NBER WORKING PAPER SERIES

INCENTIVES AND PROSOCIAL BEHAVIOR

Roland Bénabou
Jean Tirole

Working Paper 11535
http://www.nber.org/papers/w11535

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2005

We thank for useful comments George Akerlof, Roland Fryer, Timur Kuran, Bentley MacLeod, Tom Romer, Armin Falk, participants at various seminars and conferences and three anonymous referees. We are especially indebted to Ian Jewitt for valuable suggestions. Bénabou gratefully acknowledges support from the John Simon Guggenheim Memorial Foundation in 2004 and from the National Science Foundation. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Incentives and Prosocial Behavior
Roland Bénabou and Jean Tirole
NBER Working Paper No. 11535
August 2005
JEL No. D64, D82, H41, Z13

ABSTRACT

We develop a theory of prosocial behavior that combines heterogeneity in individual altruism and greed with concerns for social reputation or self-respect. Rewards or punishments (whether material or image-related) create doubt about the true motive for which good deeds are performed and this "overjustification effect" can induce a partial or even net crowding out of prosocial behavior by extrinsic incentives. We also identify settings that are conducive to multiple social norms and those where disclosing one's generosity may backfire. Finally, we analyze the choice by public and private sponsors of incentive levels, their degree of confidentiality and the publicity given to agents' behavior. Sponsor competition is shown to potentially reduce social welfare.

Roland Bénabou
Department of Economics
Woodrow Wilson School
Princeton University
Princeton, NJ 08544
and NBER
rbenabou@princeton.edu

Jean Tirole
Université des Sciences Sociales
Manufacture des Tabacs
Aile Jean-Jacques Laffont
Accueil MF 404
21, allée de Brienne
F-31000 Toulouse
FRANCE
tirole@cict.fr

People commonly engage in activities that are costly to themselves and mostly benefit others. They volunteer, help strangers, vote, give to political or charitable organizations, donate blood, join rescue squads and sometimes sacrifice their life for strangers. Many experiments and field studies confirm that a significant fraction of people engage in altruistic or reciprocal behaviors. A number of important phenomena and puzzles, however, cannot be explained by the sole presence of individuals with other-regarding preferences.

First, providing rewards and punishments to foster prosocial behavior sometimes has a perverse effect, *reducing* the total contribution provided by agents. Such a crowding-out of "intrinsic motivation" by extrinsic incentives has been observed in a broad variety of social interactions (see Bruno S. Frey (1997) and Frey and Reto Jegen (2001) for surveys). Studying schoolchildren collecting donations for a charitable organization, Uri Gneezy and Aldo Rustichini (2000b) thus found that they collected less money when given performance incentives (see also Frey and Lorenz Götte (1999) on volunteer work supply). These findings are in line with the ideas in Richard Titmuss (1970), who argued that paying blood donors could actually reduce supply. On the punishment side, George A. Akerlof and William T. Dickens (1982), suggested that imposing stiffer penalties could sometimes undermine individuals' "internal justification" for obeying the law. Frey (1997) provided some evidence to that effect with respect to tax compliance and Gneezy and Rustichini (2000a) found that fining parents for picking up their children late from day-care centers resulted in more late arrivals. In experiments on labor contracting, subjects provided less effort when the contract specified fines for inadequate performance than when it did not (Fehr et al. (2001) and Fehr and Gächter (2002)) and they behaved much less generously when the principal had simply removed from their choice set the most selfish options (Armin Falk and Michael Kosfeld (2004)). These findings extend a large literature in psychology documenting how explicit incentives can lead to decreased motivation and unchanged or reduced task performance (see, e.g. Edward Deci (1975), Deci and Richard Ryan (1985)). In studying this class of phenomena, however, one cannot simply assume that rewards and punishments systematically crowd out spontaneous contributions. Indeed, there is also much evidence to support the basic premise of economics that incentives are generally effective, for instance in workplace contexts (e.g., Robert Gibbons (1997), Canice Prendergast (1999) and Edward P. Lazear (2000a,b)). A more discriminating analysis is thus required.

A second set of issues relates to the fact that people commonly perform good deeds and refrain from selfish ones because of social pressure and norms that attach honor to the former and shame to the latter

(e.g., Dan Batson (1998), Richard B. Freeman (1997)). Charitable and non-profit institutions make ample use of donors' desire to demonstrate their generosity and selflessness (or at least the appearance thereof), with displays ranging from lapel pins and T-shirts to plaques in opera houses or hospitals and buildings named after large contributors**.** The presence of a social signalling motive for giving is also evident in the fact that anonymous donations are both extremely rare –typically, less than 1 percent of the total number[2]– and widely considered to be the most admirable. Conversely, boasting of one's generous contributions is often self-defeating. Codes of honor, whose stringency and scope varies considerably across time and societies, are another example of norms enforced largely through feelings of shame (losing face) or glory. To understand these mechanisms it is again important to not posit exogenous social constraints, but rather to model the inferences and market conditions involved in sustaining or inhibiting them.

Finally, as much as people care about the opinion others have of them, they care about their own *self-image*. In the words of Adam Smith (1776), they make moral decisions by assessing their own conduct through the eyes of an "impartial spectator", an "ideal mate within the breast":

> *"We endeavour to examine our own conduct as we imagine any other fair and impartial spectator would examine it. If, upon placing ourselves in his situation, we thoroughly enter into all the passions and motives which influenced it, we approve of it, by sympathy with the approbation of this supposed equitable judge. If otherwise, we enter into his disapprobation, and condemn it."*

In more contemporary terms, psychologists and sociologists describe people's behavior as being influenced by a strong need to maintain conformity between one's actions, or even feelings, and certain values, long-term goals or identities they seek to uphold.[3] Recent empirical studies confirm the importance of such self-image concerns and their contribution to prosocial behavior.[4] A very telling experiment by Jason Dana, Jason

---

[2] See, e.g., the studies reported in Glazer and Konrad (1996, p. 1021). Note that anonymous contributions have the same tax-deduction benefits as nonanonymous ones.

[3] Thus Batson (1998) writes that *"The ability to pat oneself on the back and feeling god about being a kind, caring person, can be a powerful incentive to help";* he also discusses the anticipation of guilt. Daniel Kahneman and Jack Knetsch (1992) find that subjects' stated willingness to pay for alternative public goods is well predicted by independent assessments of the associated "moral satisfaction". Michèle Lamont (2000) documents the importance attached by her interviewees to the presence or absence of the "caring self" not just in others, but also in themselves.

[4] For instance, in a transportation-related survey of about 1,300 individuals, Olof Johansson-Stenman and Peter Martinsson (2003) find that people who are asked which attributes in a car are most important to them systematically put environmental performance near the top and social status near the bottom; but when asked

Kuang, and Roberto Weber (2003) thus reveals that when people are given the opportunity to remain ignorant of how their choices affect others, or of their precise role in the outcome (as with firing squads, which always have one blank bullet), many "altruists" choose not to know and revert to selfish choices.[5]

To examine this broad array of issues, we develop a theory of prosocial behavior that combines heterogeneity in individuals' degrees of altruism and greed with a concern for social reputation or self-respect. The key property of the model is that agents' pro- or anti-social behavior reflects an endogenous and unobservable mix of three motivations: intrinsic, extrinsic, and reputational, which must be inferred from their choices and the context. We obtain four main sets of results.

– *Rewards and punishments.* The presence of extrinsic incentives spoils the reputational value of good deeds, creating doubt about the extent to which they were performed for the incentives rather than for themselves. This is in line with what psychologists term the "overjustification effect" (e.g., Mark R. Lepper et al. (1973)), to which we give here a formal content in terms of a signal-extraction problem.[6] Rewards act like an increase in the noise-to-signal ratio, or even reverse the sign of the signal, and the resulting crowding out of the reputational (or self-image) motivation to contribute can make aggregate supply downward-sloping over a wide range, with possibly a sharp drop at zero.

– *Publicity and disclosure.* The prominence and memorability of contributions strengthen the signaling motive and thus generally encourage prosocial behavior. When individuals are heterogeneous in their image concerns, however, a greater prominence also acts like an increase in the noise-to signal-ratio: good actions become suspected of being motivated by appearances, which limits the effectiveness of policies based on "image rewards" such as praise and shame. The same concern can lead individuals to refrain from overtly disclosing their good deeds and from turning down any rewards that are offered. Sponsors may respond to contributors' desire to appear intrinsically rather than extrinsically motivated by publicly announcing low rewards, but then find it profitable to offer higher ones in private, creating a commitment problem.

about the true preferences of their neighbors or average compatriots, they give dramatically reversed rankings. Interviews with car dealers show intermediate results.

[5] In a related vein, J. Keith Murnighan et al. (2001) find that the fairness of offers in dictator games is significantly decreased when the precision with which offerers can split the cake is decreased, allowing them to construe the outcomes as largely outside their control.

[6] It is also consistent with the informal explanation provided by Frey and Jegen (2001), namely that *"An intrinsically motivated person is deprived of the chance of displaying his or her own interest and involvement in an activity when someone else offers a reward, or orders him/her to do it".*

– *Spillovers and social norms.* The inferences that can be drawn from a person's actions depend on what others choose to do, creating powerful spillovers that allow multiple norms of behavior to emerge as equilibria. More generally, individuals' decisions will be strategic complements or substitutes, depending on whether their reputational concerns are (endogenously) dominated by the avoidance of stigma or the pursuit of distinction. The first case occurs when there are relatively few types with low intrinsic altruism and when valid excuses for not contributing are more rare than events that make participation inevitable, or unusually easy. The second case applies in the reverse circumstances.

– *Welfare and competition.* When setting rewards and publicizing contributions, sponsors will exploit these complementarities or substitutabilities, which respectively increase or decrease the elasticity of the supply curve. Because they do not internalize the reputational spillovers that fall on non-participants or on those who contribute through other sponsors, however, their policies will generally be inefficient. Thus, even a monopoly sponsor may offer rewards and "perks" (preferred seating, meetings with famous performers, valuable social networking opportunities, naming rights to a building, stadium or professorial chair, etc.) that are too generous from the point of view of social welfare, and sponsor competition may further aggravate this inefficiency. The socially optimal incentive scheme, by contrast, subtracts from the standard Pigouvian subsidy for public goods provision a "tax" on reputation-seeking, which, per se, is socially wasteful. In the market for prosocial contributions, finally, a form of holier-than-thou competition can also lead sponsors to offer agents opportunities for reputationally motivated sacrifices that will again reduce social welfare, without any increase in the supply of public goods..

While a number of related themes have been examined in the literature, none of the existing models provides a unified account of this broad range of phenomena. Standard models of public goods provision or altruistic behavior, whether based on a concern for others' welfare, a pure joy of giving, or reciprocity, are not consistent with a (locally) downward-sloping response of prosocial behavior to incentives, nor with people choosing not to know how their actions will affect others and reverting to selfish behavior when such ignorance is feasible. Models of giving as a signal of wealth explain monetary donations but not in-kind prosocial acts such as volunteering, helping, giving blood, etc. (these should instead be avoided, as they signal a low opportunity cost of time), the greater admiration reserved for anonymous contributions, or people's choosing to be modest about their good deeds. Models that postulate a reduced-form crowding out

(or in) of intrinsic motivation by incentives do not really explain its source and miss its dependence on the informational environment, such as the observability of actions and rewards or the distribution of preferences in the population. The same is true for models of social norms that assume complementarities in payoffs.

The papers most closely related to the present one take a signaling approach to social interactions, although none share with it the structure of multidimensional uncertainty that is essential to generating overjustification effects and net crowding out. In Bénabou and Tirole (2003), a potential conflict between extrinsic and intrinsic motivation arises from the fact giving an agent high-powered incentives may convey bad news about the task or his ability. The idea that the principal has private information about these variables applies well to child-rearing, education and empowerment versus monitoring of employees, but not to activities such as contributing to a charitable cause, donating blood, voting, etc., which are our focus here. In B. Douglas Bernheim (1994), individuals take actions designed to signal that their tastes lie close to "the mainstream", leading to conformity in behavior and multiple social norms. When reputation bears on prosocial orientation, however, what is valuable is not to resemble the average but to appear as altruistic as possible. Such is the case in Corneo's (1997) signaling model of union membership, with which our analysis of social norms shares some important insights. On the other hand, Corneo's model does not give rise to crowding out, and while Bernheim does not consider the effects of incentives, the similarly unidimensional structure of his model will also lead to a standard upward-sloping response. Jerker Denrell (1998) shows how the presence of monetary or side benefits in some activity can destroy the separating equilibrium that would otherwise obtain. While this again does not lead to crowding out, a principal may obtain higher profits with a zero reward than with a positive one. Closest to our paper is that of Paul Seabright (2002), where individuals derive from participating in a "civic activity" both a direct benefit that depends on their private type and a reputation that will make them more desirable partners in a later matching market. Under a sorting condition that makes high types care more about reputation, a "payment" discontinuity arises at zero, in that total participation can be greater when no reward is offered than with a small positive one.[7]

[7] Our paper naturally also ties in to the large literature on gifts and donations, such as James Andreoni (1993) Amihai Glazer and Kai A. Konrad (1996), William Harbaugh (1998), Andrea Buraschi and Francesca Cornelli (2002) and Prendergast and Lars A. Stole (2001). Other related papers include Bodner and Prelec (2003) and Bénabou and Tirole (2004a) on self-signaling, Akerlof and Rachel E. Kranton (2000) on identity, Kjell Arne Brekke, Snorre Kverndokk, and Karine Nyborg (2003) on moral motivation, Maarten Janssen and Ewa Mendys-Kamphorst (2004) on rewards and the evolution of social norms, and Wolfgang Pesendorfer (1995) and Laurie Simon Bagwell and Bernheim (1996) on ostentatious consumptions as signaling devices.

The paper is organized as follows. Section I presents the model and an intuitive illustration of the image-spoiling effect of rewards. Section II formally demonstrates the crowding-out phenomenon, as well as a related form of the overjustification effect. Section III deals with social norms and more generally identifies the features of the market that make individual decisions strategic complements or substitutes. Section IV explores issues of confidentiality and disclosure with respect to rewards or actions. Section V examines the setting of incentives by public or private sponsors and the effects of competition on social welfare. Section VI concludes with possible directions for further research. All proofs are gathered in the Appendix.

I. The Model

A. Preferences and information

We study the behavior of agents who choose the extent of their participation in some prosocial activity: contributing to a public good or worthy cause, engaging in a friendly action, refraining from imposing negative externalities on others, etc. Each selects a participation level a from some choice set $A \subset \mathbb{R}$ that can be discrete (voting, blood donation) or continuous (time or money volunteered, fuel efficiency of car purchased). Choosing a entails a utility cost $C(a)$ and yields a monetary or other material reward ya. The incentive rate $y \gtrless 0$ may reflect a proportional subsidy or tax faced by agents in this economy, or the fact that participation requires a monetary contribution. It is set by a principal or "sponsor" and, for now, individuals take it as given.

Denoting by v_a and v_y an agent's intrinsic valuations for contributing to the social good and for money (consumption of market goods), participation at level a yields a direct benefit

$$(v_a + v_y y)\, a - C(a). \tag{1}$$

Each individual's preference type or "identity" $\mathbf{v} \equiv (v_a, v_y) \in \mathbb{R}^2$ is drawn independently from a continuous distribution with density $f(\mathbf{v})$, marginal densities $g(v_a)$ and $h(v_y)$ and mean $(\bar{v}_a, \bar{v}_y)$. Its realization is private information, known to the agent when he acts but not observable by others.

Social signaling. In addition to these direct payoffs, decisions carry reputational costs and benefits,

Our work is also technically related to a recent literature on signals that convey diverging news about different underlying characteristics (Aloisio Pessoa de Araújo et al. (2004), Philipp Sadowski (2004), David Austen-Smith and Roland G. Fryer (2005)).

reflecting the judgements and reactions of others –family, friends, colleagues, employers. The value of reputation can be instrumental (making the agent a more attractive match, as in Denrell (1998), Herbert Gintis et al. (2001) or Seabright (2002)) or purely hedonic (social esteem as a consumption good). For simplicity, we assume that it depends linearly on observers' posterior expectations of the agent's type $\mathbf{v}$, so that the reputational payoff from choosing a, given an incentive rate y is

$$R(a,y) \equiv x\left[\gamma_a E\left(v_a|a,y\right) - \gamma_y E\left(v_y|a,y\right)\right], \quad \text{with } \gamma_a \geq 0 \text{ and } \gamma_y \geq 0. \tag{2}$$

[8]

The signs of γ_a and γ_y reflect the idea that people would like to appear as *prosocial* (public-spirited) and *disinterested* (not greedy), while the factor $x > 0$ measures the visibility or salience of their actions: probability that it will be observed by others, number of people who will hear about it, length of time during which the record will be kept, etc. Defining $\mu_a \equiv x\gamma_a$ and $\mu_y \equiv x\gamma_y$, an agent with preferences $\mathbf{v} \equiv (v_a, v_y)$ and reputational concerns $\boldsymbol{\mu} \equiv \left(\mu_a, \mu_y\right)$ thus solves

$$\max_{a \in A} \left\{\left(v_a + v_y y\right) a - C(a) + \mu_a E\left(v_a|a,y\right) - \mu_y E\left(v_y|a,y\right)\right\}. \tag{3}$$

In the basic version of the model, $\boldsymbol{\mu}$ is taken to be common to all agents and thus public knowledge. In the full version we also allow for unobserved heterogeneity in image-consciousness, with $\boldsymbol{\mu}$ distributed independently of $\mathbf{v}$ according to a density $m(\boldsymbol{\mu})$. Note, finally, that while we shall generally cast the analysis in terms of effortful or time-consuming prosocial actions such as volunteering, voting, etc., it is equally applicable to purely monetary (e.g., charitable) donations.[9]

Self-signaling and identity. The model admits an important reinterpretation in terms of self-image. Suppose that, at the time he makes his decision, the individual engages in a self-assessment or receives

[8] This payoff is defined net of the constant $(1-x)\left(\gamma_a \bar{v}_a - \gamma_y \bar{v}_y\right)$, which corresponds to the case where a remains unobserved. Note that a value of reputation that is a linear functional of the posterior distribution over the agent's type (such as its expectation) avoids building into his preferences either information-aversion (concave functional) or information-loving (convex functional). The more restrictive assumption, which we make for tractability, is that the coefficients in (2) are independent of the agent's type $\mathbf{v}$.

[9] Let a now be the number of dollars contributed by an individual with a known, concave utility over income, represented by the term $-C(a)$. Each dollar generates one unit of public good and entitles the contributor to y units of perks and privileges (meeting with performers, gala events, networking, etc.), a "currency" for which he has utility v_y. This alternative interpretation of (3) is fully consistent with the analysis in Section II. One can also capture the case where instead of perks, the sponsor offers a matching rate y : let $v_y \equiv 1$ and $C(a) = ca$, where c is the cost of providing a unit of public good, so the net cost to the contributor is only $c-y$. This corresponds to the specification of (3) used in most of Sections III-V.

some external signal about his type: "How important is it for me to contribute to the public good? How much do I care about money? What are my real values?" This information, however, may not be perfectly recalled or "accessible" later on –in fact, there will often be strong incentives to remember it in a self-serving way. Actions, by contrast, are much easier to encode and remember than the underlying motives, making it rational to define oneself partly through ones' past choices: "I am the kind of person who behaves in this way". Suppose therefore that the feelings or signal motivating the participation decision are forgotten with some probability proportional to x and that, later on, the agent cares about "what kind of a person he is".[10] If, for simplicity, this utility from self-image is linear in beliefs, with weights γ_a and $_-\gamma_y$ on perceived social orientation and greediness, the model is formally equivalent to the social-signaling one.

Relation to altruism and public goods. An agent's intrinsic motivation to behave prosocially, v_a, can stem from two sources. First, he may care about the overall level of a public good to which his action contributes but that is enjoyed by others as well, such air quality. Let this component of utility be $w_a\left(n\bar{a}/n^{\kappa}\right)$, where $\bar{a}$ represents the average contribution, n the size of the group and $\kappa \geq 0$ the degree of congestion; w_a then measures the intensity of the individual's "pure" altruism.[11] Second, he may experience a "joy of giving" u_a (independent of social- or self-esteem concerns) that makes him value his own contribution to $\bar{a}$ more than someone else's.[12] Combining these "pure" and "impure" forms of altruism (Andreoni (1988)) yields $v_a = u_a + w_a/n^{\kappa}$; in large groups with $\kappa > 0$, the second term vanishes. The simplest interpretation of our model is thus one where there is a unit continuum of agents, so that $v_a = u_a$, but where $\kappa = 1$ so that the average contribution still generates a public good, which individuals value as $w_a\bar{a}$. The model applies equally well to finite groups of any size n and value κ, however. All that matters is that there be heterogeneity

[10] This may reflect a hedonic motive (people enjoy feeling generous or disinterested, e.g. Akerlof and Dickens (1982) or Botond Köszegi (2000)), an instrumental purpose (providing motivation to undertake and persevere in long-term tasks or social relationships, e.g. Juan D. Carrillo and Thomas Mariotti (2000) or Bénabou and Tirole (2002)), or both. The idea that individuals take their actions as diagnostic of their preferences originated in psychology with Daryl J. Bem (1972) and relates closely to cognitive dissonance theory (Leon Festinger and James Carlsmith (1959)). The link between imperfect recall and intertemporal self-signaling is analyzed in Bénabou and Tirole (2004a), while Bodner and Prelec (2003) examine contemporaneous self-signaling in a split-self model.

[11] At the cost of some additional complexity, one could make agents care about social welfare (which is then defined as a fixed point) rather than about the level of the public good per se.

[12] Such would be the effect of feelings of empathy (emphasized by Batson (1998)) or reciprocity. Equivalently, the marginal cost of participation may include an individual component equal to $-u_a$. The term u_a could also arise from agents' following the Kantian imperative to evaluate their actions as if they would lead everyone to make those same choices (Brekke et al. (2003)).

in the intrinsic propensity to contribute or reciprocate, v_a, no matter its source, and that agents value being perceived, or perceiving themselves, as having a high v_a. This (self) esteem benefit, $\mu_a E(v_a|a,y)$, is perhaps what corresponds best to the idea of a "warm glow" of giving: gaining social approval, feeling good about oneself, etc. Finally, note that the action a chosen by agents and giving rise to reputation could be their reaction to someone else's behavior, such as cooperation or defection. The model is thus applicable to reciprocity as well as to unconditional prosocial behavior.

We now turn to the terms in (3) relating to material compensation. That in $v_y y$ requires no explanation, except to note that if the individual believes that his receiving y reduces the resources available to the sponsor for supporting other activities he cares about, it will be attenuated by an "eviction effect".[13] Consider next the potential negative reputation attached to "greed" or money-orientation, $-\mu_y E(v_y|a,y)$. Note first that all the paper's results but one (Proposition 3) obtain with $\mu_y \equiv 0$ as well. It is nonetheless natural to allow for such an effect –"greedy" is no compliment. Someone who has a high valuation for money relative to effort and / or public goods is not a very attractive partner in friendship, marriage, hiring to a position of responsibility, electing to office and other situations where it is difficult to always monitor behavior or write complete contracts. Demonstrating a low marginal utility for money v_y can also be valuable because it signals high wealth, a motive that figures prominently in the literatures on charitable contributions and on conspicuous consumptions (e.g., Glazer and Conrad (1996), Bagwell and Bernheim (1996)).

B. The image-spoiling effect of rewards: basic insights

We begin with an intuitive presentation of some key mechanisms. Consider the first-order condition for an agent's choice of a, assuming a well-behaved decision problem over a continuous choice set. By (3), an individual with type $(\mathbf{v}, \boldsymbol{\mu})$ who faces a price y equates

$$C'(a) = v_a + v_y y + r(a, y; \boldsymbol{\mu}), \tag{4}$$

[13] In experiments on charitable giving (e.g., Gneezy and Rustichini (2000b)), it is typically emphasized to subjects that any rewards will come from an entirely separate research budget and therefore not reduce the amount actually donated. In the real world, the presence and magnitude of an eviction effect will depend on individuals' beliefs about the level at which the budget constraint binds and how they value the alternative uses of funds. Suppose, for instance, that a charity has a fixed budget and will use any funds left over to hire "professionals" who produce τ units of a per dollar, or some other public good of equivalent value. An individuals' valuation of a reward y for his contribution will now be $(v_y - \tau w_a / n^\alpha)y$. This simply amounts to a redefinition of v_y, in a way that contributes to making it negatively correlated with v_a.

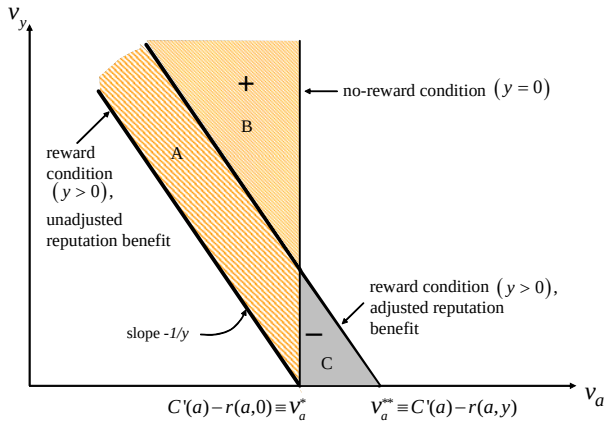

Figure 1: the effects of rewards on the pool of participants

where the last term is his (marginal) reputational return from contributing at level a :

$$r(a, y; \boldsymbol{\mu}) \equiv \mu_a \frac{\partial E\left(v_a | a, y\right)}{\partial a} - \mu_y \frac{\partial E\left(v_y | a, y\right)}{\partial a}. \tag{5}$$

Three important points are apparent from (4). First, observing someone's choice of a reveals the sum of his three motivations to contribute (at the margin): *intrinsic, extrinsic, and reputational.* In general all three vary across people, so that learning about v_a or v_y corresponds to a *signal-extraction* problem. Second, a higher incentive rate y will reduce the informativeness of actions about v_a, and the converse for v_y. Third, heterogeneity in agents' image concerns $\boldsymbol{\mu}$ represents an additional source of noise that makes inferences about both v_a and v_y less reliable, and that is amplified when actions become more visible (higher x).

To gain further insight into the impact of incentives on inferences and behavior, let us now focus on the benchmark case where v_a and v_y are independent random variables, while μ_a and μ_y are fixed and omitted from the notation. Figure 1 then shows, for any $a > 0$, how the set of agents who contribute *at least* a varies with the reward y. This group, which we shall term "high contributors", comprises all agents with

$$v_a + v_y y \geq C'(a) - r(a, y), \tag{6}$$

so its boundary is a straight line corresponding to (4), along which agents choose exactly a. The same condition applies when the participation decision is discrete, $a \in \{0, 1\}$, as will be the case in the second half of the paper, provided we denote $C'(1) \equiv C(1) - C(0)$ and $r(1, y) \equiv R(1, y) - R(0, y)$. Along the boundary, agents are now indifferent between participating and abstaining.

When no reward is offered, $y = 0$, the separating locus is vertical: an agent's contribution reveals nothing

about his v_y, but is very informative about his v_a. In the continuous case prosocial orientation is learned perfectly, in the discrete case one learns whether it is above or below a known cutoff.

When a reward $y > 0$ is introduced, the slope of the separating locus becomes $-1/y < 0$. If we ignore, in a first step, any changes in the inferences embodied in the intercept, the original boundary simply pivots to the left, as shown in Figure 1 (everything works symmetrically for a fine or penalty, $y < 0$). The set of agents contributing at least a thus expands, as types in the hatched area $(A + B)$ are drawn in. Since this occurs at every level of a, the distribution of contributions shifts up (stochastically), resulting in a higher total supply; this is the standard effect of incentives. In equilibrium, however, there are two reputational effects:

a) The new members of the high-contributors' club have lower v_a's than the old ones, so they drag down the group's reputation for prosocial orientation. The reputation of the low-contributors' group also declines, however, so in the discrete-choice case the net effect on the reputational incentive to participate can clearly go either way. Similarly, in the continuous case the reputation $E(v_a|a, y)$ attached to contributing exactly a declines (as that locus pivots to the left), but so does the reputation attached to contributing exactly $a' = a - da$, where da is small; the effect on the marginal return $\partial E(v_a|a, y)/\partial a$ is thus generally ambiguous.

b) The new high contributors are "greedy" types (have a v_y above the mean), whereas those who still contribute below a after the reward is introduced reveal that they care less about money than average. This unambiguously reduces the reputational incentive to participate, as is clear in the discrete case. In the continuous case this follows from the fact that, after the rotation, the locus for contributing at $a - da$ lies below that for contributing a.[14]

If the overall impact of these changes in inferences is negative, $r(a, y) < r(a, 0)$, as drawn in Figure 1, the reward *attracts* some new participants (more greedy agents in area B) to contributing a or more, but *repels* some existing ones (more public-spirited agents in area C). This matches precisely William Upton's (1973) findings that offering a monetary reward for giving blood led to reduced donations by those who had regularly been giving for free and increased donations from those who never had. Overall, the number of agents who contribute at least a may increase or decrease, depending on the weights given to B and C by the distribution $f(\mathbf{v})$. If a net decrease occurs at every a, the distribution of contributions shifts down (stochastically) and

[14] This is due to the fact that $C'(a) - r(a, y)$ is increasing in a, by the second-order condition for (3).

total supply actually declines when a reward $y > 0$ is introduced, starting from a no-reward situation.

II. The overjustification effect and crowding out

We now turn to the formal analysis, establishing three main results. First, we show how the "overjustification effect" discussed by psychologists can be understood as a signal-extraction problem in which *rewards amplify the noise,* leading observers (or a retrospecting individual) to attribute less of role to intrinsic motivation in explaining variations in behavior. We then identify the conditions under which monetary incentives crowd out reputational motivation, resulting in a supply curve that is downward-sloping over a potentially wide range, or exhibits a sharp drop at zero. Finally, we assess the effectiveness of *non-material* rewards such as praise and shame, showing in particular that it is also limited by a form of overjustification effect.

We use here a specification of the model that builds on the familiar normal-learning setup. Let actions vary continuously over $A = \mathbb{R}$, with cost $C(a) = ka^2/2$.[15]

$$\begin{pmatrix} v_a \\ v_y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{v}_a \\ \bar{v}_y \end{pmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ay} \\ \sigma_{ay} & \sigma_y^2 \end{bmatrix} \right), \quad \bar{v}_a \gtrless 0, \quad \bar{v}_y > 0, \tag{7}$$

and at first we continue to focus on the case where everyone has the same reputational concerns, $\boldsymbol{\mu} \equiv (\bar{\mu}_a, \bar{\mu}_y)$. We then extend the analysis to the case where $\boldsymbol{\mu}$ is also normally distributed across individuals.[16]

A. Material rewards

With fixed $\boldsymbol{\mu}$'s, the reputational return (5) is constant across agents and equal to

$$\bar{r}(a, y) \equiv \bar{\mu}_a \frac{\partial E\left(v_a | a, y\right)}{\partial a} - \bar{\mu}_y \frac{\partial E\left(v_y | a, y\right)}{\partial a}. \tag{8}$$

Thus, by (4), an agent's choice of a reveals his $v_a + y v_y$, equal to $C'(a) - \bar{r}(a, y)$. Standard results for normal random variables then yield

[15] The case of a general convex function $C(a)$ is treated in Bénabou and Tirole (2004b). Both here and there, we focus attention on equilibria in which the reputation vector, $E\left(\mathbf{v}|a, y\right)$, is differentiable in a. .

[16] As is often the case, normality yields great tractability at the cost of allowing certain variables to take implausible negative values. By choosing the relevant means large enough, however, one can make the probability of such realizations arbitrarily small; but (7) and (17) below should really be interpreted as local approximations, consistent with the linearity of preferences assumed throughout the paper.

$$E\left(v_a|a,y\right) = \bar{v}_a + \rho(y)\cdot\left(ka - \bar{v}_a - \bar{v}_y\ y - \bar{r}(a,y)\right) \tag{9}$$

$$E\left(v_y|a,y\right) = \bar{v}_y + \chi(y)\cdot\left(ka - \bar{v}_a - \bar{v}_y\ y - \bar{r}(a,y)\right), \tag{10}$$

where

$$\rho(y) \equiv \frac{\sigma_a^2 + y\sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2} \quad \text{and} \quad y\chi(y) \equiv 1 - \rho(y). \tag{11}$$

Intuitively, the posterior assessment of an agent's intrinsic motivation, $E\left(v_a|a,y\right)$, is a weighted average of the prior $\bar{v}_a$ and of the marginal cost of his observed contribution, *net* of the average extrinsic and reputational incentives to contribute at that level.

Finally, substituting (8) into (9)-(10) shows that an equilibrium corresponds to a pair of functions $E\left(v_a|a,y\right)$ and $E\left(v_y|a,y\right)$ that solve a system of two linear differential equations.

Proposition 1 *Let all agents have the same image concern $(\bar{\mu}_a, \bar{\mu}_y)$. There is a unique (differentiable-reputation) equilibrium, in which an agent with preferences (v_a, v_y) contributes at the level*

$$a = \frac{v_a + v_y\ y}{k} + \bar{\mu}_a\rho(y) - \bar{\mu}_y\chi(y), \tag{12}$$

where $\rho(y)$ and $\chi(y)$ are defined by (11). The reputational returns are $\partial E(v_y|a,y)/\partial a = \rho(y)k$ and $\partial E(v_y|a,y)/\partial a = \chi(y)k$, resulting in a net value $\bar{r}(y) = k\left(\bar{\mu}_a\rho(y) - \bar{\mu}_y\chi(y)\right)$.

The effects of extrinsic incentives on inferences and behaviors can now be analyzed. While a higher y increases agents' direct payoff from contributing, $v_a + v_y\ y$, it also tends to reduce the associated signaling value along both dimensions. In the benchmark case of no correlation ($\sigma_{ay} = 0$), for instance,

$$\rho(y) = \frac{1}{1 + y^2\sigma_y^2/\sigma_a^2} \quad \text{and} \quad \chi(y) \equiv \frac{y\sigma_y^2/\sigma_a^2}{1 + y^2\sigma_y^2/\sigma_a^2}, \tag{13}$$

so a higher y acts much like an increase in the *noise-to-signal ratio* $\theta \equiv \sigma_y/\sigma_a$, leading observers who parse out the agent's motives to decrease the weight attributed to social orientation, $\rho(y)$, and increase its counterpart for greediness, $\chi(y)$.[17] When $\sigma_{ay} \neq 0$, a positive correlation tends to amplify the decline in $\rho(y)$, a negative one works to weaken it.[18] Indeed, the more v_a and v_y tend to move together, the less observing

[17] More precisely, $y\chi(y) = 1 - \rho(y)$ rises with y everywhere, but the same is true of $\chi(y)$ only for $|y| \leq 1/\theta$.

[18] For instance, as the correlation between v_a and v_y rises from -1 to 0 to 1, the function $\rho(y)$ pivots downwards over the range $0 < y < 1/\theta$, from $1/(1-\theta y)$ to $1/(1+\theta^2 y^2)$ and then to $1/(1+\theta y)$. The effect of σ_{ay} on the slope $\chi'(y)$ is more complex, as it depends on σ_{ay}^2; the formula is provided in the Appendix.

a high contribution a, or equivalently a high $v_a + v_y y$, represents good news about the agent's intrinsic valuation v_a; and the larger is y, the stronger is this "discounting" effect.

Summing (12) over agents yields the (per capita) aggregate supply of the public good $\bar{a}(y)$, whose slope,

$$\bar{a}'(y) = \frac{\bar{v}_y}{k} + \bar{\mu}_a \rho'(y) - \bar{\mu}_y \chi'(y), \tag{14}$$

reflects both the standard effect of incentives and the crowding out or in of reputational motivation that they induce. Since the general expression (provided in the appendix) is a bit complicated, we focus here on two benchmark cases that make clear the main factors at play. The first one is that of independent values, for which we show that as long as the reputational concern over either prosocial orientation or money-orientation is above some minimum level, there exists a range over which incentives backfire.

Proposition 2 ***(overjustification and crowding out).*** *Let $\sigma_{ay} = 0$ and define $\theta \equiv \sigma_y/\sigma_a$. Incentives are counterproductive, $\bar{a}'(y) < 0$, at all levels such that*

$$\frac{\bar{v}_y}{k} < \bar{\mu}_a \cdot \frac{2y\theta^2}{\left(1 + y^2\theta^2\right)^2} + \bar{\mu}_y \cdot \frac{\theta^2\left(1 - y^2\theta^2\right)}{\left(1 + y^2\theta^2\right)^2}. \tag{15}$$

Consequently, for all $\bar{\mu}_a$ above some threshold $\mu_a^* \geq 0$ there exists a range $[y_1, y_2]$ such that $\bar{a}(y)$ is decreasing on $[y_1, y_2]$ and increasing elsewhere on $\mathbb{R}$. If $\bar{\mu}_y < \bar{v}_y/k\theta$, then $\mu_a^* > 0$ and $0 < y_1 < y_2$; as $\bar{\mu}_a$ increases, y_1 rises and y_2 falls, so $[y_1, y_2]$ widens. If $\bar{\mu}_y > \bar{v}_y/k\theta^2$, then $\mu_a^* = 0$ and $y_1 < 0 < y_2$; as $\bar{\mu}_a$ increases both y_1 and y_2 rise and, for $\bar{\mu}_a$ large enough, $[y_1, y_2]$ again widens.

The role of $\bar{\mu}_a$ is illustrated in Figure 2a. Crowding out can occur over a fairly wide range, making all but very large rewards inferior to none.[19]

The second case we highlight is that of "small rewards", which is interesting for two reasons. First, some studies find crowding out ($\bar{a}(y)$ decreasing) to occur mostly at relatively low levels, and it is sometimes even suggested that the main effect is a *discontinuity at zero* in subjects' response to incentives (Gneezy and Rustichini (2000b), Gneezy (2003)). Is there something qualitatively different between "unrewarded" and "rewarded" activities that could cause rational agents to behave in this way? We show that there is, and explain when it will matter. The second reason why "small rewards" are of interest is that in real-world

[19] The values used in Figure 2a are $\bar{v}_a = 4$, $v_y = 1$, $\mu_y = 0, \theta = .2$ and $\mu_a \in \{0, 6.7, 8.3, 10, 12, 14.6\}$. In Figure 2b they are $\bar{v}_a = 3$, $\bar{v}_y = 1$, $\mu_a = \mu_y = 1$ and $\theta \in \{0, 1, 2, 3, 5\}$.

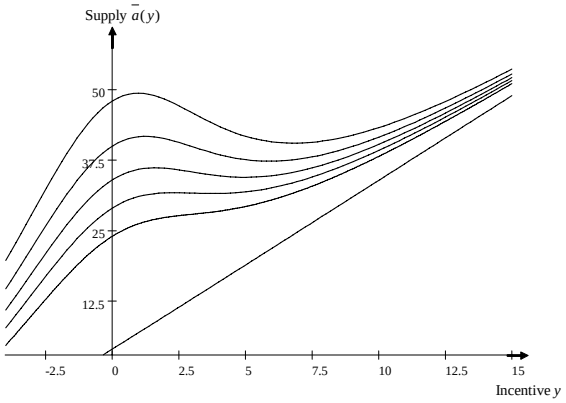

Figure 2a: varying $\bar{\mu}_a$ (with $\bar{\mu}_y = 0$). The straight line corresponds to $\bar{\mu}_a = 0$ (no reputation concern).

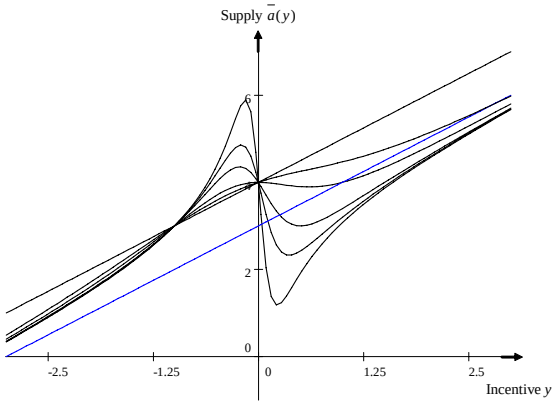

Figure 2b: varying $\theta = \sigma_y/\sigma_a$ (with $\bar{\mu}_a = 0$). The lower straight line corresponds to $\bar{\mu}_y = 0$ (no reputation concern), the upper one to $\theta = 0$ (standard one-dimensional signaling model).

situations where time has an opportunity cost, they will actually correspond to substantial values of y.

Proposition 3 ***(small net incentives and signal-reversal).*** *(1) Small rewards or punishments are counterproductive,* $\bar{a}'(0) < 0$, *whenever*

$$\frac{\bar{v}_y}{k} < \bar{\mu}_a \left(\frac{\sigma_{ay}}{\sigma_a^2} \right) + \bar{\mu}_y \left(\frac{\sigma_y^2 - 2\sigma_{ay}^2/\sigma_a^2}{\sigma_a^2} \right). \tag{16}$$

(2) Let $\bar{\mu}_y > 0$ *and assume that* v_a *and* v_y *are uncorrelated, or more generally not too correlated. Then, as* σ_a/σ_y *becomes small, the slope of the supply function at* $y = 0$ *tends to* $-\infty$.

(3) Suppose that participation entails a unit opportunity cost with monetary value $\hat{y}$. *Then* $\bar{a}'(\hat{y}) < 0$ *and* $\bar{a}'(\hat{y}) \to -\infty$ *under the conditions stated in (1) and (2) respectively.*

The first term on the right-hand side of (16) reflects the intuition given earlier about the role of correlation in generating crowding out -or in. Most important is the second term, whose dependence on the noise-to-signal ratio is illustrated in Figure 2b: letting $\sigma_{ay} = 0$, for instance, shows that $\bar{a}'(0) = \bar{v}_y/k - \bar{\mu}_y(\sigma_y/\sigma_a)^2$. Thus, when individuals' desire for money becomes much more uncertain (to observers) than their motivation for the specific task at hand, and even if they have only a minimal concern about appearing greedy ($\bar{\mu}_y$ is small), *the supply response becomes discontinuous (downward) at zero.* The intuition for why "zero is special" is that, at that point, participation switches from being an "unprofitable" to a "profitable" activity and thus comes to be interpreted as a signal of greed rather than disinterestedness. This *signal reversal* effect, operating specifically around a zero net reward, creates an additional source of crowding out on top

of the general *signal-jamming* effect (decrease in $\rho(y)$) that was shown to operate at all levels of y.[20]

If the empirical validity of this signal reversal was restricted to very small prizes and fines, it would be of somewhat limited interest. The third result, shows, however, that the relevant "tipping point" is not really zero (except in laboratory experiments, where subjects, once there, have no profitable alternative uses of their time) but agents' monetary value of time, which can be quite substantial.

B. Image rewards

Public authorities and private sponsors aiming to foster prosocial behavior make heavy use of both public displays and private mementos conveying honor or shame. Nations award medals and honorific titles, charitable organizations send donors pictures of "their" sponsored child, non-profits give bumper stickers and T-shirts with logos, universities award honorary "degrees" to scholars, etc. Conversely, the ancient practice of the pillory has been updated in the form of televised arrests and publishing the names of parents who are delinquent on child support, or the licence plate numbers of cars photographed in areas known for drug trafficking or prostitution. Peer groups also play an important role by creating a rehearsal mechanism: if acquaintances all contribute to a cause, one is constantly reminded of one's generosity, or lack thereof.[21]

Formally, greater publicity or prominence corresponds to a homothetic increase in (μ_a, μ_y). Our model then confirms the above intuitions, but also delivers important caveats. In particular, when agents are heterogeneous in their reputational concerns, giving greater scrutiny to their behavior may not work that well, as good actions *come to be suspected of being image-motivated.* To analyze these issues we now allow agents' image concerns, like their valuations, to be normally distributed:

$$\begin{pmatrix} \mu_a \\ \mu_y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{\mu}_a \\ \bar{\mu}_y \end{pmatrix}, \begin{bmatrix} \omega_a^2 & \omega_{ay} \\ \omega_{ay} & \omega_y^2 \end{bmatrix} \right), \quad \bar{\mu}_a \geq 0, \quad \bar{\mu}_y \geq 0, \tag{17}$$

with $\mathbf{v}$ and $\boldsymbol{\mu}$ independent. In the first-order condition (4), the reputational return $r(a, y; \boldsymbol{\mu})$ is now also normal and independent of $\mathbf{v}$ (conditionally on a), with mean $\bar{r}(a, y)$ given by (8) and variance

[20] When the two effects are combined it is easy to get supply curves that have a sharp local minimum at $y = 0$, so that neither offering rewards (up to a point) nor requiring sacrifices raises supply.

[21] People indeed volunteer more help in response to a request to do so, especially when it comes from a friend, a colleague or family (Freeman 1997), whose opinion of them they naturally care about more than that of strangers.

$$\Omega(a,y)^2 \equiv \begin{pmatrix} \frac{\partial E(v_a|a,y)}{\partial a} & -\frac{\partial E(v_y|a,y)}{\partial a} \end{pmatrix} \begin{bmatrix} \omega_a^2 & \omega_{ay} \\ \omega_{ay} & \omega_y^2 \end{bmatrix} \begin{pmatrix} \frac{\partial E(v_a|a,y)}{\partial a} \\ -\frac{\partial E(v_y|a,y)}{\partial a} \end{pmatrix}. \tag{18}$$

The signal-extraction formulas (9)-(10) thus remain unchanged, except that the updating coefficients $\rho(y)$ and $\chi(y)$ are respectively replaced by

$$\rho(a,y) \equiv \frac{\sigma_a^2 + y\sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2 + \Omega(a,y)^2} \text{ and } \quad \chi(a,y) \equiv \frac{y\sigma_y^2 + \sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2 + \Omega(a,y)^2}. \tag{19}$$

An equilibrium then corresponds again to a pair of functions $E(v_a|a,y)$ and $E(v_y|a,y)$ that solve the differential equations (9)-(10), but this system is now *nonlinear,* due to the term $\Omega(a,y)^2$ in ρ and χ. We are able to solve it for the intuitive and important class of solutions where Ω is independent of a, so that reputations remain linear in a. We cannot a priori exclude the existence of other, nonlinear, equilibria.

Proposition 4 *(1) A linear-reputation equilibrium corresponds to a fixed-point* $\Omega(y)$*, solution to:*

$$\Omega(y)^2/k^2 \equiv \omega_a^2\ \rho(y)^2 - 2\omega_{ay}\ \rho(y)\chi(y) + \omega_y^2\ \chi(y)^2, \tag{20}$$

where $\rho(y)$ *and* $\chi(y)$ *are given by (*19*) with* $\Omega(a,y) \equiv \Omega(y)$*. The optimal action chosen by an agent with type* $(\mathbf{v}, \boldsymbol{\mu})$ *is then*

$$a = \frac{v_a + y \cdot v_y}{k} + \mu_a\rho(y) - \mu_y\chi(y) \tag{21}$$

and the marginal reputations are $\partial E(v_a|a,y)/\partial a = \rho(y)k$ *and* $\partial E(v_y|a,y)/\partial a = \chi(y)k$*, with a net value of* $r(y;\boldsymbol{\mu}) = (\mu_a\rho(y) - \mu_y\chi(y))k$ *for the agent.*

(2) There always exists such an equilibrium, and if $\omega_{ay} = 0$ *it is unique (in the linear-reputation class).*

A greater variability of image motives, $\Omega(y)^2 = Var\left(r(y;\boldsymbol{\mu})\right)$, makes individuals' behavior a more noisy measure of their true underlying values (v_a, v_y), reducing both $\rho(y)$ and $\chi(y)$. This variance is itself endogenous, however, as agents' reputational calculus takes into account how their collective behavior affects observers' signal-extraction-problem. This is reflected in the fixed-point nature of equation (20).[22]

Proposition 4 also allows us to demonstrate how increased publicity gives rise to an offsetting *overjustification effect.* Let all the reputational weights $\boldsymbol{\mu} = (\mu_a, \mu_y)$ be scaled up by some prominence or memorability

[22] When $\omega_{ay} \neq 0$ there could be multiple equilibria, with different degrees of informativeness. Since the general theme of multiplicity is investigated in Section III.A, we do not pursue it here.

factor, x; the material incentive y remains constant. Aggregate supply is now

$$\bar{a}(y,x) = \frac{\bar{v}_a + y \cdot \bar{v}_y}{k} + x\left(\bar{\mu}_a \rho(y,x) - \bar{\mu}_y \chi(y,x)\right), \tag{22}$$

where the dependence on x indicates that all the covariance terms $\left(\omega_a^2, \omega_{ay}, \omega_y^2\right)$ in the original equation (20), corresponding to $x = 1$, are now multiplied by x^2. A greater visibility of actions (and of any rewards attached to them) thus has two offsetting effects on the reputational incentive to contribute:

a) a direct *amplifying* effect, the sign of which is that of $\mu_a \rho(y,x) - \mu_y \chi(y,x)$ for an individual and $\bar{\mu}_a \rho(y,x) - \bar{\mu}_y \chi(y,x)$ on average. For people who are mostly concerned about appearing socially-minded ($\mu_a \gg \mu_y$) this increases the incentive to act in a prosocial manner, whereas for those most concerned about not appearing greedy ($\mu_y \gg \mu_a$) it has the reverse effect.[23]

b) a *dampening* effect, as reputation becomes less sensitive to the individual's behavior, which observers increasingly ascribe to image concerns. Formally, the "effective noise" $\Omega(y,x)$ increases with x (in any stable equilibrium) and $\rho(y,x)$ and $\chi(y,x)$ consequently tend to decrease with it.

This tradeoff implies that giving increased publicity to pro- or anti-social behavior may be of somewhat limited effectiveness, even when it is relatively cheap to do. Consider for instance the case where μ_y is known ($\omega_y = 0$), possibly equal to zero. As x becomes large (more generally, $xk\omega_a^2 >> 1$), equation (20) yields

$$\rho(y,x) \approx \left(\frac{\sigma_a^2 + y\sigma_{ay}}{k^2 \omega_a^2}\right)^{1/3} x^{-2/3}. \tag{23}$$

The aggregate social benefit from publicity $\bar{\mu}_a x \rho(y,x)$ thus grows only as $x^{1/3}$, implying that it is optimal to provide only a finite level of x even when it has a constant marginal cost, or even a marginal cost that declines slower than $x^{-2/3}$.[24] Policies by parents, teachers, governments and other principals that rely on the "currency " of praise and shame are thus effective up to a point, but eventually self-limiting.

III. Honor, stigma, and social norms

The second main issue we explore is that of social and personal norms. We first show how multiple

[23] We are focussing this discussion, for simplicity, on the "natural" case where ρ and χ are both positive, which occurs as long as σ_{ay} is not too negative; see (19).

[24] On the other hand there cannot be full crowding out, namely $x\rho(y,x)$ actually decreasing with x : otherwise, by (19) and (20) $\rho(y,x)$ would be increasing in x, a contradiction.

standards of "acceptable" behavior can arise from the interplay of honor and shame, then examine what characteristics of the "market", such as the distribution of social preferences, the availability of excuses or the observability of action and inaction, facilitate or impede their emergence.

For the remainder of the paper we focus on the case of a binary participation decision, $A = \{0, 1\}$, in which the notions of honor and stigma are most sharply apparent. Unless otherwise specified (Sections IV.B and IV.C) we also assume that all agents share the same reputational concern $\boldsymbol{\mu} \equiv (\mu_a, \mu_y)$ and the same valuation for money, which we normalize to $v_y \equiv 1$. Their prosocial orientation v_a, by contrast, is distributed on some interval $[v_a^-, v_a^+]$.[25] Indeed, whereas two-dimensional uncertainty is essential to the overjustification and backfiring-incentives effects analyzed earlier, it is not needed for most of the other results. This simplification also removes any potential incentive for agents to "burn money" in order to signal a low v_y.

We again denote $r(y) \equiv R(1, y) - R(0, y)$ and let $c \equiv C(1) - C(0)$. Thus, an agent now participates if $v_a \geq c - y - r(y) \equiv v_a^*(y)$. To determine the equilibrium threshold of altruism let us define, for any candidate cutoff v_a, the conditional means in the upper and lower tails:

$$\mathcal{M}^-(v_a) \equiv E(\tilde{v}_a \,|\tilde{v}_a \leq v_a), \tag{24}$$

$$\mathcal{M}^+(v_a) \equiv E(\tilde{v}_a \,|\tilde{v}_a \geq v_a) \ . \tag{25}$$

The first expression governs the *"honor"* conferred by participation, which is the difference between $\mathcal{M}^+(v_a)$ and the unconditional mean $\bar{v}_a$. The second one governs the *"stigma"* from abstention, which is $\bar{v}_a - \mathcal{M}(v_a)$. Since both are nondecreasing functions, the net reputational gain $\mathcal{M}^+(v_a) - \mathcal{M}^-(v_a)$ and the marginal agent's total *non-monetary* return to contributing,

$$\Psi(v_a) \equiv v_a + \mu_a \left[\mathcal{M}^+(v_a) - \mathcal{M}^-(v_a)\right] \equiv v_a + \Delta(v_a), \tag{26}$$

may increase or decrease with overall participation, $[v_a, v_a^+]$. The slopes of these two functions will play central roles in what follows.[26]

[25] The results generalize to the case where v_a and v_y are independently distributed and reputation bears only on the former ($\mu_y = 0$).

[26] Recall also that, in the discussion of Figure 1, it was argued that the reputation for prosociality of contributors may worsen either more or less than that of non-contributors when the separating locus pivots to the left due to the presence of a reward $y > 0$. Indeed, for any given value of v_y (over which one then

A. Endogenous social norms

What makes a given behavior socially or morally unacceptable is often the very fact that "it is just not done", meaning that only people whose extreme types make them social outliers would not be dissuaded by the intense shame attached to it. In other places or times different norms or codes of honor prevail, and the fact that "everyone does it" allows the very same behavior to be free of all stigma. Examples include choosing surrender over death, not going to church, not voting, divorce, bankruptcy, unemployment, welfare dependency, minor tax evasion, and conspicuous modes of consumption.

We show here that such interdependencies between agents' choices arise *endogenously* through the inferences made from observed behaviors, creating the potential for multiple norms of social responsibility. In particular, no assumption of complementarity in payoffs (e.g., between v_a and the average contribution $\bar{a}$, representing a form of "reciprocity") is required to explain the common finding that individuals contribute more to public goods when they know that others are also giving more.[27]

Proposition 5 *(1) When Ψ is increasing, there is a unique equilibrium, which varies with y as described in Figure 3a.*

(2) When Ψ is decreasing, the equilibrium set varies with y as described in Figure 3b. Thus, for all $y \in (c - \Psi(v_a^-), c - \Psi(v_a^+))$, there are three equilibria: $v_a^* = v_a^-$ (full participation), $v_a^* = v_a^+$ (no participation) and an interior one defined by $\Psi(v_a^*) = c - y$ that is unstable (in the usual tâtonnement sense).

(3) When Ψ is non-monotonic, there exists a range of values of y for which there are at least two stable equilibria, of which one at least is interior.

We provide two examples. When v_a is uniformly distributed on $[0,1]$, $\Psi(v_a) = v_a + \mu_a/2$ so the supply curve is a familiar, upward-sloping one, as in Figure 3a. When v_a has density $g(v_a) = 2v_a$ on $[0,1]$, by contrast, $\Psi(v_a) = v_a + (2\mu_a/3)(1+v_a)^{-1}$ is decreasing for all $\mu_a > 6$, resulting in three equilibria as in Figure 3b. For $\mu_a \in (3/2, 6)$, Ψ is hump-shaped, making the high-participation equilibrium interior.

integrates), these reputations respectively correspond to $\mathcal{M}^+(v_a^* - v_y y)$ and $\mathcal{M}^-(v_a^* - v_y y)$, whose difference may increase or decrease with y depending on the slope of $\mathcal{M}^+ - \mathcal{M}^-$.

[27] For instance, James H. Bryan, and M.A. Test (1967) found that motorists were more likely to stop and help someone with a flat tire, and walkers-by more likely to put money into a Salvation Army kettle, when they had observed earlier someone else (a confederate) doing so a few minutes before. See also Jan Potters, Martin Sefton and Lise Vesterlund (2001) on charities' frequent strategy of publicly announcing "leadership" contributions and the higher yields achieved when donors act sequentially rather than simultaneously.

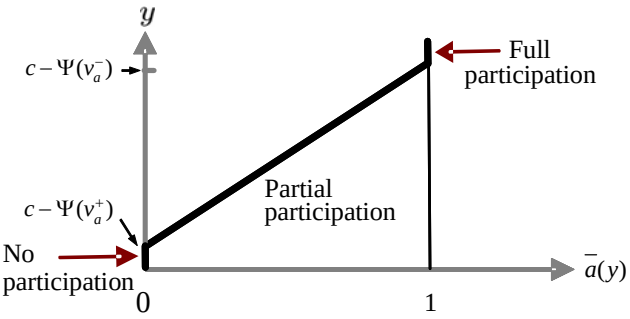

Figure 3a: unique equilbrium

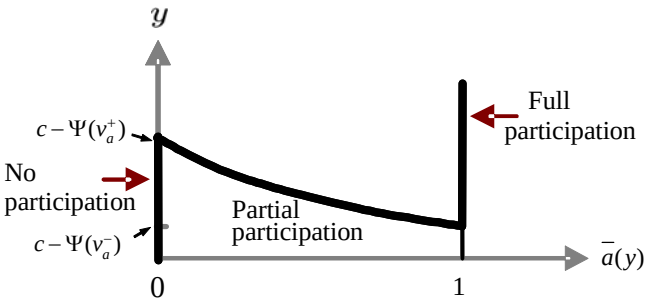

Figure 3b: multiple equilbria

B. Strategic complementarity and substitutability

The intuition for these results is that agents' actions will (endogenously) be strategic complements or substitutes, depending on whether it is *stigma* or *honor* that is *most responsive* to the extent of participation. This same condition will also turn out to play a key role in other results, such as those relating to the disclosure or confidentiality of rewards and the socially optimal level of incentives.

Definition 1 *Participation decisions exhibit strategic complementarities if* $\Delta'(v_a) \equiv \mu_a(\mathcal{M}^+ - \mathcal{M}^-)' < 0$ *for all* v_a.

When $\Delta' < 0$, a wider participation ($dv_a < 0$) worsens the pool of abstainers more than that of contributors, so that the stigma from abstention $\bar{v}_a - \mathcal{M}^-(v_a)$ rises faster than the honor from participation $\mathcal{M}^+(v_a) - \bar{v}_a$ fades. When $\Delta' < -1$, or $\Psi' < 0$, the resulting net increase in reputational pressure is strong enough that the marginal agents in $[v_a^* - dv_a, v_a^*]$, who initially preferred to abstain, now feel compelled to contribute. This further increases participation and confines abstention to an even worse pool, etc., leading to corner solutions as the only stable equilibria, as in Figure 3b. When $\Delta' \in (-1, 0)$, complementarity is weak enough that the marginal agents still prefer to stay out, hence stability obtains. This is a fortiori the case when there is susbtitutability, $\Delta' > 0$.

Equipped with this general intuition, we now investigate the main factors that make strategic complementarity –and thus the existence of multiple social norms– more or less likely.

Distribution of social preferences. One expects that stigma considerations will be dominant when the population includes only a few "bad apples" with very low intrinsic values, which most agents will be eager to differentiate themselves from. Formally, an increasing density $g(v_a)$ makes it more likely that $\mathcal{M}^+ - \mathcal{M}^-$

is declining: a rise in v_a hardly increases $E(v_a|\tilde{v}_a \geq v_a)$ but substantially increases $E(v_a|\tilde{v}_a \leq v_a)$, since the weight reallocated at the margin is small relative to that in the upper tail, but large relative to that in the lower tail. Conversely, honor will dominate when there are only a few heroic or saintly types, whom the mass of more ordinary individuals would like to be identified with.[28]

Proposition 6 *(1) (Jewitt (2004)) If the distribution of v_a has a density that is (a) decreasing, (b) increasing, (c) unimodal, then $(\mathcal{M}^+ - \mathcal{M}^-)(v_a)$ is respectively (a) increasing, (b) decreasing, (c) quasi convex. (2) If the distribution of v_a has a log-concave density (more generally, a log-concave distribution function), then for all $\mu_a \in [0,1]$ the supply function is everywhere upward-sloping.*

The first set of results provide sufficient conditions for the monotonicity of $\mathcal{M}^+ - \mathcal{M}^-$, which defines complementarity or substitutability. What ultimately matters for uniqueness or multiplicity and the slope of the supply curve, on the other hand, is the behavior of $\Psi(v_a) = v_a + \mu_a(\mathcal{M}^+ - \mathcal{M}^-)(v_a)$, for which the strength of reputational concerns, μ_a, is also relevant. The second result thus shows that for $\mu_a \in [0,1]$ uniqueness obtains as long as g does not increase too fast –a much weaker condition than (1b). No simple analogue is available for the case of multiplicity, but it is clear that it corresponds to situations where complementarity obtains and μ_a is high enough (as in the example given earlier).

Excuses, forced participation, and observability. We have so far assumed that observers (other agents, future "self") know for sure that the individual had an opportunity to contribute and whether or not he did. This is often not the case.

Suppose that with probability $\delta \in [0,1]$, an individual faces (unverifiable) circumstances that *preclude* participation: not being informed, having to deal with some emergency, etc. For any potential cutoff v_a, the honor conveyed by participation is unchanged, $\mathcal{M}^P(v_a) = \mathcal{M}^+(v_a)$, while the stigma conveyed by non-participation is lessened, taking the form of a weighted average

$$\mathcal{M}^{NP}(v_a;\delta) = \frac{\delta\bar{v}_a + (1-\delta)G(v_a)\mathcal{M}^-(v_a)}{\delta + (1-\delta)G(v_a)}. \tag{27}$$

The same expressions are easily seen to apply if abstention never gives rise to a signal that the individual

[28] Corneo (1997) provides related insights (but no general result) based on whether the value of reputation is a concave ("conformist") or a convex ("elitist") function of someone's perceived rank (which, by definition, is uniformly distributed) in the distribution of altruism.

contributed, but a contribution may go unnoticed (fail to generate such a signal) with probability δ.

Conversely, suppose that with probability $\delta' \in [0,1]$, an individual is *forced* to contribute, or draws a temporarily low cost c The stigma from abstention is now unchanged, $\mathcal{M}^{NP}(v_a) = \mathcal{M}^{-}(v_a)$, but the distinction conveyed by participation is dulled, and given by

$$\mathcal{M}^{P}\left(v_a;\delta'\right) = \frac{\delta'\bar{v}_a + \left(1-\delta'\right)\left[1-G\left(v_a\right)\right]\mathcal{M}^{+}\left(v_a\right)}{\delta' + \left(1-\delta'\right)\left[1-G\left(v_a\right)\right]}. \tag{28}$$

The same expressions apply if participation always gives rise to a signal suggesting that the individual contributed, but non-participation can go undetected (also lead to such a signal) with probability δ'.

Proposition 7 *1) An increase in the probability of unobserved forced participation facilitates the emergence of strategic complementarities and multiple social norms, whereas an increase in the probability of (unobserved) involuntary non-participation inhibits it.*
2) The same results hold for, respectively, an increase in the probability that abstention may escape detection and for an increase in the probability that a good deed goes unnoticed.

The results of this section have empirical and policy implications. First, for behaviors such as crime, from which most people are deterred by either a strong intrinsic distaste (the density of v_a is increasing) or strong extrinsic constraints (a high δ'), stigma-avoidance will be the dominant reputational concern (by contrast, having no criminal record is not particularly glorious) and actions will be strategic complements, potentially leading to substantial variations over time and space. Conversely, opportunities to engage in heroic behaviors (risking one's life for someone else, donating an organ or significant wealth) are relatively rare (high δ) and few people are intrinsically motivated to such great feats of abnegation. The signaling motive will therefore be dominated here by the pursuit of distinction, making noble acts strategic substitutes and their prevalence much less variable than that of (comparably rare, on average) criminal acts.[29] Second, even absent multiplicity, the two types of behaviors will respond quite differently to public intervention. For crime-like behaviors the effect of rewards and punishments (y) is amplified by the response of social pressure (crowding in), whereas for self-sacrifices it moves in the opposite direction (partial crowding out). We shall

[29] A more general intuition can also be offered. For all distributions that have a standard "bell shape", or a full support on $\mathbb{R}$, a sufficiently high cost of c of behaving prosocially (heroic deeds) will place the cutoff v_a^* in the upper tail, where Δ is decreasing (by Part 1.c of Proposition 6), whereas a sufficiently high cost of behaving antisocially (crime, persistent non-employment) will place it in the lower tail, where Δ increases.

come back to this point when analyzing the socially optimal level of incentives.

IV. Disclosure

Since the presence of material rewards spoils the reputational value of good deeds, it is natural to examine what will occur when sponsors can keep them confidential, or when agents have the opportunity to turn them down. Similarly, given that explicit publicity also leads to a discounting of intrinsic motivation, we will examine the extent to which agents may want to be "modest" about their generosity.

A. Should the fee remain confidential?

We consider here a sponsor (NGO, government agency, religious organization, etc.) that derives from each agent's participation a benefit with equivalent monetary value B, relative to its opportunity cost of funds This could reflect the premium placed on a public good by a particularly motivated constituency (friends of the arts, environmentalists) or some private benefits tied to the delivery of the public good (rents derived by a government agency, bundling of a religious message together with schooling or poverty relief); B could also represent, in reduced form, the sponsor's own signaling or career concerns. We focus here on monopolistic or specialized sponsors, deferring to the next section the analysis of competition. A sponsor's expected payoff from setting a reward rate y is thus equal or proportional to $\pi(y) \equiv n\bar{a}(y)(B-y)$.[30]

We assume that the sponsor can commit to either of two incentive policies: *confidentiality* (C), under which only the agent knows the level of y offered (but participation is publicly observable), or *public disclosure* (D). We maintain the same specification of preferences as above (unknown v_a's and $v_y \equiv 1$) and assume $\Psi' > 0$ to avoid a multiplicity of participation equilibria. We also assume that the sponsor's objective function is quasiconcave in y under both policies.

Confidentiality. The target audience rationally expects a fee and cutoff $\left(y^C, v_a^C\right)$ satisfying $v_a^C - c + y^C + \Delta\left(v_a^C\right) \equiv 0$. If the sponsor secretly deviates and offers y, it thus faces the *ex-post* supply curve

$$\bar{a}_C(y) = 1 - G\left(c - y - \Delta\left(v_a^C\right)\right), \tag{29}$$

[30] In the next section we consider sponsors who also internalize (part or all of) social welfare, maximizing the a weighted sum $\alpha\bar{U} + \pi$, where $\bar{U}$ is agents' average utility and $\alpha \in [0,1]$. One can show that Proposition 8 extends to this case as well, for all $\alpha < 1$. For $\alpha = 1$ the sponsor behaves like a social planner and is indifferent between confidentiality and disclosure.

and chooses y to maximize $\pi_C(y) \equiv \bar{a}_C(y)(B-y)$. The equilibrium fee y^C is then defined by $\pi'_C(y^C) = 0$.

Public disclosure. The difference is that the fee is now credibly announced and therefore affects the reputational value of contributions. For any choice of y, the sponsor thus faces the *ex-ante* supply curve

$$\bar{a}_D(y) = 1 - G\left(c - y - \Delta\left(v_a^*(y)\right)\right) \tag{30}$$

and chooses y to maximize $\pi_D(y) \equiv \bar{a}_D(y)(B-y)$. The equilibrium fee y^D is then defined by $\pi'_D(y^D) = 0$.

Proposition 8 *(1) It is optimal for the sponsor to publicly disclose and commit to the fee.*
(2) With strategic complements $(\Delta' < 0)$, *the sponsor offers a higher fee and elicits a higher participation under disclosure than under public confidentiality. The reverse holds for strategic substitutes* $(\Delta' > 0)$.
(3) The optimal reward under disclosure y^D *is immune to secret renegotiation between agents and sponsor when* $\Delta' < 0$. *By contrast, when* $\Delta' > 0$, *the equilibrium reward when secret renegotiation is feasible is* y^C.

Under public disclosure (but not confidentiality), strategic complementarity creates a "bandwagon effect" that *raises the slope of the supply curve* and therefore makes announcing higher fees profitable. Ex-post, the sponsor would like to lower the fee to y^C but participants would not agree, so the announced price is renegociation-proof. Strategic substitutability has the converse effect on supply, leading to $y^D < y^C$. In this case, the sponsor and participants would agree to secretly increase the reward ex-post; anticipating this collusive renegotiation, the audience properly expects that the actual fee will be y^C and not y^D.

B. Turning down rewards

An agent may be eager to participate but concerned that his image will be tainted by an inference that money played a role in the decision. So even when the sponsor offers y, the agent could turn down part or all of the reward (assuming $y > 0$), or even complement his participation (such as giving blood) with a net monetary contribution. Is this possibility damaging to our results?

Note first that the issue does not arise if give-backs are not observable by the audience to whom agents are trying to signal, or if the sponsor can reward them secretly. On the other hand, taking secret rewards does not help with self-image and may even damage it.

Suppose now that the realized transfer from the sponsor to the agent is effectively observed. When the uncertainty is about v_a, the net reputational gain from participating for $y' \leq y$, relative to not participating,

is $r(y') = \mu_a \left(E\left(v_a|1, y'\right) - E\left(v_a|0, y'\right)\right)$. The agent therefore cannot signal his type by turning down any part of the reward, or even giving money to the sponsor: the loss of monetary income, $v_y\left(y - y'\right)$, and the net reputational benefit, $r(y') - r(y)$, are both type-independent.

Proposition 9 *Let $v_y \equiv 1$, while v_a is unknown. The equilibria studied in Sections* III *and* IV.A *are still equilibria of the enlarged game in which the individual can turn down part or all of the reward. For the same reason, offering menus of rewards cannot benefit the sponsor.*[31]

By contrast, when the uncertainty is (also) about v_y, which is needed to obtain net crowding-out, turning down the reward or part of it could be used to signal the absence of greed. Yet even in this case it may be that all agents either just accept y or do not participate, but *never turn down* rewards. The intuition is that doing so could lead the audience to question an agent's motivation along another dimension: is he genuinely disinterested, or merely concerned about his social (or self) image? It is thus linked to the general idea that good deeds that are "too obvious" may backfire, which was first encountered when studying public prominence in Section II.B and will recur again when examining private disclosure.

To capture this idea, we allow again uncertainty about $\mathbf{v} = (v_a, v_y)$ to combine with uncertainty about agents' degree of image-consciousness $\boldsymbol{\mu} = (\mu_a, \mu_y)$ but focus here on a very simple case, to avoid what would otherwise be a rather technical analysis. Suppose that $(\mu_a, \mu_y) = \tilde{x}(\gamma_a, \gamma_y)$, where (γ_a, γ_y) is fixed and thus known to the audience, whereas $\tilde{x}$ is independently distributed from (v_a, v_y) and takes one of two extreme values: agents are either *image indifferent* ($\tilde{x} = 0$) or *image driven* ($\tilde{x} = +\infty$). Image-indifferent individuals participate if and only if $v_a - c + v_y y \geq 0$; when they do, they clearly never turn down the reward (or part of it), as this would be a strictly dominated strategy. We assume that if the population consisted only of image-indifferent individuals, participation would yield a better reputation than non-participation (this always holds for y below some threshold). Turning now to image-driven individuals, they all pool on the actions that yield the highest reputation, choosing an $a \in \{0, 1\}$ and a reward $y' \leq y$ that maximize $R(a, y') = \mu_a E\left(v_a|a, y'\right) - \mu_y E\left(v_y|a, y'\right)$. If, in equilibrium, a positive fraction of them chose to participate and receive $y' < y$, they would be identified as image-driven types, and so their reputation would correspond

[31] It can also be verified that these equilibria satisfy the Never-a-Weak-Best-Response criterion of In-Koo Cho and David M. Kreps (1987).

to the prior mean $(\bar{v}_a, \bar{v}_y)$.[32] But they would then be strictly better off pooling with those image-indifferent agents who participate at price y. The unique equilibrium thus consists in participation, at the offered price y, by all image-driven individuals and by those image-indifferent individuals for whom $v_a - c + v_y y \geq 0$.

Proposition 10 *Agents may never turn down the reward, or part of it, even when this would be publicly observed and there is uncertainty about* v_y.

It is worth pointing out that in deriving this result, we did not assume any social opprobrium on image-consciousness; presumably, this would only reinforce agents' reluctance to turn down rewards.[33]

C. Conspicuous versus anonymous generosity

People often react with disapproval when someone tries to buy social prestige by revealing how generous, disinterested, well-thinking, etc., they are. Conversely, the most admired contributions and sacrifices are anonymous ones. To analyze this phenomenon, let us assume that if an agent participates others will normally learn of it only with probability $x < 1$. He can, however, make sure that they find out by verifiably *disclosing* his action, by incurring a time or resources cost d. Agents differ again both in their valuation v_a for the public good (whereas $v_y \equiv 1$, for simplicity) and in their concern for image γ_a, with the two characteristics being independent.

In the *symmetric information* case where each agent's γ_a is observable, one can show that there exist thresholds $0 < \gamma_a^* < \gamma_a^{**}$ such that agents with $\gamma_a < \gamma_a^*$ never disclose, those with $\gamma_a > \gamma_a^{**}$ always do and for $\gamma_a \in [\gamma_a^*, \gamma_a^{**}]$ there are multiple norms: both disclosure and non-disclosure are equilibrium behaviors, because the absence of information about an agent's contribution carries a lower stigma if comparable others do not disclose than if they do.[34]

We assume here instead that γ_a, like v_a, is unobservable and show that this can reduce disclosure, which now *itself* carries a stigma, even though there is no social opprobrium on image-consciousness per se. The

[32] If they pooled at multiple values y', all these would need to deliver the same average reputation, which would therefore again correspond to the prior mean.

[33] The result also implies that offering menus of rewards along which agents with different v_y's could sort themselves, which is optimal when $\boldsymbol{\mu}$ is known (see Bénabou and Tirole (2004b) for an analysis), may still not benefit sponsors when people also differ in their in image-consciousness.

[34] Thus, if most people belong to some church, synagogue or mosque, anyone who does not risks being seen as a selfish materialist, since "doing good" through other channels is less easily demonstrable. The symmetric-information case is omitted here for brevity but can be found in Bénabou and Tirole (2004b).

idea is that since the people most prone to advertise their good deeds are those with a high concern for image, disclosure of a prosocial act makes it more likely that it was motivated by image-seeking (a high γ_a) rather than genuine altruism (a high v_a). Formally, suppose that γ_a takes values γ_a^H for a fraction θ of agents and $\gamma_a^L \leq \gamma_a^H$ for the remaining $1-\theta$, such that, under symmetric information, it is an equilibrium for type γ_a^H to disclose and type γ_a^L not to do so. We show the following results.

Proposition 11 *Under asymmetric information about the extent of image-consciousness* γ_a :
(1) In a separating equilibrium where the γ_a^H *types disclose while the* γ_a^L *ones do not, disclosure of one's contribution to the public good carries a stigma, in that the inferences about the individual's prosocial orientation are not as favorable as when participation is revealed through other channels:* $v_a^H < v_a^L$.
(2) Asymmetric information about the extent of image-consciousness can reduce disclosure: for some range of values of d, the γ_a^H *type no longer discloses when* γ_a *is unobservable.*
3) Suppose that $\gamma_a^L = 0$ *(so type* γ_a^L *never discloses) and that disclosure (by the agent or the sponsor) is relatively cheap. Then, starting from either a separating equilibrium as in (1), or one with pooling at no-disclosure as in (2), the sponsor gains from a policy under which contributions are systematically disclosed.*

A commitment to automatic disclosure acts as a remedy to asymmetric information about γ_a, as it relieves agents from the suspicion of image-seeking attached to discretionary disclosure. This leaves only the direct effect of an increased visibility of actions, which raises the reputational incentive to participate and thereby increases the sponsors' payoff at any given level of y.

V. Welfare and Competition

We now examine the way in which public or private sponsors will set incentives and the welfare properties of the resulting equilibrium. For these purposes, we need to make explicit again the public-good aspects of agents' contributions. Recall from Section I.A that an individual's intrinsic motivation can, in general, have two components: $v_a = u_a + w_a/n^\kappa$, where u_a is a pure "joy of giving" whereas w_a is the marginal utility of a public good $n\bar{a}/n^\kappa$ generated by total contributions $n\bar{a}$. To simplify the analysis, we take here u_a and w_a to be independently distributed (with again $v_y \equiv 1$) and denote the mean of w_a as $\bar{w}_a$.

Given an incentive rate y, an equilibrium (unique or not) is determined by a cutoff v_a^*. Agents' (expected)

average welfare is thus

$$\begin{aligned}(31)\qquad \bar{U}(v_a^*;y) &\equiv E\left[w_a\left(n\bar{a}/n^\kappa\right)\right]+E\left[a\left(u_a-c+y\right)+\mu_a v_a\right]\\ &= \int_{v_a^*}^{v_a^+}\left[\left(n-1\right)\left(\bar{w}_a/n^\kappa\right)+v_a-c+y\right]g\left(v_a\right)dv_a+\mu_a\overline{v}_a.\end{aligned}$$

This expression embodies three effects. First, each agent who contributes enjoys a direct utility $v_a - c + y$ and additionally generates for the $n-1$ others a positive spillover, equal to $\bar{w}_a/n^\kappa$ on average. Second, the pursuit of esteem is a zero-sum game: the average reputation in society remains fixed at $\mu_a\overline{v}_a$, reflecting the martingale property of beliefs.[35] Third, because an agent's participation decision is based on the private reputational return rather the social one (which is zero), it inflicts an externality onto others. Thus, starting from equilibrium, the welfare impact of a marginal increase in participation is

$$(32)\qquad -\frac{\partial\bar{U}(v_a^*;y)}{\partial v_a^*}=\left[\left(n-1\right)\left(\bar{w}_a/n^\kappa\right)+v_a^*-c+y\right]g\left(v_a\right)=\left[\left(n-1\right)\left(\bar{w}_a/n^\kappa\right)-\Delta(v_a^*)\right]g\left(v_a\right).$$

The first term is the standard public-goods externality, which we shall denote as $\bar{e} \equiv (n-1)(\bar{w}_a/n^\kappa)$. The second term reflects the fact that each marginal participant brings down the "quality" of the pool of contributors as well as that of non-contributors: by the martingale property, the reputational losses of inframarginal agents on both sides must add up to the gains of the marginal participant, $\Delta(v_a^*)$. Equivalently, we can think of (32) as the difference between a *free-riding effect* and a *reputation-stealing* effect.

A. Sponsors' choice of incentives and the social optimum

Let B again denote the *private* monetary value of the benefit that participation by an agent confers to a sponsor. In addition, the sponsor could also internalize part or all of agents' welfare. The general form of a monopolistic sponsor's payoff is thus

$$(33)\qquad \bar{W}(y)\equiv\alpha\bar{U}(v_a^*(y);y)+\pi(y),$$

where $\pi(y) = n\bar{a}(y)(B-y)$ and $\alpha \in [0,1]$. For a social planner whose preferences mirror the ex-ante utility of the n potential contributors and who has access to lump-sum taxes, $\alpha = 1$ and $B = 0$. More

[35] That is, $E\left[E\left[v_a|a,y\right]\right] = \bar{v}_a$. It thus does not matter whether or not we include agents' utilities from reputation (e.g., vanity) in the definition of social welfare. Note that the zero-sum property also relies on the linearity of the reputational payoff and the independence of μ_a from v_a. When these assumptions do not hold, the distribution of reputation across agents will have allocative and efficiency consequences – for instance, through subsequent matching patterns.

generally, $B \geq 0$ could reflect a different discounting of the welfare of future generations (e.g., with pollution or biodiversity) and $\alpha \leq 1$ the presence of a shadow cost of public funds: clearly, replacing $B - y$ by $B - (1+\lambda)y$ in $\pi(y)$ is equivalent to dividing both B and α in (33) by $1+\lambda$. For other actors such as NGO's or specialized government agencies, B reflects the purely private benefits (material or reputational) that the sponsor derives from contributions transiting through it and α the weight it places on social welfare, both normalized by the sponsors' own opportunity cost of funds.[36]

Since rewards that lead to net crowding out, $\bar{a}'(y) < 0$, are never optimal, we assume that $\Psi' > 0$, resulting in a unique equilibrium $v_a^*(y)$ and supply curve $n\bar{a}(y) = n\left[1 - G(v_a^*(y))\right]$, with elasticity $\varepsilon(y) \equiv y\bar{a}'(y)/\bar{a}(y) > 0$. We also assume that $\bar{W}$ is strictly quasiconcave in all cases (it always is for $\alpha = 1$). Using (32) and noting that $\bar{a}'(y) = -\left(v_a^*\right)'(y) \cdot g(v_a^*(y))$, we have

$$\bar{W}'(y) = \left[\alpha\left(\bar{e} - \Delta(v_a^*(y))\right) + B - y\right] \cdot \bar{a}'(y) - (1-\alpha)\bar{a}(y). \tag{34}$$

For (symmetric) competitive sponsors, the term $\pi(y)$ in (33) is replaced by $\pi_i(y) \equiv n\bar{a}_i(y)\left(B - y_i\right)$, where $\bar{a}_i(y)$ is the share of total supply specifically channeled to sponsor i; in equilibrium, all rewards are driven to B.[37] We shall denote the values of $\alpha, B, \bar{W}$ and y for the social planner, monopolistic and competitive sponsors by the superscripts s, m and c respectively, with $\alpha^s > \max\left\{\alpha^m, \alpha^c\right\}$.

Proposition 12 *1) The socially optimal incentive rate,*

$$y^s = \frac{\alpha^s\left[\bar{e} - \Delta(v_a^*(y^s))\right] + B^s}{1 + (1-\alpha^s)/\varepsilon(y^s)}, \tag{35}$$

is strictly less than the standard Pigouvian subsidy $y^P \equiv \bar{e} + B^s$ that leads agents to internalize the full public-good value of their contribution, even when taxation is non-distortionary ($\alpha^s = 1$).

2) A monopoly sponsor with $\alpha^m < \alpha^s$ may offer contributors a reward y^m that is too generous (or, require of them too low a monetary donation) from the point of view of social welfare, resulting in excess participation. This is true even when the benefits it derives from agents' participation coincide exactly with the gap between

[36] It is worth recalling here that the model also applies to charitable monetary donations; see footnote 9.

[37] While this is the standard result, it depends here crucially on the fact that $v_y = 1$ is known. Otherwise, there is a reputational payoff to participating for a lower fee and sponsor competition will then lead to rewards being bid *down* rather than up, leaving firms with positive profits. This "reversal" of Bertrand competition is analyzed in Bénabou and Tirole (2004b) and shares important similarities with Bagwell and Bernheim's (1996) analysis of the pricing of conspicuous-consumption goods.

their social and private contributions to the public good ($B^m + \alpha^m \bar{e} = B^s + \alpha^s \bar{e}$).

3) Competition between sponsors increases rewards (or, reduces required monetary contributions) and may thus reduce social welfare, compared to a monopoly (with the same $\alpha^c = \alpha^m$ *and* $B^c = B^m$*).*

The first result shows, most transparently when $\alpha^s = 1$, that the optimal incentive scheme should include a tax that corrects for the reputation-seeking motive to contribute, which in itself is socially wasteful. This reputational rent is endogenous to the reward, however. Thus with $\alpha^s = 1$, when individual contributions are complements (resp., substitutes) y^s responds less (resp., more) than y^P to changes in B^s (which leave the function Δ unchanged). Similarly, the optimal penalty for antisocial activities such as littering, polluting, etc., should "leave space" for the effect of opprobrium, which itself depends on the fine. As to a higher shadow cost of public funds (a proportional reduction in α^s and B^s), it naturally tends to reduce y^s; when contributions are substitutes, some of this reduced public intervention is made up by increased social pressure, as Δ rises in response to the decline in participation. With complements, however, the reputational incentive to contribute is also weakened. These results provide both some support and an important qualification to arguments (e.g. Brennan and Pettit (2004)) calling for a shift in public policy from the use of fines and other costly sentences to a greater reliance on public praise and shame. Esteem-based incentives can adequately replace material rewards and punishments in spheres where gaining distinction is the dominant reputational concern (self-sacrifice, heroism, great inventions), but not in those where avoiding stigma is most important (crime, welfare dependency). This point, in turn, suggests that scarce public funds should be allocated much more to fostering prosocial behaviors (and discouraging antisocial ones) of the latter kind than of the former.

The intuition for the second result in Proposition 12 is that a monopolist setting y^m does not not internalize the reputational losses of inframarginal agents to the same extent as a planner would. This gives it an incentive to attract too many "customers", which works against the standard monopolistic tendency to serve too few. The tension between these two forces can be seen from the fact that $\left(\bar{W}^s\right)'(y^m) < 0$ if

$$(\alpha^s - \alpha^m)\left[y^m/\varepsilon(y^m) - \Delta(v_a^*(y^m))\right] + B^s + \alpha^s \bar{e} - B^m - \alpha^m \bar{e} < 0. \tag{36}$$

A low supply elasticity ε causes the monopolist to offer too low a price, as usual. When reputational concerns are important enough, however (a high μ_a and therefore a high Δ), the informational externality can dominate, making the monopolist too "generous" or not demanding enough in the standards it sets for

monetary donations. The last two terms in (36), finally, represent the the total benefits (private benefit plus internalized contribution to social welfare) derived by each sponsor from a marginal agent's participation, each normalized by the corresponding shadow cost of funds. The effect of their difference on the sign of $y^m - y^s$ is straightforward, and Part (2) of the proposition normalizes it to zero as a benchmark.

Sponsor competition, finally, further exacerbates the above inefficiency, because each firm now has a much higher incentive to raise its offer than a monopolist (it takes the whole market), but still inflicts the same reputational cost on all inframarginal non-contributors. This suggests, for instance, that universities may sell the naming rights to professorial chairs and buildings too cheaply, relative to the social optimum.

Quality of participation. Sponsors often care about "high-quality" participation, not just total enrollment. This arises when participation is an open-ended contract, subject to adverse selection or moral hazard. Thus, one argument for relatively low pay for the military is to select true patriots rather than people whose main loyalty is to money (e.g., mercenaries who may find out one day that the enemy pays better). Similarly, it is often argued that not paying for blood reduces the fraction of donors with hepatitis and other diseases. These ideas can be captured by introducing a hidden action (beyond $a \in A$, which is observed) whose marginal cost to the individual decreases with v_a, leading to a benefit for the sponsor $B(v_a)$, with $B' > 0$. The theory is then the same, with for example a private sponsor ($\alpha = 0$) now maximizing $\pi(y) \equiv E_{\mathbf{v},\boldsymbol{\mu}}\left[\left(B(v_a) - y\right) a(\mathbf{v}, \boldsymbol{\mu};\ y)\right]$.

B. Holier-than-thou competition

We saw that competition may reduce welfare by inducing excessive participation in prosocial activities that generate only moderate public-good benefits but have a high visibility. We will now see that it can reduce welfare (relative to a monopolist) even without any change in participation, by leading sponsors to screen contributors in inefficient ways. This result formalizes in particular the idea of religions and sects competing on orthodoxy, asceticism and other costly requirements for membership (e.g., Eli Berman (2000)). Another example of rapidly growing importance is that of charities sponsoring events where agents, instead of simply donating or raising money (or on top of it), engage time-intensive, strenuous activities such as a day-long walk, marathon or other test of endurance.

To capture this phenomenon most simply, let v_a take values v_a^H with probability ρ or $v_a^L < v_a^H$ with

probability $1-\rho$, while maintaining $v_y \equiv 1$. Assume, furthermore, that the non-monetary cost of contributing is c (possibly zero) unless the sponsor demands a "sacrifice", which it is able to verify and publicly certify. The cost then becomes c^H for the high type and c_L for the low type, where

$$c < c^H < c^L. \tag{37}$$

A sacrifice is a pure deadweight loss, whose only benefit is to help screen agents' motivation. The assumption that $c^L > c^H$ reflects the idea that such a sacrifice is less costly to a more motivated agent. For simplicity, we will assume that c^L is so large that the low type is never willing to sacrifice and will focus on deterministic contracts offered by sponsors who seeking to maximize their private payoff $\pi(y)$; that is, we set $\alpha = 0$ (the results would extend to any $\alpha < 1$).

Proposition 13 *In the two-type case described above, a monopoly sponsor who wants both types to contribute does not screen contributors inefficiently. By contrast, competing sponsors may require high-valuation individuals to make costly sacrifices that represent pure deadweight losses, thereby reducing social welfare.*

The intuition for this result is that non-price screening imposes a negative externality on low-type agents, the cost of which a monopolist must fully bear but which competitive sponsors do not internalize. Indeed, screening through costly sacrifices has two effects: a) it inflicts a deadweight loss $c^H - c$ on the high type, which the sponsor must somehow pay for; b) it boosts the high type's reputation and lowers that of the low type. When the high-type's reputational gain exceeds the cost of sacrifice, the sponsor through which he contributes can appropriate the surplus, in the form of a lower reward. If this sponsor is a monopolist who finds it profitable to serve the whole market (which is always the case when ρ is low enough), he must also compensate the low type for his reputational loss. By a now familiar argument, these losses must exactly offset the high type's reputation gains, so the net effect of (b) on agents' average utility, as well as on the monopolist's payoff, is nil. This leaves only the net cost corresponding to (a), implying that a sponsor serving the whole market will never require sacrifices.

Things are quite different under free entry. First, since v_y is known, price competition again drives all sponsors to offer B. Second, by requiring a sacrifice, entrants can now attract the high types away from competitors who impose no such requirement, leaving low-types (or their sponsors) with the resulting reputational loss. This "cream-skimming" leads inevitably to an equilibrium where a proportion ρ of the

contracts offered by active sponsors require an inefficient sacrifice and attract only high-types, while the remaining $1-\rho$ require only the normal contribution c and attract the low types.[38]

Turning finally to welfare, one can show that both types of agents are better off under competition than under monopoly (see the appendix). The sponsors or their underlying beneficiaries, however, must necessarily lose more than all contributors gain: total participation remains unchanged (both types still contribute), the same is true of average reputation (by the martingale property), and rewards are pure transfers. There is now, however, a deadweight loss of $\rho(c^H - c)$, corresponding to the wasteful sacrifices made by the high-types to separate. Therefore, competition unambiguously reduces welfare.

VI. Conclusion

To gain a better understanding of prosocial behavior we sought, paraphrasing Adam Smith, to "*thoroughly enter into all the passions and motives which influence it*". People's actions indeed reflect a variable mix of altruistic motivation, material self-interest and social or self image concerns. Moreover, this mix varies across individuals and situations, presenting observers seeking to infer a person's true values from his behavior (or an individual judging himself in retrospect) with a signal-extraction problem. Crucially, altering any of the three components of motivation, for instance through the use of extrinsic incentives or a greater publicity given to actions, *changes the meaning* attached to prosocial (or antisocial) behavior and hence feeds back onto the reputational incentive to engage in it.

This simple mechanism lead to many new insights concerning individuals' contributions to public goods as well as the strategic decisions of public or private sponsors seeking to increase or capture these contributions. This line of research could be extended in several interesting directions. A first one concerns organizations, where high-powered incentives or performance pay could potentially conflict with agents' signaling motives that arise from teamwork or career concerns. A second relates to the role and objectives of sponsors, who in practice often have their own signaling concerns. A third one, linked to the self-image interpretation of the model, is to the topic of identity and the many instances where people refuse transactions that seem to be in their best economic interest, but which they judge to be insulting to their dignity.

[38] As long as ρ is not too large, this is the only equilibrium that is robust to the Cho-Kreps (1987) criterion.

Appendix

Proof of Proposition 1: Since y is simply a fixed parameter, in what follows we will temporarily omit from the notation the dependence of all functions on this argument. Differentiating (9)-(10) with respect to a yields

$$\frac{dE\left(v_a|a\right)}{da} = \rho\left[k - \bar{r}(a)\right] \quad \text{and} \quad \frac{dE\left(v_y|a\right)}{da} = \chi\left[k - \bar{r}(a)\right]. \tag{A.1}$$

Therefore, $\bar{r}(a)$ is a solution to the linear differential equation $\bar{r}(a) = \mu\left(k - \bar{r}'(a)\right)$, where $\mu \equiv \bar{\mu}_a\rho - \bar{\mu}_y\chi \gtrless 0$. The generic solution is $\bar{r}(a) = k\left(\mu + \zeta e^{-a/\mu}\right)$, where ζ is a constant of integration. Equation(4) yields $a^*(\mathbf{v}) = \left(v_a + y \cdot v_y\right)/k + \mu + \kappa e^{-a/\mu}$ and substituting into (9)-(10), we obtain:

$$E\left(v_a|a^*(\mathbf{v})\right) = \bar{v}_a + \rho \cdot \left(v_a - \bar{v}_a - y \cdot \left(v_y - \bar{v}_y\right) - \kappa e^{-a/\mu}k\right) \tag{A.2}$$

$$E\left(v_y|a^*(\mathbf{v})\right) = \bar{v}_y + \chi \cdot \left(v_a - \bar{v}_a - y \cdot \left(v_y - \bar{v}_y\right) - \kappa e^{-a/\mu}k\right) \tag{A.3}$$

Applying the law of iterated expectations over $\mathbf{v}$ shows that $\kappa = 0$, concluding the proof. ■

Proof of Propositions 2 and 3: From (11), we have

$$\rho'(y) = -\frac{2y\sigma_a^2\sigma_y^2 + \sigma_{ay}\left(\sigma_a^2 + y^2\sigma_y^2\right)}{\left(\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2\right)^2}, \tag{A.4}$$

$$\chi'(y) = \frac{\sigma_y^2\left(\sigma_a^2 - y^2\sigma_y^2\right) - 2\sigma_{ay}\left(y\sigma_y^2 + \sigma_{ay}\right)}{\left(\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2\right)^2}. \tag{A.5}$$

Substituting into (14) immediately yields Part (1) of Proposition 3 in the case $y = 0$, and Part (1) of Proposition 2 when $\sigma_{ay} = 0$. This last inequality can be rewritten as

$$Q(y) = \left(\bar{v}_y/k\right)\left(1 + y^2\theta^2\right)^2 + \bar{\mu}_y\theta^4y^2 < 2\bar{\mu}_a\theta^2y + \bar{\mu}_y\theta^2 \equiv L(y). \tag{A.6}$$

The left hand side is a second order polynomial in y^2, hence convex and symmetric over all of $\mathbb{R}$, with value $Q(0) = \bar{v}_y/k > 0$ at the origin. The right-hand side is an increasing linear function with $L(0) = \bar{\mu}_y\theta^2$. Consequently, if $L(0) \geq Q(0)$, then for any $\bar{\mu}_a > 0$, $L(y)$ intersects $Q(y)$ once on at some $y_1 < 0$ and once at some $y_2 > 0$. If $L(0) < Q(0)$, on the other hand, then there exists a unique $\mu_a^* > 0$ for which $L(y)$ has a (single) tangency point $y^* > 0$ with $Q(y)$. For all $\bar{\mu}_a < \mu_a^*$, $\dot{Q(y)} > L(y)$ on all of $\mathbb{R}^*$, so $\bar{a}'(y) > 0$ everywhere. For all $\bar{\mu}_a > \mu_a^*$, however, $L(y)$ intersects $Q(y)$ twice, at points $0 < y_1 < y_2$. These properties, together with the linearity of L in $\bar{\mu}_a y$ and the convexity of $Q(y)$, conclude the proof of Proposition 2.

Part (2) of Proposition 3 follow from the fact that, given Part (1), as $\theta = \sigma_y/\sigma_a \to +\infty$ the dominant term in $\bar{a}'(0)$ is asymptotically equivalent to $-\bar{\mu}_y\theta^2\left[1-2\left(\sigma_{ay}/\sigma_a\sigma_y\right)^2\right]$, which tends to $-\infty$ as long as the correlation between v_a and v_y is less than $1/\sqrt{2}$ in absolute value. ■

Proof of Proposition 4: The only difference with Proposition 1 is the presence of the term $\Omega(y)^2 = k^2\, Var[r(y;\boldsymbol{\mu})]$ in the denominator of ρ and χ (see (19)), leading to the fixed-point equation defining $\Omega(y)$:

$$\Omega^2 = k^2 Var\left[\mu_a\left(\frac{\sigma_a^2+y\sigma_{ay}}{\sigma_a^2+2y\sigma_{ay}+y^2\sigma_y^2+\Omega^2}\right)-\mu_y\left(\frac{y\sigma_y^2+\sigma_{ay}}{\sigma_a^2+2y\sigma_{ay}+y^2\sigma_y^2+\Omega^2}\right)\right] \equiv Z(\Omega^2). \tag{A.7}$$

Since $Z(\Omega^2)$ is always positive but tends to zero as Ω^2 becomes large, there is always at least one solution. When $\omega_{ay}=0$, moreover, $Z(\Omega^2)$ is the sum of two squared terms that are decreasing in Ω^2, so the solution is unique. When $\omega_{ay}\neq 0$, one cannot rule out multiple equilibria; note, however, that those that are stable (in a standard, tâtonnement sense) are those where Z cuts the diagonal from above. Therefore, in any stable equilibrium Ω is increasing in k, which in turn implies that $\rho(y)$ and $\chi(y)$ are decreasing in k, as long as σ_{ay} is not too negative. Finally, multiplying all the $\left(\mu_a,\mu_y\right)$'s by a common "publicity factor" x has the same effect on (A.7) as multiplying k^2 by x, which concludes the proof. ■

Proof of Proposition 6: Part (1) is due to Jewitt (2004). For Part (2), we can write:

$$v_a+\mu_a\left[\mathcal{M}^+(v_a)-\mathcal{M}^-(v_a)\right] = v_a-\mathcal{M}^-(v_a)+\mu_a\mathcal{M}^+(v_a)+(1-\mu_a)\mathcal{M}^-(v_a),$$

then observe that both $\mathcal{M}^+$and $\mathcal{M}^-$ are increasing functions, and so is $v_a-\mathcal{M}^-(v_a) = \left(\int_{-\infty}^{v_a} G(v)\,dv\right)/G(v_a)$ if the integral of G is log-concave. Since log-concavity is preserved by integration over convex sets, it suffices that G itself be log-concave. In turn, a sufficient condition for this is that g be log-concave. ■

Proof of Proposition 7: To show (1), rewrite $\left(\mathcal{M}^P-\mathcal{M}^{NP}\right)(v_a;\delta) = \left[\mathcal{M}^+(v_a)-\bar{v}_a\right]/\left[1-(1-\delta)(1-G(v_a)\right]$ and observe that if $(\mathcal{M}^P-\mathcal{M}^{NP})'(v_a;\delta)>0$, this expression is also positive for all $\delta'>\delta$, since

$$\frac{1}{\left(\mathcal{M}^P-\mathcal{M}^{NP}\right)(v_a;\delta')} = \frac{1}{\left(\mathcal{M}^P-\mathcal{M}^{NP}\right)(v_a;\delta)}+\frac{\left(\delta'-\delta\right)(1-G(v_a))}{\mathcal{M}^+(v_a)-\bar{v}_a}$$

and the last term is clearly decreasing in v_a. Similarly, to show (2) note that in this case $\left(\mathcal{M}^P-\mathcal{M}^{NP}\right)(v_a;\delta) = \left[\bar{v}_a-\mathcal{M}^-(v_a)/\right]\left[1-(1-\delta)G(v_a)\right]$ and that if $\left(\mathcal{M}^P-\mathcal{M}^{NP}\right)'(v_a;\delta)<0$, it is also negative for all $\delta'>\delta$.■

Proof of Proposition 8: (1) A sponsor with the ability to credibly commit to the terms of the contract he offers can always replicate the equilibrium choice of one without commitment. Since we show below that he chooses a different fee (as long as $\Delta' \neq 0$), he must in fact do strictly better.

(2) and (3). With disclosure, if the sponsor still chooses $y = y^C$ the reservation value and level of supply that result remain the same as in the confidentiality equilibrium: since $v_a^C - c + y^C + \Delta\left(v_a^C\right) \equiv 0$ by definition, $v_a^*(y^C) = v_a^C$ and therefore $\bar{a}_D(y^C) = \bar{a}_C(y^C)$. The elasticity (or slope) of supply at y^C is different, however:

$$\bar{a}'_C(y^C) = g\left(c - y^C - \Delta\left(v_a^C\right)\right), \text{ whereas} \tag{A.8}$$

$$\begin{aligned} \bar{a}'_D(y^C) &= g\left(c - y^C - \Delta\left(v_a^C\right)\right)\left[1 + \Delta'\left(v_a^*(y)\right) v_a^{*\prime}(y\right] \\ &= g\left(c - y^C - \Delta\left(v_a^C\right)\right)\left[1 + \Delta'\left(v_a^*(y)\right)\right]^{-1}, \end{aligned} \tag{A.9}$$

by the definition of $v_a^*(y)$. Therefore, if $\Delta' < 0$ we have $\bar{a}'_D(y^C) > \bar{a}'_C(y^C)$, hence

$$\begin{aligned} \pi'_D(y^C) &= \bar{a}'_D(y^C)(B - y^C) - \bar{a}_D(y^C) = \bar{a}'_D(y^C)(B - y^C) - \bar{a}_C(y^C) \\ &> \bar{a}'_C(y^C)(B - y^C) - \bar{a}_C(y^C) = \pi'_C(y^C) \equiv 0. \end{aligned}$$

Since π_D is assumed to be quasiconcave, this implies $y^D > y^C$ and therefore also $\bar{a}_D(y^D) > \bar{a}_D(y^C) = \bar{a}_C(y^C)$. The same reasoning works in reverse when $\Delta' > 0$. Part (3), finally, was proved in the text. ■

Proof of Proposition 11: We shall assume all supply curves to be uniquely defined and upward sloping ($\Psi' > 0$, for the relevant Ψ).

(1) Let $\hat{v}_a^L$ and $\hat{v}_a^H$ be the valuation cutoffs used under *symmetric information* by types γ_a^L and γ_a^H respectively (in the equilibrium under consideration). In particular, $\widehat{v}_a^H$ must satisfy

$$\widehat{v}_a^H + \gamma_a^H\left[\mathcal{M}^+\left(\widehat{v}_a^H\right) - \mathcal{M}^-\left(\widehat{v}_a^H\right)\right] = c + d - y, \tag{A.10}$$

$$\gamma_a^H(1 - x)\left[\mathcal{M}^+(\widehat{v}_a^H) - \mathcal{M}^-(\widehat{v}_a^H)\right] \geq d. \tag{A.11}$$

For $\hat{v}_a^L$, in the first equation γ_a^H is replaced by $x\gamma_a^L$ and d by zero. In the second one, γ_a^H is replaced by γ_a^L and the inequality is reversed. Consider now a separating equilibrium under asymmetric information, in which types γ_a^H and γ_a^L participate when v_a is above the cutoffs v_a^H and v_a^L respectively. The posterior expectations of v_a, conditioned respectively on disclosure and on the information that the individual participated but did not disclose, are $E\left(v_a \,|\, D; v_a^H\right) = \mathcal{M}^+(v_a^H)$ and $E\left(v_a \,\middle|\, N; v_a^L\right) = \mathcal{M}^+(v_a^L)$, while

$$E\left(v_a\left|\phi, v_a^H, v_a^L\right.\right) \equiv \frac{\theta \int_0^{v_a^H} v\, g(v)\, dv + (1-\theta)\left[\int_0^{v_a^L} v\, g(v)\, dv + (1-x)\int_{v_a^L}^{\infty} v g(v) dv\right]}{\theta G(v_a^H) + (1-\theta)[G(v_a^L) + (1-x)(1-G(v_a^L)]} \tag{A.12}$$

is the updated reputation in the absence of information. Thus v_a^H and v_a^L are defined by

$$v_a^H + \gamma_a^H\left[\mathcal{M}^+(v_a^H) - E\left(v_a\left|\phi, v_a^H, v_a^L\right.\right)\right] = c + d - y, \tag{A.13}$$

$$v_a^L + \gamma_a^L x\left[\mathcal{M}^+(v_a^L) - E\left(v_a\left|\phi, v_a^H, v_a^L\right.\right)\right] = c - y, \tag{A.14}$$

where the Ψ-type functions in (A.13) and (A.14) are assumed to be increasing. These two inequalities, together with the image-conscious type γ_a^H's willingness to disclose,

$$\gamma_a^H\left[\mathcal{M}^+(v_a^H) - x\mathcal{M}^+(v_a^L) - (1-x)\, E\left(v_a\left|\phi, v_a^H, v_a^L\right.\right)\right] \geq d, \tag{A.15}$$

imply that $v_a^H < v_a^L$.

(2) We demonstrate the claim by way of an example: suppose that $x = 0$ (generally, x is not too large). Then (A.15) and (A.10) respectively reduce to:

$$\begin{aligned}
\Delta_D^H(v_a^H) &\equiv \gamma_a^H\left(\frac{\theta G(v_a^H)\left[\mathcal{M}^+(v_a^H) - \mathcal{M}^-(v_a^H)\right] + (1-\theta)[\mathcal{M}^+(v_a^H) - \bar{v}_a]}{\theta G(v_a^H) + (1-\theta)}\right) \geq d, \\
\widehat{\Delta}_D^H(\widehat{v}_a^H) &\equiv \gamma_a^H\left[\mathcal{M}^+(\widehat{v}_a^H) - \mathcal{M}^-(\widehat{v}_a^H)\right] \geq d,
\end{aligned}$$

Note that $\widehat{\Delta}_D^H(v_a) > \Delta_D^H(v_a)$ for all v_a. Assuming that $1 + \left(\widehat{\Delta}_D^H\right)' > 0$ and using the fact that $v_a^H + \Delta_D^H(v_a^H)$ and $\widehat{v}_a^H + \widehat{\Delta}_D^H(\widehat{v}_a^H)$ both equal $= c + d - y$, we obtain $\widehat{v}_a^H < v_a^H$, hence $\Delta_D^H(v_a^H) < \widehat{\Delta}_D^H(\widehat{v}_a^H)$. Hence, for $\Delta_D^H(v_a^H) < d < \widehat{\Delta}_D^H(\widehat{v}_a^H)$, disclosure by γ_a^H types no longer occurs under asymmetric information about γ_a.

(3) Let x_L and $x_H \in \{x, 1\}$ denote the two types' "visibility" parameters. When $\gamma_a^L = 0$, v_a^L does not vary with x_L and/or x_H. In a pooling equilibrium $x_L = x_H = x$, so a systematic disclosure policy ($x_L = x_H = 1$) alters neither $E\left(v_a\left|1, v_a^H, v_a^L\right.\right)$ nor $E\left(v_a\left|\phi, v_a^H, v_a^L\right.\right)$. Thus, type γ_a^H's reputational incentive is multiplied by $1/x$ and v_a^H decreases. In a separating equilibrium, $x_L = x$ and $x_H = 1$. Keeping $x_H = 1$, $E\left(v_a\left|1, v_a^H, v_a^L\right.\right)$ increases with x_L, while $E\left(v_a\left|\phi, v_a^H, v_a^L\right.\right)$ decreases with it. Hence v_a^H decreases when x_L increases. ■

Proof of Proposition 12: Part (1) follows from (34) and the assumed strict quasiconcavity of $\bar{W}^s$. The additional properties stated in the text for the case $\alpha^s = 1$ follow from the fact that we then have $\left(\bar{W}^s\right)'(y) = [B^s + \bar{e} - \varphi(y)] \cdot \bar{a}'(y)$, where $\varphi(y) \equiv y + \Delta(v_a^*(y))$ is such that

(A.16) $$\varphi'(y) = 1 - \frac{\Delta'(v_a^*(y))}{\Psi'(v_a^*(y))} = \frac{1}{\Psi'(v_a^*(y))} = \frac{1}{1+\Delta'(v_a^*(y))} > 0.$$

and $\varphi(-\infty) = -\infty = -\varphi(+\infty)$. Therefore, $\bar{W}(y)$ is strictly concave and maximized at the point where $y^s = B^s + \bar{e} - \Delta(v_a^*(y^s))$, which is such that $dy^s/dB \gtrless 1$ as $\Delta'(v_a^*(y^s)) \lesseqgtr 0$.

For Part (2), note that (36) holds for all $B^m + \alpha^m \bar{e} \leq B^s + \alpha^s \bar{e}$ as long as $\Delta(v_a^*(y^m)) > \bar{a}(y)/\bar{a}'(y)$, or

(A.17) $$\frac{\Delta(v_a^*)}{\Psi'(v_a^*)}\left(\frac{g(v_a^*)}{1-G(v_a^*)}\right) > 1,$$

where $v_a^{*\cdot}$ stands for $v_a^*(y^m)$. For instance, for $\alpha^m = 0$ and v_a uniformly distributed on $U[0,1]$, we have $v_a^*(y^m) = (c - \mu_a/2 + 1 - B)/2 \in (0,1)$ and $y^m = (B - 1 + c - \mu_a/2)/2$ as long as $-\mu_a/2 < 1 + B - c < 2 - \mu_a/2$. Thus $y^m > y^s = B + \bar{e} - \mu_a/2$ whenever $\mu_a > 1 + B - c + 2\bar{e}$, which is consistent with the previous inequalities as long as $\mu_a > 2\bar{e}$. Part (3), finally, is implied by Part (2), since $y^c = B^c = B^m > y^m$ and $\bar{W}^s$ is declining to the right of y^s. ■

Proof of Proposition 13: (1) As long as ρ is not too small, it is optimal for the monopolist to get both types on board. If he does not demand any sacrifice, he sets y so as to make the low type indifferent:: $y = c - v_a^L - \mu_a\left(\bar{v}_a - v_a^L\right)$, where $\bar{v}_a \equiv \rho v_a^H + (1-\rho) v_a^L$ is the prior mean. The sponsor's payoff is then:

(A.18) $$\pi_1 \equiv B - y = B - c + v_a^L + \mu_a\left(\bar{v}_a - v_a^L\right).$$

Suppose now that the high type is asked to sacrifice. Rewards are then $y^L = c - v_a^L$ and (from incentive compatibility) $y^H = y^L + c^H - c - \mu_a\left(v_a^H - v_a^L\right)$. The sponsor's payoff is then only

(A.19) $$\pi_2 = B - \rho y^H - (1-\rho) y^L = \pi_1 - \rho\left(c^H - c\right) < \pi_1.$$

(2) Under free entry all sponsors offer, and all contributors accept, $y = B$. Moreover, if $c^H - c \leq \mu_a\left(v_a^H - v_a^L\right)$, it is now an equilibrium for the high type to separate from the low type by opting for a sponsor who requires a sacrifice. In the resulting equilibrium (described in the text), both types of agents are better off than under monopoly: the low type's payoff rises from $\mu_a v_a^L$ to $\mu_a v_a^L + v_a^L - c + B$, while the high type's payoff increases by at least $v_a^L - c + B$, which is positive from the condition that the monopoly prefers to enlist both types. The fact that sponsors must necessarily lose more than the agents gain, resulting in a net welfare loss from competition, was established in the text. ■

References

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

Akerlof, George A. and Dickens, William T. "The Economic Consequences of Cognitive Dissonance." *American Economic Review,* June 1982, 72(3), pp. 307–319.

—— and Kranton, Rachel E. "Economics and Identity." *Quarterly Journal of Economics*, August 2000, 115(3), pp. 716–753.

Araújo, Aloisio Pessoa de; Gottlieb, Daniel and Moreira, Humberto. "A Model of Mixed Signals with Applications to Countersignaling and the GED." Mimeo, Getulio Vargas Foundation, 2004.

Austen-Smith, David and Fryer, Roland G. "An Economic Analysis of "Acting White"," *Quarterly Journal of Economics,* May (2005), 120(2), pp. 551-583.

Bagwell, Laurie Simon and Bernheim, Douglas B. "Veblen Effects in a Theory of Conspicuous Consumption." *American Economic Review,* June 1996 86(3), pp. 349-373.

Batson, Dan. "Altruism and Prosocial Behavior," in D. Gilbert, S. Fiske, and G. Lindzey eds., *Handbook of Social Psychology* vol. II. McGraw Hill, 1998, pp. 282–316.

Bénabou, Roland and Tirole, Jean. "Self Confidence and Personal Motivation." *Quarterly Journal of Economics*, August 2002, 117(3), pp. 871–915.

—— . "Intrinsic and Extrinsic Motivation." *Review of Economic Studies,* 2003, 70(3), pp. 489-520.

—— ."Willpower and Personal Rules." *Journal of Political Economy,* August 2004a, 112(4),pp. 848-887.

——. "Incentives and Prosocial Behavior." CEPR Discussion Paper 4633, September 2004b.

Bem, Daryl. J. "Self-Perception Theory."in L. Berkowitz , ed., *Advances in Experimental Social Psychology*, Vol. 6, New York: Academic Press,1972, pp. 1-62.

Berman, Eli. "Sect, Subsidy and Sacrifice: An Economist's View of Ultra-Orthodox Jews." *Quarterly Journal of Economics, Summer* 2000, pp. 905-953.

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

Bodner, Ronit and Prelec, Drazen. "Self-Signaling and Diagnostic Utility in Everyday Decision Making," in I. Brocas and J. Carrillo eds. *The Psychology of Economic Decisions.* Vol. 1: *Rationality and Well-Being,*

Oxford University Press, 2003, pp.105-126.

Brekke, Kjell Arne; Kverndokk, Snorre and Nyborg, Karine. "An Economic Model of Moral Motivation." *Journal of Public Economics,* September 2003, 87(9-10), pp. 1967-83.

Brennan, Geoffrey and Pettit, Philipp. The Economy of Esteem. Oxford, U.K.: Oxford University Press, 2004.

Bryan, James H. and Test, M.A. "Models and Helping: Naturalistic Studies in Aiding Behavior." *Journal of Personality and Social Psychology,* 1967, 6, pp. 400-407.

Buraschi, Andrea and Cornelli, Francesca. "Donations." CEPR Discussion Paper No. 3488, August, 2002.

Carrillo, Juan D. and Mariotti, Thomas. "Strategic Ignorance as a Self Disciplining Device." *Review of Economic Studies*, Spring 2000, 67(3): 529–544.

Cho, In-Koo and Kreps, David M. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics*, May 1987, 102(2), pp. 179–221.

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

Dana, Jason; Kuang, Jason and Weber, Roberto. "Exploiting Moral Wriggle Room: Behavior Inconsistent with a Preference for Fair Outcomes." Carnegie Mellon Behavioral Decision Research Working Paper No. 349, June 2003.

Deci, Edward. *Intrinsic Motivation.* New York: Plenum, 1975.

—— and Ryan, Richard. *Intrinsic Motivation and Self-Determination in Human Behavior.* New York: Plenum Press, 1985.

Denrell, Jerker. *Essays on the Economic Effects of Vanity and Career Concerns.* Stockholm, Sweden: Stockholm Institute of International Economics Press, 1998.

Falk, Armin and Kosfeld, Michael. "Distrust - The Hidden Cost of Control." IZA Discussion Paper 1203, July 2004.

Fehr, Ernst and Simon Gächter. "Do Incentive Contracts Undermine Voluntary Cooperation?" Institute for Empirical Research in Economics (Zurich University) Working Paper No. 34, 2002.

——; Klein, Alexander and Schmidt, Klaus. "Fairness, Incentives, and Contractual Incompleteness." Mimeo, University of Munich, 2001.

Festinger, Leon and Carlsmith, James. "Cognitive Consequences of Forced Compliance." *Journal of Abnormal and Social Psychology,* 1959 58, pp. 203–210.

Freeman, Richard B. "Working for Nothing: the Supply of Volunteer Labor." *Journal of Labor Economics*, January 1997, 15(1), pp. 140–166.

Frey, Bruno S. *Not Just for the Money: An Economic Theory of Personal Motivation.* Edward Elgar, Cheltenham,1997.

—— and Götte, Lorenz. "Does Pay Motivate Volunteers?" Institute for Empirical Economic Research, University of Zurich, Working Paper 7, 1999.

—— and Jegen, Reto. "Motivation Crowding Theory: A Survey of Empirical Evidence." *Journal of Economic Surveys*, December 2001, 15(5), pp. 589– 611.

Gibbons, Robert. "Incentives and Careers in Organizations," in: David Kreps and Ken Wallis, eds., *Advances in Economic Theory and Econometrics*, vol.2. Cambridge University Press, 1997.

Gintis, Herbert; Smith, Eric and Bowles, Samuel. "Costly Signaling and Cooperation." *Journal of Theoretical Biology*, 2001, 213, pp.103–119.

Glazer, Amihai and Konrad, Kai A. "A Signaling Explanation of Charity." *American Economic Review,* September 1996, 86(4), pp.1019-28.

Gneezy, Uri ."The W effect of Rewards." Mimeo, University of Chicago, 2003.

—— and Rustichini, Aldo. "A Fine is a Price." *Journal of Legal Studies*, January 2000a, 29(1), pp. 1-17.

——. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics*, August 2000b, 115(3), pp. 791-810.

Harbaugh, William. "What Do Donations Buy? A Model of Philanthropy Based on Prestige and Warm Glow." *Journal of Public Economics,* February 1998, 67(2), 169-284.

Jewitt, Ian. "Notes on the Shape of Distributions." Mimeo, Oxford University, Summer 2004.

Janssen, Maarten C.W. and Mendys-Kamphorst, Ewa. "The Price of a Price: On the Crowding Out of Social Norms." *Journal of Economic Behavior and Organization* 2004, 55:3, pp.377-395.

Johansson-Stenman, Olof and Martinsson, Peter. "Honestly, Why Are You Driving a BMW?" Mimeo, Göteborg University, 2003. Forthcoming in the *Journal of Economic Behavior and Organization.*

Kahneman, Daniel and Knetsch, Jack. "Valuing Public Goods: the Purchase of Moral Satisfaction." *Journal*

of Environmental Economics and Management, January 1992, 22(1), pp. 57-70.

Köszegi, B. (2004) "Utility from Anticipation and Personal Equilibrium," Mimeo, U.C. Berkeley , June.

Lamont, Michèle. *The Dignity of Working Men.* New York: Russel Sage Foundation Press, 2000.

Lazear, Edward. "Performance Pay and Productivity." *American Economic Review*, December 2000a, 90(5), pp.1346–1361.

—. "Personnel Economics and Economic Approaches to Incentives." *HKCER Letters*, September/October 2000b, 61.

Lepper, Mark; Greene, D. and Nisbett, Richard. "Undermining Children's Interest with Extrinsic Rewards: A Test of the 'Overjustification Hypothesis'." *Journal of Personality and Social Psychology,* 1973, 28, pp. 129– 137.

Murnighan, J. Keith; Oesch, John M. and Pillutla, Madan. "Player Types and Self-Impression Management in Dictatorship Games: Two Experiments." *Games and Economic Behavior,* November 2001, 37(2), pp. 388-394.

Pesendorfer, Wolfgang. "Design Innovation and Fashion Cycles." *American Economic Review*, September 1995, 85(4), pp. 771–792.

Potters, Jan; Sefton, Martin and Vesterlund, Lise. "Why Announce Leadership Contributions? An Experimental Study of the Signaling and Reciprocity Hypotheses." Mimeo, University of Pittsburgh, 2001.

Prendergast, Canice. "The Provision of Incentives in Firms." *Journal of Economic Literature*, March 1999, 37(1), pp. 7–63.

—— and Stole, Lars. "The Non-Monetary Nature of Gifts." *European Economic Review*, December 2001, 45(10), pp. 1793–1811.

Sadowski, Philipp. "Overeagerness."Mimeo, Princeton University, 2004.

Seabright, Paul. "Continuous Preferences Can Cause Discontinuous Choices: an Application to the Impact of Incentives on Altruism." Mimeo, IDEI, Toulouse, 2002.

Smith, Adam. *The Theory of Moral Sentiments. New York:* Prometheus Books, 1759

Titmuss, Richard. *The Gift Relationship.* London: Allen and Unwin, 1970.

Upton, William. "Altruism, Attribution and Intrinsic Motivation in the Recruitment of Blood Donors," in *Selected Readings in Donor Motivation and Recruitment*, Vol. II, American Red Cross ed., 1973.