicity could offer a powerful tool to 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 the consumption of publicly provided or funded health care, there is no policy to "out" those who impose the highest costs as the result of partially controllable behaviors such as smoking, poor diet, or addictions. On the contrary, there are

[2] In Greece, tax authorities have released lists of major corporate and individual tax evaders. 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)). In France, a July 2014 law allows judges finding a firm or an individual guilty of undeclared employment to post, for up to two years, their names and professional addresses on an internet "black list" hosted by the Ministry of Labor. Shaming can also be spontaneously organized by activists, as with the "Occupy Wall Street" movement, or the hacking of Ashley Madison's list of user identities. There is even a growing movement of frustrated parents posting videos on the internet and social media to publicly shame their misbehaving children.

[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 and municipalities all across the country, then asking the people involved for a hefty fee in order to take down the post concerning them. (Segal (2013)). There are also more established companies serving businesses by "managing" their on-line reputations in consumer forums, blogs, etc.

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 obviously a case for protecting individuals' information from the eyes of parties with potentially malicious intent: undemocratic government repressing dissenters, firms using data on consumer's habits and spending patterns to engage in price discrimination or exploitation, hackers intent on identity theft and rivals seeking to steal trade secrets. While these issues are undeniably important, we focus here on identifying very different costs of transparency, related to *evolving social norms* and the *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 always leads agents to provide more public goods, it is optimal to maintain or protect a certain degree of privacy. This remains a fortiori true under less ideal conditions.

1.2 Our Framework

The key idea 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, most importantly, for those who *wield it.* Our argument builds on two complementary mechanisms:

1. *Inefficient variability in the power of social image.* The rewards and sanctions generated by publicizing an individual's actions stem from the reactions that this knowledge elicits from his family, peers, or neighbors. These social incentives thus involve the *emotional responses* of many people as well as their degree of *coordination,* which makes their severity hard to predict and fine-tune *a priori* (E. Posner (2000)). Depending on place, time, group, offense and individual contingencies, the feared response may go from mild ostracism to mob action, be easy or hard to escape, etc.[4] Variability in the strength of 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 individual behaviors are made more visible or salient.[5]

2. *Rigid and maladaptive public policy.* Public stigmatization has long been used to repress non-believers, mixed-race relationships, single-mothers, homosexuals, etc. But, of course, the purpose at the time was precisely to discourage such behaviors, widely considered immoral and socially nefarious, and accordingly also punished by the law. The real problem

[4] On such instability and indeterminacy in collective-action outcomes, see Lohmann (1994) and Kuran (1997). The current "explosion" of shaming on social media is a good example of this variability. In many instances, the resulting costs to the punished party (loss of job and family, suicide) ended up being wildly disproportionate to the perceived offense. Sometimes there is even a backlash, where individuals who played a key role in coordinating a shaming that "went too far" are themselves publicly shamed on the same media (Ronson (2015)).

[5] Similar effects of variability occur if social sanctioning involves (convex) resource costs, or if agents are risk averse. We abstract from these channels, since they would lead to very similar results as those we focus on.

is that *societal preferences change* unpredictably due to technology, enlightenment, migration, trade, etc. In order to learn how policy –the law and other institutions, taxes and subsidies, etc.– should be adapted to recent evolutions, an imperfectly informed principal must assess societal preferences from prevailing behaviors and mores. If individuals feel too constrained by the fear of social stigma and sanctions from others, these preference shifts will remain hidden, or be revealed too slowly. The result will be a rigidification and maladaptation not only in *private conduct* –excessive conformity– but also in *public policy*, doubly impacting the efficiency of resource allocation.

Full reversals of societal preferences, where some behavior like overt racism, sexism or domestic violence goes from "normal" to deeply scorned, or on the contrary from intensely stigmatized to widely acceptable, like divorce, cohabitation ("living in sin"), homosexuality or drug use, are relatively common in modern societies and sometimes quite sudden. It is thus all but certain that some conducts generally seen as abhorrent and shameful today will become perfectly mundane within a couple of decades, and vice-versa. Uncertainty lies only in which ones it will be and which way the cursor will move. Possible candidates of both types include organ sales, prostitution, extramarital and other parallel relationships, eating meat and wearing animal products, atheism and apostasy, transhumanistic enhancements and others we cannot yet conceive of. Ignoring that what constitutes a public good, a heinous deed or a wide-ranging externality (and hence also a proper signal of prosociality) is subject to important and unpredictable shifts, is the fundamental error of "minimal-privacy" advocates such as R. Posner, who in the process explicitly equate social conformity with the common good:

> *"At one level, Bloustein [an earlier law scholar advocating for the right to privacy] 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 apparently... for reasons he must consider self-evident since he does not attempt to explain them." (R. Posner (1977), pp. 401-405).*

We will show that the problem with low privacy lies not in increased uniformity of individual behaviors (it may in fact have the opposite effect, generating inefficient image-seeking variations) but in the reduced informativeness of *aggregate* behavior, which impedes everyone –in particular, the legislator– from learning and adapting to recent evolutions in societal preferences. Even for behaviors that remain unambiguously bad or good from a social point view (drunk driving, tax compliance, etc.), moreover, the tradeoff we identify remains. As long as their *relative* importance to social welfare fluctuates over time (with sign unchanged), a policymaker will need to learn of these evolutions through shifts in aggregate behavior, so as to appropriately redirect limited financial or enforcement resources.[6]

[6] Only for actions which the Principal *always* wants to deter maximally (corner solution), obviating the need to ascertain anything about how they affect agents, does the effect vanish; heinous murders may be such a case.

Formally, we study 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 the public good, and their collective information, suitably aggregated, is a precise signal of its social value. Each chooses how much to contribute, based on his own mix of public-spiritedness, information and reputational concern for appearing prosocial. The Principal can amplify or dampen these reputational payoffs, and hence total contributions, by making individual behavior more or less visible to the community. While this entails little cost (none, for simplicity), she faces an informational 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. Learning about public good quality or externalities is important for choosing her own (e.g., tax-financed) contribution, matching rate or other policy, such as the law. If the Principal suppresses image motivations by making contributions anonymous, she can precisely infer societal preferences from agents' aggregate behavior. However, each individual will then free-ride to a greater extent, leaving her with a greater share of the burden in achieving the desired level 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 esteem. The Principal thus faces a tradeoff between using *image as an incentive* and gaining better *information* on societal preferences.[7]

We first study agents' equilibrium behavior and learning under any fixed level of publicity, establishing a surprisingly simple *benchmarking* result for how social inferences take place in this complex environment (multidimensional signaling and heterogenous, higher-order beliefs). We then consider the Principal's optimal choice of publicity and the tradeoffs arising from agents' responses. We show that a positive level of privacy must always be maintained and fully characterize its comparative statics, as well as those of the Principal's second-stage policy (matching rate) with respect to her cost of funds, the idiosyncratic and aggregate variances of agents' preferences, and both sides' quality of information. 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 be greater* than in a more static or "traditional" one, where preferences vary mostly across individuals but are stable at the societal level.

1.3 Applications

From social norms to formal institutions. Formal 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 change over time, sometimes quite radically and very fast. Where behavior is highly constrained by the fear

[7] 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.

of social stigma, assessing social preferences and shaping laws by what people do (*"descriptive norm"*) can be a very poor indicator of what they really value (*"prescriptive norm"*).

Similar issues arise in the debate over freedom of speech versus "political correctness." Reputational concerns lead people to refrain from acts and speech considered to be offensive (Loury (1994), Morris (2001)), so activist groups, media outlets and institutions like universities commonly use publicity (and rules) to curtail such behaviors.[8] Without sufficient privacy protections, however, what people have really come to think will be learned only too late.

Public good provision, charitable donations. We cast the model in terms of a classical benchmark: providing the "right kind" of public goods in a cost-effective manner. This also facilitates comparison with previous work. 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 parts of the world. 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.

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 conformity 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.

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.

Leadership. As emphasized in the literature on corporate culture, a key role of leadership is to coordinate expectations and efforts toward goals that reflect shared objectives and beliefs (Kreps (1990), Hermalin and Katz (2006), Bolton, Brunnermeier, and Veldkamp (2013)). Our analysis highlights how a leader faces the challenge of using publicity to align agents' goals and values with those of the organization, while also allowing enough dissent and contrarian behavior for her (and others) to learn how these should adapt over time.

Political activism. The Principal can also stand for an electorate, while agents are activists and informational lobbies exerting effort to persuade voters about some drastic reform. When

[8] Thus, a recent activist campaign in Brazil tracks down the geotagged locations of people who post racist comments on social media, then reposts them on giant billboards and public buses in the immediate neighborhood of the source (with names and profile pictures blurred, however).

the media makes their actions more visible, activists are willing to take more costly steps, so publicity again provides incentives. At the same time, activism is discounted to a further extent as being "attention-seeking," and indeed may not offer much useful information.[9]

1.4 Related Literature

Our study relates to several parts of the large literature examining the impact of transparency on individual and collective decision-making, and thus ultimately on institutional design.

A first strand focuses on signaling, especially in a public-goods context.[10] Our model builds on Bénabou and Tirole (2006) who study how incentives, whether material or social, can undermine individual's reputational returns derived from a prosocial activity. We develop this basic framework in two important, more *aggregate* directions. First, a Principal explicitly chooses how much agents know about each other's behavior, internalizing their equilibrium responses. Second, she is imperfectly informed about the social value of the activity, generating a tradeoff between image incentives and information aggregation that is a novel feature of our model.[11] Also closely related is Daughety and Reinganum (2010), who study how making actions fully public can result in the overprovision of public goods, whereas making them fully private can result in underprovision. We consider the problem of a Principal who can adjust continuously how much privacy to accord individuals, faces uncertainty about they will respond to it and, most importantly, cares about the informational content of their behavior.[12]

Transparency is also a central issue when experts, judges, or committee members have career concerns over the quality of their information (rather than their prosociality), 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, excessively contradict seniors or 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

[9] Lorentzen (2013), for instance, studies how China's government relies on public protests as a signal of local corruption. Our point is that media attention to these protests helps mitigate collective action problems, but also interferes with information transmission when activism becomes image-driven.

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

[11] Excessive constraints on behavior (commitment devices, monitoring with threats of punishment) can also interfere with learning about agents' individual types, rather than with information aggregation over the state of the world; see, e.g., Bénabou and Tirole (2004), Ichino and Muehlheusser (2008) and Ali (2011).

[12] 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 fact that here the Principal's not knowing the exact value of image is important here.

Swank (2007)). On the normative side, whether conformity or exaggeration is worse for the Principal, and whether she prefers transparency or 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)). In our framework, agents' incentives to signal their types increase rather than decrease conformity, and the latter has simultaneously positive (mean-contribution) and negative (excessive variance and information-garbling) effects. Another key difference is that the strength of image concerns, which is common knowledge in nearly all of the signaling literature, is here one of the key sources of uncertainty.[13]

As stated earlier we do not deal here with government snooping, consumer exploitation by firms, the theft of identity or trade secrets, which all involve principals seeking to misuse agents' data (see Acquisti, Taylor, and Wagman (2016) for a survey). Our focus is instead on what private citizens *know about each other's behaviors* and on the social value of privacy even when the Principal may be benevolent (though we also allow her not to be).[14] Finally, in emphasizing how laws emerge from evolving social norms our paper relates to a growing literature on how formal and informal institutions shape each other (e.g., Bénabou and Tirole (2011), Jia and Persson (2013), Besley, Jensen, and Persson (2014), Acemoglu and Jackson (2016)).

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

[13] Bénabou and Tirole (2006) study signaling agents with heterogenous (privately known) image-concerns, and Fischer and Verrecchia (2000) and Frankel and Kartik (2014) agents with heterogenous payoffs to misrepresenting their actions. In such settings, greater visibilty makes each individual's observed behavior less informative about his true motivations. In none of these papers is there any aggregate uncertainty, nor a Principal who seeks to incentivize agents and/or learn from their behavior.

[14] For the same reason we abstract from concerns that public shaming constitutes cruel humiliation that negates other important societal values such as human dignity. See, e.g., Posner (1998) for such arguments and Bénabou and Tirole (2011) for an analysis of expressive law, including the case of "cruel and unusual punishments."

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.[15] 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. 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.[16] 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.[17] 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 the 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 net 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 $x \geq 0$ parametrizes the degree

[15] We model agents' preferences as separable in intrinsic motivation and quality for analytical tractability, but the basic insights are robust to relaxing this assumption; see Section 2.1 for a discussion.

[16] 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.

[17] 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.

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" utility, 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.[18]

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 the Principal, for instance by polling agents about the quality of the public good or the importance of social image.[19]

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 signal about quality, θ_i, and how important social esteem is to him, 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 focus, for tractability, on Perfect Bayesian Equilibria in which an agent's contribution is linear in his type, (v_i, θ_i, μ_i). This will be shown to imply that an equivalent formulation of the Principal's decision problem is:

(a) Given any x, optimally set a baseline investment level she will provide (based on her own signal) and a *matching rate* on private contributions: $a_P = \underline{a}_P(x, \theta_P) + m(x)\bar{a}$.

(b) Based on ex-ante information only, set x optimally.

[18] 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 Principal who intrinsically *dislikes* the activity that agents consider socially valuable –political opposition, cultural resistance, etc.

[19] This information is typically limited: polling is costly (see Auriol and Gary-Bobo (2012) on the optimal sample size or number of representatives) and invites strategic responses from agents who would like the Principal to contribute more (Morgan and Stocken (2008), Hummel, Morgan, and Stocken (2013)). Allowing the Principal to obtain an independent, noisy signal of μ would also not affect our analysis.

2.1 Discussion of the Model

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

1. Agent's contributions become driven in larger part by variations in their social-image concerns, rather than by their signals concerning 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 conformist –that is, too uncorrelated with the true quality of the public good, and too difficult to for her to learn from.

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

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 θ_i and strong image concerns μ_i. Moreover, each agent lacks information about others' signals and so cannot perfectly anticipate how they will interpret his actions. Social incentives thus involve both multidimensional heterogeneity and higher-order uncertainty, making the problem a complex one. Specifying agents' preferences as separable in intrinsic motivation and public-good quality allows us to keep it tractable and derive simple, closed-form solutions. The basic 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 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 detached 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 6.2, 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.[20] The choice of

[20] 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.*

publicity / privacy would then also commingle the Principal's motive to learn from agents with her incentive to signal to them.

Principal's Policy The Principal chooses her provision level 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.[21] For simplicity, and to focus squarely on the effects of publicity, we abstract from the use of price incentives to induce compliance. More generally, material incentives entail both direct and indirect costs that limit the extent to which they can be used.[22]

Private vs. Common Values In the benchmark case, we model agents' ex-post payoffs from contributing to and consuming the public good as reflecting some objective, universally agreed-upon quality or social value θ; this corresponds to a setting with *common values*. The model also applies, however, to the *private values* case, in which each agent's ex-post payoff depends on his own perspective or taste θ_i concerning the public good. In that case, s_θ^2 reflects the *dispersion of views* and the Principal cares about θ as reflecting the *average preference* for the public good or externalities-generating action.[23] We show in Section 6.1 that all key results in this case closely parallel those of the benchmark one, with all formulas being either identical to, or a simpler special case of, the corresponding ones in the original specification. This generality is important, as some of the applications discussed earlier fit better the common-values setting, others the private-values one. The former include charitable contributions and clearly prosocial (or antisocial) behaviors affecting a group's safety or the extent of rent-seeking within it, as well as incentives and leadership in a firm. Among the latter are most social mores and norms, as well as political and religious opinions.

3 Equilibrium Behavior: How Agents Respond to Publicity

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:

[21] 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.

[22] We can thus also define ω as the wedge left after the Principal has already used any standard incentives at her disposal. Note also that it will always be optimal to use some positive level of publicity as an additional incentive: the gain is initially first-order, whereas the induced distortions are second-order.

[23] This distinction similar to the one discussed within the context of global games (Carlsson and Van Damme (1993), Morris and Shin (2006)).

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

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*

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

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

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

A sufficient-statistic result. Greater intrinsic motivation v_i, better perceived quality θ_i and a stronger image concern μ_i naturally lead an agent to contribute more. Most remarkable, however, is the *simplicity* of the social-image computations that emerge from this complex setting, as reflected by the common marginal impact $\xi(x)$ that an additional unit of contribution has on one's expected image. This is also, intuitively, the *signal-to-noise ratio* faced by an *observer* when trying to infer someone's type v_i from their action, knowing that behavior reflects private preferences, private signals and image concerns according to (B.3).

Strikingly, the expected image return is the *same for all agents,* even though they have very 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 (B.3). 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 $\rho\theta_i + \mu_i x\xi\left(x\right)$, and for this purpose $\rho\theta + \mu x\xi\left(x\right)$, hence also $\bar{a}$, is a *sufficient statistic.* Put differently, whereas $E\left[\theta_i|a, \bar{a}, \theta_j, \mu_j\right]$ and $E\left[\mu_i|a, \bar{a}, \theta_j, \mu_j\right]$ both depend on j's private type, $E[a_i - \rho\theta_i - \mu_i x\xi\left(x\right)|a_i, \bar{a}, \theta_j, \mu_j]$ does not, so all observers

of agent i again share the *same beliefs* about his motivation: $E[v_i|a,\bar{a},\theta_j,\mu_j] = E[v_i|a,\bar{a}]$.[24] Agent i's *ex-post* reputation will thus be a linear function of $a_i - \bar{a}$ only, 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: $\partial R\left(a,\theta_i,\mu_i\right)/\partial a = \xi\left(x\right)$.

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:

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

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

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

where $\xi(x)$ is given, as a fixed point, by (B.4).

In anonymous settings, $x = 0$, or 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, $x^2 s_\mu^2 > 0$, 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 $\xi\left(x\right) < \xi(0) \equiv \xi$, with a slight abuse of notation. This "over-justification 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 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, $\beta(x) \equiv x\xi\left(x\right)$, is strictly increasing in x, with $\lim_{x\to\infty}\beta(x) = +\infty$, and it shares the properties of $\xi\left(x\right)$ with respect to all variance parameters. The same is true of the aggregate contribution $\bar{a}(x)$, as long as $\mu > 0$.[25]

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

[24] 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.

[25] 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.

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.[26]

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.

Exogenous variations in privacy. Inspection of (B.5) 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.

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. [27]

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 image motives, $\sigma_\mu^2 = 0$.

[26] This somewhat subtle comparative static emerges in the common-value environment, where each agent estimates the quality of the "true" public good, based on his signal. By contrast, in the private-values environment studied in Section 6.1, agents effectively behave as if $\rho = 1$; in that case, the social image return $\xi(x)$ is strictly decreasing in s_θ^2.

[27] 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.

(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 $x\mu$ 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 will focus until Section 4.4 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. Agents' social-learning problems, meanwhile, become simpler, with the image return $\xi(x)$ reducing to the constant ξ given by (14).

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

Consider first the simple case where agents' image motive is invariant: both they and the Principal believe with probability 1 that $\mu = \bar{\mu}$, so $\sigma_\mu^2 = 0$. Upon observing the aggregate contribution $\bar{a}$, the Principal perfectly infers θ by inverting (B.5), allowing her to optimally set

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

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 with respect to x leads to an optimal level of

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

where the superscripts stands for "First Best" and the wedge $\omega > 0$ was defined in (6).[28]

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 $\lambda = \frac{1}{2}$. 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

[28] This result is a special case in the proof of Proposition 3 below.

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 Accounting for Variability in the Image Motive

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 achieve precise control of agents' compliance and achieve her first-best through (15)-(16). We show below that this leads her, *even if she observes* the realization of μ *ex post*, to moderate her use of visibility as an incentive mechanism.

A principal who learns the realization of μ (once x has been set) is again able, upon observing $\bar{a}$, to infer the true θ by inverting (B.5). 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 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, which embodies two opposing effects:

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

The two 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.*

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 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 (B.5), 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 with mean $\bar{\theta}_P$ and variance $\sigma^2_{\theta,P}$:

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

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

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

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

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.[29] 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, (B.5), (20) and (23) imply that the Prin-

[29] 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.

cipal'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

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

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*

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

In general, (29) could have multiple solutions, because the cost of information distortion is not globally convex: the marginal loss, proportional to $\gamma(x)^2x$, is hump-shaped in x.[30] 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, $s^2_\mu>0$, 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 (B.5), idiosyncratic differences in μ_i's wash out in the aggregate

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

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 $\tilde{\alpha}$. 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 the yields the following results.

Proposition 6. *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*

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

where $\xi(x)$ is given by (B.4) 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.[31]

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 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.[32]

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.

[31] Throughout much of our analysis below, in describing how x varies with a primitive, we establish that the objective function is supermodular in x and that primitive. It follows by Topkis's Theorem that the set of global maximizers is then increasing in x and that primitive. If the gusual case where the global maximizer is unique, then optimal publicity is increasing in that primitive in the usual way; if there are multiple global maximizers, then the set of maximizers is increasing in the strong set order.

[32] 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 $(-1/2)\int_0^1 a_i^2 d_i$.

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 perfect, or more generally 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 (B.4), (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.*

We note that the assumption that $s_\theta^2/\sigma_\theta^2$ is sufficiently small is restrictive, but captures the environment in which the Principal has a lot to learn from agents, or where the prior is extremely uninformative. Moreover, as we discuss in Section 6.1, these comparative statics always emerge in the setting with private values, regardless of the value of $s_\theta^2/\sigma_\theta^2$. Accordingly, in a "fast-changing" society, where the Principal ex ante knows little about the quality of the public good, she shall wish to have greater privacy if agents are relatively better informed than she is, or if the setting is that in which agents have heterogeneous views on the public good—as in a private values environment—as we discuss further in Section 6.1.

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 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 first-best level x^{FB} that solve $\beta(x) = \bar{\mu}\omega / \left[\lambda\left(\bar{\mu}^2 + s_\mu^2\right)\right]$.

2. Heterogeneity in image concerns. For given x, an increase in s_μ^2 influences the image incentive $\beta(x)$ in complex ways (see (B.4)), 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 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 agents' common prior over θ is far less informative than their private signal: 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.*

The assumption that $s_\theta^2/\sigma_\theta^2$ is sufficiently small is somewhat restrictive, but corresponds to the most relevant environments for our purposes, namely those in which the Principal has "enough" to learn from agents' compliance. Moreover, in the closely parallel setting with *private values*, the above comparative-statics hold regardless of the value of $s_\theta^2/\sigma_\theta^2$ (see Section Section 6.1 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 Extensions

6.1 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 is now certain of 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. 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.

The analysis of private values turns out to be quite simple at this stage, as it virtually reduces to a special case of the common-values setting. Relegating details to the Supplementary Appendix, we provide here only the key point: because each agent's view of the public good is driven solely by her perspective on it, θ_i, she contributes as if she were getting a perfect

signal about a common value, that is, as if $\rho = 1$.[33] The entire analysis is thus the same as in Sections 2-5 but replacing ρ by 1 in all formulas in the text. An important implication is that, in a private values environment, all of the prior comparative-statics results carry over, but with those relative to s_θ^2 and σ_θ^2 now holding *without any restriction* on the ratio s_θ/σ_θ.

6.2 Noisy Observability

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

$$\xi(x) = \frac{s_v^2}{\xi(x)^2 s_\mu^2 + s_\varepsilon^2/x^2 + s_v^2 + \rho^2 s_\theta^2}, \tag{33}$$

which remains quite similar (though of course not identical) to the earlier expression, (B.4). Consequently, we show in the supplementary Appendix that all key implications remain unchanged [To be completed].

6.3 Simple Dynamics

Our model has highlighted the effects of aggregate variability (and idiosyncratic differences) in societal preferences on agent's behavior and the optimal decisions of a Principal. While the model is a static one, we occasionally interpreted the results in dynamic terms, e.g. a fast- or slow-changing society, formal institutions adapting to those changes or remaining rigid, etc. Formally extending the results to a simple dynamic environment is straightforward, but we do so here for the sake of completeness.

Let each generation of agents live for one period, subdivided into two subperiods during which they interact among themselves and with a Principal just as before (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 (e.g., their parents') 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 is long-lived, discounting payoffs at rate δ.

[Preliminary] It is straightforward to see that, 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

[33] 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.

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.

6.4 Publicity and Formal Law

In our motivating examples, we mentioned the fact 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 gradually 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^* 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 a^* should ultimately be determined. All the results, including the Principal's choice of publicity and its comparative statics properties, remain closely analogous to the earlier ones, as we show in the Supplementary Appendix. In the process we also derive the optimal mandate a^*, and compare it to the Principal's a_P in the original specification.

6.5 Publicity and Material Incentives

6.5.1 First-period incentives

Consider a Principal who, in the first period of the baseline model (agents only contribute once, then the Principal chooses her own a_P), can provide agents with monetary or other material incentives of y per unit of compliance a_i, but faces a cost of funds of $(1 + \kappa)$ dollars per dollar of payment, where κ represents, e.g., the deadweight loss from taxation. The Principal can also set the level of publicity x, as before, at no cost. We show in the Supplementary appendix that the baseline model's main results, and in particular those concerning the information effect, the optimal level of privacy and its comparative statics, remain unchanged. We also derive the optimal incentive rate y, and show precisely how the principal's mix of publicity and incentives (x, y) depends on their relative costs and how much agents value reputation.

6.5.2 Second- period incentives

The case of a law or mandate a^* can be thought of as one where agents are under extremely strong incentives (e.g., prohibitively high fines, prison sentences), which the Principal can deliver at (implicitly) very low cost. In practice, enforcement is costly, and many policies also take the

form of subsidies –or, in a firm, bonuses– which also entail non-trivial resource costs. To deal with this set of applications (discussed earlier), one can in a sense combine the two previous extensions:

(i) In period 1 the Principal sets publicity x, then agents choose their contributions a_i (unincentivized, for simplicity).

(ii) In period 2 the principal chooses an incentive rate y', with opportunity cost $(1+\kappa)y'$, then agents choose individual contributions a'_i

(iii) For simplicity, no reputational payoffs are attached to second-period contributions: agents "die" or the group dissolves shortly after, so that (unlike in period 1) there is no subsequent continuation game in which image would matter.

The same core insights apply here again, in particular, on how the optimal x will reflect the extent to which formal incentives y' will need to be adapted to likely shifts in societal preferences, agents' or the principal's information, or other parameters. Correspondingly, we show in the Supplementary Appendix that the form of all equilibrium solutions, as well as their comparative statics, remain again unchanged.

7 Conclusion

We studied the 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. Among other results, we thus showed 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.

There are several directions in which the analysis could be further developed. A first one is an overlapping-generations environment in which the value of the public good and the strength of reputational concerns evolve stochastically, and where agents receive signals and care about their reputations in both periods of life. over time. Compared to the simpler dynamic extension considered above, this will introduce interesting lifecyle effects. First, older agents are less responsive to publicity (and more to fundamental information), since their past record is already indicative of their type, whereas younger agents are more keen to signal their motivation through their actions. Second, older agents may be better informed than young ones (provided they

observe signals in both periods of life), which as we saw has a U-shaped effect on reputational incentives.

A second extension would be to examine mechanisms by which principals may alleviate the informational problem we identify. This could for instance involve a two-stage procedure, in which agents first choose their participation levels anonymously –thereby revealing the state– then, in a second stage, are asked to contribute. Such dynamic procedures may lead to efficiency gains because: (a) information is better revealed in the first stage; (b) in the second stage, image is even more responsive to contributions than before, as the informational effect (rationalizing a low contribution as possibly reflecting a low private signal) is eliminated. Of course, such mechanisms may not be feasible in all contexts.

8 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 over θ and μ, we obtain from (4):

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

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, therefore the first-order condition described in (18) characterizes the unique optimum. Equating the right-hand-side to zero yields (19), which simplifies to (16) when $\sigma_\mu^2 = 0$. ■

Proof of Proposition 4 on p. 19

The formula for $m(x)$ follows directly from the reasoning in the text. As to the baseline investment,

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

it 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, recall that she would choose a contribution level of $(w+\theta)\varphi/(1-\lambda)k_P$. 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

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

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

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

Therefore

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

which corresponds to (29). Recall now that $E\tilde{V}(x)$ is strictly concave in x and maximized at $\tilde{x} > 0$. Therefore, $\partial EV/\partial x < \partial E\tilde{V}/\partial x < 0$ 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. 21

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*

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

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

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

or

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

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:

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

Recalling that

$$\gamma\left(x\right)\equiv\frac{\rho^{2}\sigma_{\theta,P}^{2}}{\rho^{2}\sigma_{\theta,P}^{2}+\beta(x)^{2}\sigma_{\mu}^{2}}\quad\Rightarrow\quad\gamma'\left(x\right)=-\frac{2\sigma_{\mu}^{2}}{\rho^{2}\sigma_{\theta,P}^{2}}\beta(x)\beta'(x)\gamma(x)^{2},$$

this yields

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

which is equivalent to (32). ■

Proof of Proposition 7 on p. 22

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

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

in which

$$\tilde{\gamma}(z)\equiv\frac{\rho^{2}\sigma_{\theta,P}^{2}}{\rho^{2}\sigma_{\theta,P}^{2}+z^{2}\sigma_{\mu}^{2}},\tag{A.17}$$

and C is a constant that is independent of s_{v}^{2} and z. Therefore, the optimal z solves the first-order condition

$$\bar{\mu}\omega-\lambda z(\bar{\mu}^{2}+\sigma_{\mu}^{2}+(1-2\tilde{\alpha})s_{\mu}^{2})+\frac{\varphi^{2}\sigma_{\theta,P}^{2}}{2(1-\lambda)k_{P}}\tilde{\gamma}'(z)=0.\tag{A.18}$$

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 $x^{*}(s_{v})\xi(x^{*}(s_{v}),s_{v})$, it follows from (A.4) that increases in s_{v} must 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, it follows from (A.16) 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. 23

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

$$\frac{\partial^2 EV}{\partial x \partial \sigma_\mu^2} = -\lambda \beta'(x)\beta(x) - \frac{\varphi^2 \beta'(x)\beta(x)}{\rho^2(1-\lambda)k_P}\left(\gamma^2 + 2\gamma\sigma_\mu^2 \frac{d\gamma}{d\sigma_\mu^2}\right), \tag{A.19}$$

in which

$$\frac{\partial \gamma}{\partial \sigma_\mu^2} = -\frac{\rho^2\sigma_\theta^2\beta(x)^2}{\left(\rho^2\sigma_\theta^2 + \beta(x)^2\sigma_\mu^2\right)^2} = -\frac{\gamma\beta(x)^2}{\rho^2\sigma_\theta^2 + \beta(x)^2\sigma_\mu^2} = -\frac{\gamma\,(1-\gamma)}{\sigma_\mu^2}. \tag{A.20}$$

Thus,

$$\frac{\partial^2 EV}{\partial x \partial \sigma_\mu^2} = -\beta'(x)\beta(x)\left[\lambda + \frac{\varphi^2\gamma^2}{\rho^2(1-\lambda)k_P}\,(2\gamma - 1)\right]. \tag{A.21}$$

This expression is non-positive if and only if

$$\frac{\lambda(1-\lambda)\rho^2 k_P}{\varphi^2} \geq \gamma^2(1-2\gamma). \tag{A.22}$$

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

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

(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.22) 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.22) 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. 24

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

$$\xi = \frac{s_v^2}{x^2\xi^2 s_\mu^2 + s_v^2 + s_\theta^2}, \tag{A.23}$$

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.23) that $\xi(x^*(s_\theta),s_\theta)$ must decrease in s_θ. To compensate, $x^*(s_\theta)$ must then be increasing in s_θ. ■

9 Supplementary Appendix B: Formal Extensions

9.1 Analysis of Private Values in Section 6.1

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 the agent may estimate 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 the average sentiment towards the public good but also weights the agent's utilities, and therefore, in this private values environment, her final payoff is

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

We first describe how agents behave with respect to a particular choice of visibility x, and then we proceed to characterize the optimal degree of publicity and derive relevant comparative statics.

Proposition 13. ***(Equilibrium behavior and benchmarking)*** *Fix $x \geq 0$. There is a unique linear equilibrium, in which an agent of type (v_i,θ_i,μ_i) chooses*

$$a_i\,(v_i,\theta_i,\mu_i) = v_i + \theta_i + \mu_i x\xi\,(x)\,, \tag{B.3}$$

where $\tilde{\xi}\,(x)$ is the unique solution to

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

The resulting aggregate contribution (or compliance level) is

$$\bar{a}\,(x,\theta,\mu) = \bar{v} + \theta + \mu x\xi(x). \tag{B.5}$$

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

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

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

$$a_i=v_i+\theta_i+x\mu_i\left(\frac{Bs_v^2}{A^2s_\mu^2+B^2s_v^2+C^2s_\theta^2}\right). \tag{B.6}$$

Therefore, $B=C=1$, $D=0$, and $A=xs_v^2/\left(A^2s_\mu^2+s_v^2+s_\theta^2\right)$. Substituting $A=x\tilde{\xi}(x)$ yields (B.4). It remains to show that for each choice of x, $\tilde{\xi}\,(x)$ is unique. Given x, $\tilde{\xi}(x)$ solves the equation $\xi(x^2\xi^2s_\mu^2+s_v^2+s_\theta^2)=s_v^2$; the right-hand side is continuous and decreasing in ξ, clearly cutting the diagonal at a unique solution $\xi(x)$. Q.E.D.

Finally, equation (B.4)'s parallel with the expression in Proposition 2, easily yields the following results.

Proposition 14. ***(Comparative statics of social interactions)*** *In equilibrium:*

(1) The social-image return $\tilde{\xi}(x)$ is strictly increasing in the dispersion of agents' preferences s_v^2, and decreasing in the level of publicity x, in the dispersion of agents' image concerns s_μ^2, and in the dispersion of their opinions about the public good, s_θ^2.

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

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 $\tilde{\xi}(x)$ corresponds to the same formula as $\xi(x)$ previously, in which one replaces ρ by 1.

Principal's Problem: Her problem is unchanged, but for relevant substitutions: setting $\rho=1$, and adding to her payoff the term $\lambda\alpha s_\theta^2$, from her internalizing the aggregate payoff arising (by convexity) from the dispersion of contributions motivated by heterogeneity in 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.

9.2 Analysis of Publicity and Law in Section 6.4

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

$$U_i^1(v_i, \theta, w, a_i) = (v_i + \theta)a_i + (w + \theta)\bar{a} - \frac{a_i^2}{2}, \tag{B.7}$$

$$U_i^2(v_i, \theta, w, a^*) = (v_i + \theta)a^* + (w + \theta)\bar{a} - \frac{(a^*)^2}{2}, \tag{B.8}$$

and he solves: $\max_{a_i} \mathbb{E}\left[U_i^1 + \delta\, U_i^2 + (1+\delta)\, x\, \mu_i \left(R(a_i, \theta_i, \mu_i) - \bar{v}\right)\right]$

The Principal, similarly, maximizes $E[V^1 + \delta V^2]$, where

$$V^1 = \lambda\left(\alpha \int_0^1 (v_i + \theta)a_i\, d_i + (w + \theta)\bar{a} + \tilde{\alpha}\int_0^1 x\, \mu_i \left(\tilde{R}(a_i, \bar{a}) - \bar{v}\right) di - \int_0^1 \frac{a_i^2}{2}\, di\right) + (1-\lambda)b(w + \theta)\bar{a}, \tag{B.9}$$

$$V^2 = \lambda\left(\alpha \int_0^1 (v_i + \theta)a^*\, d_i + (w + \theta)a^* + \tilde{\alpha}\int_0^1 x\, \mu_i \left(\tilde{R}(a_i, \bar{a}) - \bar{v}\right) di - \int_0^1 \frac{(a^*)^2}{2}\, di\right) + (1-\lambda)b(w + \theta)a^*. \tag{B.10}$$

and $\tilde{R} \equiv \int_0^1 \mathbb{E}\left[v_i | a, \bar{a}, \theta_j, \mu_j\right]\, dj$.

Step 1: Solving for a_i: Following the exact same steps as in Proposition 1 leads again to $a_i = v_i + \rho\,\theta_i + (1-\rho)\bar{\theta} + x\xi(x)\mu$, where $\xi(x)$ is now defined as:

$$\xi(x) \equiv \frac{(1+\delta)s_v^2}{x^2\xi(x)^2 s_\mu^2 + s_v^2 + \rho^2 s_\theta^2}.$$

Step 2: Solving for a^*: In the second period, the Principal to chooses a^* to maximize $\mathbb{E}\left[V^2 | \bar{a}, \theta_P\right]$, leading to

$$\lambda\left(\alpha(\bar{v} + \mathbb{E}\left[\theta|\bar{a}, \theta_P\right]) + (w + \mathbb{E}\left[\theta|\bar{a}, \theta_P\right]) - a^*\right) + (1-\lambda)b(w + \mathbb{E}\left[\theta|\bar{a}, \theta_P\right]) = 0,$$

or

$$a^* = \frac{w\varphi + \lambda\alpha\bar{v}}{\lambda} + \frac{\varphi + \lambda\alpha}{\lambda}\mathbb{E}\left[\theta|\bar{a}, \theta_P\right]. \tag{B.11}$$

Step 3: Solving for x^*: If, after choosing x, the Principal is informed of the realized value of θ, or that of μ (allowing her to invert $\bar{a}$ and learn θ perfectly, this becomes $a^* = \left[(w\varphi + \lambda\alpha\bar{v}) + (\varphi + \lambda\alpha)\ \theta\right]/\lambda$.

Substituting into the objective function yields

$$
\begin{aligned}
E\tilde{V}(x) =& \lambda \left[\alpha \left((\bar{v}+\bar{\theta})(\bar{\bar{a}}+\delta\bar{a}^*) + s_v^2 + \rho\sigma_\theta^2 + \delta\frac{\varphi+\lambda\alpha}{\lambda}\sigma_\theta^2 \right) + \left((w+\bar{\theta})(\bar{\bar{a}}+\delta\bar{a}^*) + \rho\sigma_\theta^2 + \delta\frac{\varphi+\lambda\alpha}{\lambda}\sigma_\theta^2 \right) \right. \\
&+ (1+\delta)\tilde{\alpha}\frac{x^2\xi(x)^2}{(1+\delta)}s_\mu^2 - \frac{1}{2}[\bar{\bar{a}}^2 + s_v^2 + \rho^2(\sigma_\theta^2+s_\theta^2) + x^2\xi(x)^2(\sigma_\mu^2+s_\mu^2)] \\
&\left. -\frac{\delta}{2}\left((\bar{a}^*)^2 + \left(\frac{\varphi+\lambda\alpha}{\lambda}\right)^2\sigma_\theta^2 \right) \right] + (1-\lambda)\left[b(w+\bar{\theta})(\bar{\bar{a}}+\delta\bar{a}^*) + \rho\sigma_\theta^2 + \delta\frac{\varphi+\lambda\alpha}{\lambda}\sigma_\theta^2 \right],
\end{aligned}
\tag{B.12}
$$

Where:

$$
\bar{a}^* \equiv \frac{w\varphi+\lambda\alpha\bar{v}}{\lambda} + \frac{\varphi+\lambda\alpha}{\lambda}\,\bar{\theta}.
$$

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:

$$
x^* = \frac{\omega\,\bar{\mu}}{\xi(x^*)\,\lambda\,\left(\bar{\mu}^2+\sigma_\mu^2+(1-2\tilde{\alpha})s_\mu^2\right)}. \tag{B.13}
$$

Given the similarity in form with our earlier expressions, the same comparative statics follow here.

Second Case: Unknown θ: If the principal does not observe θ (or that of μ), the expectation that is part of (B.11) must be based on the information embodied in $\bar{a}$. Proposition 13 shows that this 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) $\gamma(x)$ by (24). The informational loss is thus still $V(\triangle) = \sigma_{\theta,P}^2(1-\gamma(x))$,with the only difference continuing to be the new definition of $\xi(x)$. Hence, once again:

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

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

$$
x^* = \frac{\omega\,\bar{\mu}}{\xi(x^*)\left(\lambda\left(\bar{\mu}^2+\sigma_\mu^2+(1-2\tilde{\alpha})s_\mu^2\right) + \frac{\delta}{\lambda}\left(\frac{(\varphi+\lambda\alpha)\sigma_\mu\,\gamma(x^*)}{\rho}\right)^2\right)}. \tag{B.14}
$$

Comparing to the optimal x^* found in the baseline model, namely

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

We see three intuitive and relatively minor, differences:

(i) The slightly different definition of $\xi(x)$.

(ii) The $(1-\lambda)k_P$ term no longer appears, since the Principal no longer contributes. In its place, we have agents' discounted second-period unit costs, to the extent λthat the Principal internalizes them.

(iii) The term φ becomes $(\varphi+\lambda\alpha)$, reflecting the slightly different "matching rates" of second-period (mandated) contributions a^*relative to that of the Principal in the original problem:

$$a^* = \frac{w\varphi + \lambda\alpha\bar{v}}{\lambda} + \frac{\varphi + \lambda\alpha}{\lambda}\,\mathbb{E}\left[\theta|\bar{a},\theta_P\right], \tag{B.15}$$

$$a_P = \frac{w\,\varphi}{(1-\lambda)k_P} + \frac{\varphi}{(1-\lambda)k_P}\mathbb{E}\left[\theta|\bar{a},\theta_P\right]. \tag{B.16}$$

9.3 Combining Material and Publicity Incentives in Section 6.4

9.3.1 First-period incentives

Denote $a_i(x)$ the equilibrium individual strategies of agents in the baseline model, namely (B.3). 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.[34] 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 (but not all of it, since they are costly), publicity has less of a role to play here. Second, by increasing agents levels of contributions, incentives raise their marginal costs, and thereby affect the variance effect.

Let us demonstrate these results formally, in the benchmark case of a benevolent Principal ($\lambda = 1/2$). Consider the Principal's objective function (4), and how it is affected by the presence of monetary incentives. Note first that the aggregate reputational-gains term (multiplying $\tilde{\alpha}$) is unchanged, since the presence of a known y does not affect anyone's image. Turning next to 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 really affecting any important results), we will therefore abstract from this term in what follows, assuming either that agents do not get additional intrinsic utility of the form $v_i y$, or just that $\alpha = 0$.

In the Principal's objective function, 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 $(1+\kappa)(\bar{a}(x)+y)y$

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

are thus the following:

$$\tilde{V}(x,y,a_P) = V(x,a_P) + (w+\theta)\varphi y - \frac{1}{2}\int \left[(a_i(x)+y)^2 - a_i(x)^2\right] di - \kappa y(\bar{a}(x)+y). \quad \text{(B.17)}$$

In the initial period, the Principal optimizes $E[\tilde{V}(x,y,a_P)]$ over (x,y) conditional on her priors, and 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:

$$E\left[\left((w+\theta)\,\varphi - y - \bar{a}(x) - 2\kappa y - \kappa\bar{a}(x)\right] = 0, \quad \text{or:}$$

$$y(x) = \frac{(w+\theta)\,\varphi - (1+\kappa)\bar{a}(x)}{1+2\kappa} = \frac{(w+\theta)\,\varphi - (1+\kappa)(\bar{v}+\bar{\theta}+\bar{\mu}x\xi(x))}{1+2\kappa} \equiv \tilde{y} - \delta\beta(x) \quad \text{(B.18)}$$

as long as the right-hand side is positive, which we shall assume: the cost of providing material incentives must not be too high, otherwise they will not be used and we are back to the benchmark case. This formula is quite intuitive, showing that the principal will use costly material incentives when image alone (which here has no direct costs, but does have indirect ones, as we saw), is insufficient to achieve her first best. Conversely, since $\beta(x)$ is increasing in x, (B.18) shows that, the lower is κ, the less publicity will be used.

Turning next to the optimal choice of x and using again the Envelope Theorem, we have, since $\partial a_i(x)/\partial x = \mu_i x\xi(x) = \mu_i\beta(x)$,

$$\frac{\partial E\tilde{V}(x,a_P)}{\partial x} = \frac{\partial EV(x,a_P)}{\partial x} - \int \left[(a_i(x)+y) - a_i(x)\right]\mu_i\beta'(x)di - \kappa y\bar{\mu}\beta'(x) \quad \text{(B.19)}$$

$$= \frac{\partial EV(x,a_P)}{\partial x} - (1+\kappa)\bar{\mu}y\beta'(x) \quad \text{(B.20)}$$

Evaluating the last term at the optimum $y = y(x)$) and substituting into (30) with $\lambda = 1/2$ yields the new first-order condition

$$\omega\bar{\mu} - (1+\kappa)\bar{\mu}\left[\tilde{y} - \delta\beta(x)\right] - \lambda\beta(x)\left(\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2\right) - \frac{\varphi^2\sigma_\mu^2}{\rho^2(1-\lambda)k_P}\beta(x)\gamma(x)^2 = 0. \quad \text{(B.21)}$$

Relative to the original, we see that the wedge ω is reduced to $\tilde{\omega} \equiv \omega - (1+\kappa)\tilde{y}$ and the coefficient on $-\beta(x)$ falls by $(1+\kappa)\bar{\mu}\delta = \bar{\mu}^2(1+\kappa)^2/(1+2\kappa)$. Thus, denoting $\tilde{\mu} \equiv \bar{\mu}\sqrt{1/2 - (1+\kappa)^2/(1+2\kappa)}$, the analogue of (32) is here

$$x^* = \frac{\bar{\mu}}{\xi(x^*)}\left(\frac{\tilde{\omega}}{\lambda(\tilde{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2) + \frac{(\varphi\sigma_\mu\gamma(x^*)/\rho)^2}{(1-\lambda)k_P}}\right), \quad \text{(B.22)}$$

with clearly unchanged comparative statics with respect to the key parameters of interest.

9.3.2 Second-period incentives

We now consider the case where the Principals' second-period policy is to set an incentive rate $\tilde{y}$ 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 (equivalently, no period 3 , in which agents play some continuation game were reputation is valuable). As to the Principal, she again has intertemporal objective function $V^1 + \delta V^2$, with components now given by:

$$V^1 = \lambda \left(\alpha \int_0^1 (v_i + \theta) a_i \, d_i + (w + \theta)\bar{a} + \tilde{\alpha} \int_0^1 x \, \mu_i \left(\tilde{R}(a_i, \bar{a}) - \bar{v} \right) di - \int_0^1 \frac{a_i^2}{2} \, di \right)$$
$$+ (1 - \lambda) b (w + \theta) \bar{a}, \tag{B.23}$$

$$V^2 = \lambda \left(\alpha \int_0^1 (v_i + \theta)(a_i' - y') \, d_i + (w + \theta + y')\bar{a}' - \int_0^1 \frac{(a_i')^2}{2} \, di \right)$$
$$+ (1 - \lambda) \left[b(w + \theta) - (1 + \kappa) y' \right] \bar{a}', \tag{B.24}$$

where "primes" denote second-period actions and, as in assumed 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'.[35]

Using the notation $a_i(x)$ introduced above to denote equilibrium contributions in the baseline model, 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 individuals' actions about their type, and the informativeness $\gamma(x)$ of aggregate compliance $\bar{a}(x)$ for the Principal's learning 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), that $\lambda = 1/2$. The problem of the Principal in period 2 is to choose y' to maximize $E[V^2 | \theta_P, \bar{a}]$, where $\bar{a}$ is the aggregate contribution from period 1 :

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

[35] One could of course allow for incentives y and y' in both periods, but this would not as much insight relative to the simpler treatments provided here and above, in which we look at them in turn.

Taking the first order condition with respect to y' yields

$$\begin{aligned} 0 &= E\left[(w+\theta)(1+b) - y' - \bar{a}(0) - \kappa\bar{a}(0) - 2\kappa y|\theta_P, \bar{a}\right] \iff \\ y' &= \frac{w(1+b) - (1+\kappa)E[\bar{a}(0)|\theta_P, \bar{a}] + (1+b)E[\theta|\theta_P, \bar{a}]}{1+2\kappa} \iff \\ y' &= \frac{w(1+b) - (1+\kappa)(\bar{v} + (1-\rho)\bar{\theta})}{1+2\kappa} + \frac{(1+b) - \rho(1+\kappa)}{1+2\kappa}E[\theta|\theta_P, \bar{a}] \end{aligned} \tag{B.25}$$

as the optimal level of incentive given the principal's information on θ.

Now consider 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. Therefore, there only remains to solve for the optimal level of x. As usual, consider first the case in which θ will become known to the principal at the beginning of period 2. Then, (B.25) becomes:

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

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

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

Suppose, finally, that the principal does not observe either θ or μ and thus, will need to use $\bar{a}$ and θ_P to update her prior. The optimal incentive rate in period 2 is given by (B.25), 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. Therefore, with $\lambda = 1/2$, we have

$$EV(x) = \tilde{EV}(x) - \frac{\delta}{4}\left[\frac{(1+b) - \rho(1+\kappa)}{1+2\kappa}\right]^2 \sigma^2_{\theta,P}(1-\gamma(x)), \tag{B.26}$$

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^*

$$x^* = \frac{2\omega\bar{\mu}}{\xi(x^*)\left[\bar{\mu}^2 + \sigma_\mu^2 + (1-2\tilde{\alpha})s_\mu^2 + \delta\left(\frac{[(1+b)-\rho(1+\kappa)]\sigma_\mu\gamma(x)}{\rho(1+2\kappa)}\right)^2\right]}. \blacksquare \tag{B.27}$$

References

Acemoglu, Daron and Matthew O. Jackson. 2016. "Social Norms and the Enforcement of Laws." *Journal of European Economic Association* .

Acquisti, Alessandro, Curis Taylor, and Liad Wagman. 2016. "The Economics of Privacy." *Journal of Economic Literature* (forthcoming).

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.

Ali, S Nageeb. 2011. "Learning Self-Control." *Quarterly Journal of Economics* 126 (2):857–893.

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. 2012. "No margin, No mission." *A Field Experiment on Incentives for Pro-Social Tasks, CEPR Discussion Papers* 8834.

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

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

———. 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. 2014. "Norms, Enforcement, and Tax Evasion." *IIES mimeo* .

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

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. 2014. "Muddled Information." Columbia University.

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.

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

Ichino, Andrea and Gerd Muehlheusser. 2008. "How Often Should you Open the Door?: Optimal Monitoring to Screen Heterogeneous Agents." *Journal of Economic Behavior & Organization* 67 (3):820 – 831.

Jacquet, Jennifer. 2015. *Is Shame Necessary? New Uses for an Old Tool.* Pantheon Books, Random House.

Jia, Ruexie and Torsten Persson. 2013. "Individual vs. Social Motives in Identity Choice: Theory and Evidence from China." *IIES and UCSD mimeo* .

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.

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.

Lohmann, Susanne. 1994. "Information Aggregation through Costly Political Action." *American Economic Review* 84 (3):518–30.

Lorentzen, Peter L. 2013. "Regularizing Rioting: Permitting Public Protest in an Authoritarian Regime." *Quarterly Journal of Political Science* 8 (2):127–158.

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 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.