Incentives and Prosocial Behavior

Roland Bénabou [1] and Jean Tirole [2]

First draft: May 16, 2003

This version: August 19, 2004 [3]

[1] Princeton University, CEPR and NBER.

[2] IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, and MIT.

[3] Earlier versions of this paper were presented in the Scribner lectures (Princeton University, April 2002), the ISNIE Congress (Boston, September 2002), the CEU Summer Workshop in Behavioral Economics (Budapest, July 2004) and various seminars. We are grateful to Tom Romer, Armin Falk and especially Ian Jewitt for useful comments. Bénabou gratefully acknowledges support from the John Simon Guggenheim Memorial Foundation.

Abstract

We build a theory of prosocial behavior that combines heterogeneity in individual altruism and greed with concerns for social reputation or self-respect. The presence of rewards or punishments creates doubt as to the true motive for which good deeds are performed, and this "overjustification effect" can result in a net crowding out of prosocial behavior by extrinsic incentives. The model also allows us to identify settings that are conducive to multiple social norms of behavior, and those where disclosing one's generosity may backfire. Finally, we analyze the equilibrium contracts offered by sponsors, including the level and confidentiality or publicity of incentives. Sponsor competition may cause rewards to bid down rather than up, and can even reduce social welfare by requiring agents to engage in inefficient sacrifices.

Keywords: altruism, rewards, motivation, overjustification effect, crowding out, identity, social norms.

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

Introduction

People frequently engage in activities that are costly to themselves and mostly benefit others. They vote, volunteer time, help strangers, give to political or charitable organizations, donate blood, join rescue squads and sometimes even risk or sacrifice their life for strangers. Many experiments and field studies confirm that a significant fraction of individuals engage in altruistic or reciprocal behaviors (e.g., Fehr and Gächter (2000), Buraschi and Cornelli (2002)). 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 in order to increase 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 the realms of social interactions, provision of public goods, tax compliance, volunteering, and experimental labor contracts (see Frey (1997) and Frey and Jegen (2001) for surveys). Studying schoolchildren collecting donations for a charitable organization, Gneezy and Rustichini (2000a) thus found that they collected less money when given performance incentives (see also Frey and Götte (1999) on volunteer work supply). This is in line with the idea in Titmuss (1970), who argued that paying blood donors could actually reduce supply. On the punishment side, Akerlof and Dickens (1982) suggested that imposing stiffer penalties for crimes might sometimes be counterproductive, by undermining individuals' "internal justification" for obeying the law. Frey (1997) provided evidence to that effect concerning tax compliance, and Gneezy and Rustichini found (2000a) that fining parents for picking up their children late from day-care centers resulted in more late arrivals. In experiments on labor contracting, Fehr et al. (2001) and Fehr and Gächter (2002)) found that subjects provided less effort when the contract specified fines for inadequate performance than when it did not. These findings are in line with a large literature in psychology that has documented many instances where explicit incentives for task performance lead to decreased motivation and reduced long-run performance (see, e.g. Festinger and Carlsmith (1959), Deci (1975), Deci and 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., Gibbons (1997), Prendergast (1999), and Lazear (2000a,b)).[1] A more discriminating analysis is thus required.

Second, 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., Batson (1998), Freeman(1997)). Charitable and non-profit institutions make ample use of donors' desire

[1] Even in the specific situations considered above, it is clear that rewards are not aversive *per se*. For instance, organizations that collect blood try to make it a relatively pleasant experience for donors, and increasing the discomfort or time costs is unlikely to increase supply. With respect to charitable contributions, a simple comparison of US and (say) France suggests that their tax treatment does matter in the anticipated way.

to publicly demonstrate their generosity and selflessness (or at least the appearance thereof): available displays range from lapel pins and T-shirts to plaques in opera houses, or buildings named after large contributors. The presence of a social signalling motive for giving, as distinct from pure altruism, 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 largely 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, leading individuals to engage in self-sacrifice for reputational reasons. To understand these mechanisms it is again important to not posit exogenous social constraints, but rather to model the inferences involved in sustaining such norms and the external factors facilitating or inhibiting them.

Finally, as much as people care about the *opinion others have of them,* they care about their own *self-image* –or, as the expression goes, being able to look at themselves in the mirror. Adam Smith (1776) eloquently described this motive for acting in a moral or unselfish way, in terms of individuals 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 behavior, or even feelings, and certain values, long-term goals or identities.[3] Recent empirical studies confirm the importance of such self-image concerns and their contribution to prosocial behavior.[4] In particular, a clever experiment by Dana et al. (2003) reveals that when people are given the opportunity to remain ignorant of how their choices affect others, or of their precise role in the

[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. Kahneman and Knetsch (1992) find that subjects' stated willingness to pay for alternative publics goods is well predicted by independent assessments of the associated "moral satisfaction". 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 (being sensitive to the needs of others, not taking advantage of them, trusting and being trusted).

[4] For instance, in a transportation-related survey of about 1,300 individuals, Johansson-Stenman and Svedsäter (2003) find that people who are asked which attributes are most important to them in a car systematically put environmental performance near the top and social status near the bottom; but when asked about the true preferences of their neighbors or average compatriots, they give dramatically reversed rankings. Interviews with car dealers show intermediate results.

outcome (as with firing squads, which always have one blank bullet), many choose not to know and revert to selfish choices. In a related vein, Murningham 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.[5]

In this paper, we examine the set of issues discussed above with a theory of prosocial behavior that combines heterogeneity in individuals' degrees of altruism and greed with a concern for social reputation or self-respect. In other words, prosocial actions are undertaken both because a certain fraction of individuals are genuinely other-regarding, and due to the fact that, in many cases:

– people want to signal to others that they are generous, fair, public-spirited, disinterested, courageous, etc. Prosocial behaviors are then part of a general quest for social esteem;

– people strive to maintain a certain view of "what kind of a person" they are. We use here also a cognitive approach, based on psychologists' findings that individuals commonly use their own past behavior as "diagnostic" of their deep preferences. Conversely, they alter their behavior with a view to its impact on the inferences they will later on make about themselves. Underlying this self-signaling is the fact that the actions one has taken are more memorable than their exact motivations.[6]

Our theory thus emphasizes the attributions made from individuals' investments in identity-building and demonstration goods, such as giving blood, contributing time or money to a charity, or giving one's life for familial honor or country. In particular, the crowding-out effect that we obtain is based on a very simple intuition: the presence of rewards or punishments *spoils* the reputational (or self-reputational) value of good deeds, creating doubt as to the extent to which they were performed for the incentives rather than for themselves. This effect is in line with what psychologists refer to as the "overjustification effect" (e.g., Lepper et al. (1973)), to which we give here a formal content in terms of a signal-extraction problem. It is also consistent with the informal explanation provided by the designers of several of the experiments reported above; for instance, Frey and Jegen (2001) state 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".*

[5] A related set of classical findings in social psychology concerns attitudes towards victims. People who directly directly witness abuse or injustice often tend to derogate the victims, unless they are able to either help the victim or not feel any personal responsibility for his or her suffering (see, e.g., Batson (1998, p.296) or Lerner (1980)). By trying to convince both themselves (often with the help of some form of self-deception) and others that the victim would have derived only small benefits from such help, or did not really deserve it, they seek to avoid the adverse inferences about their character that not helping might otherwise generate.

[6] For more discussion, see Bodner and Prelec (2003) and Bénabou and Tirole (2004). In psychology, the idea that individuals take their actions as diagnostic of their preferences originated with Bem (1972), but it also relates to cognitive dissonnance theory (Festinger and Carlsmith (1959)).

The model also helps explain why it is often more effective to contribute one's time than money (as with a parent trying to demonstrate her love or concern to a child), and allows us to study the disclosure of information about one's generosity, showing in particular how attempts to buy social prestige may backfire. Other insights that emerge from an information-based approach include the impact of observability to others and memorability to oneself, together with key features of the distribution of intrinsic social preferences, on the emergence of multiple social norms enforced by the interplay of honor and shame. Finally, an explicit treatment of the signaling dimension of prosocial behavior also sheds light on the nature of competition among the recipients of altruism (or their intermediaries) such as charities, NGO's, arts groups, universities, and similar non-profit organizations. Donating to worthy causes is sometimes rewarded with non-trivial "perks" such as preferred seating, meetings with famous performers, gala events which also allow valuable social networking, or naming rights to a building, stadium or professorial chair. In contrast to what would be predicted by standard Bertrand competition, we show that fund-raisers competing for donations will not end up dissipating most the proceeds on such incentives, due to what might be termed a "holier than thou" form of emulation. This same effect implies that sponsor competition can even reduce social welfare, by inducing agents to engage in more inefficient sacrifices.

The two papers most closely related to the present work are Bénabou and Tirole (2003) and Seabright (2002). In our earlier work we developed an alternative (but similarly cognitive) approach to the potential conflict between extrinsic and intrinsic motivation, based on the idea that giving an agent high-powered incentives may convey bad news about the nature of the task or his ability, whenever the principal has private information about these variables. That theory has natural applications to child-rearing, education, pay-for-performance and empowerment versus monitoring of employees. It is much less relevant for incentives that are designed for large groups and for activities –such as blood donations, voting, late arrivals at a day care center, or contributing to a charitable cause– about which the principal seems unlikely to have superior knowledge. In Seabright's paper, as here, an individual's direct benefit from contributing to a "civic activity" depends on his private type. Agents then overinvest in this activity in order to gain a reputation that will make them more desirable partners in a later matching market. Most importantly, Seabright shows that if each agent sets his own price for participating in the civic activity (subject to a cap set by some public authority), and if this price is constrained to be non-negative, a "payment discontinuity"arises, whereby small rewards are never observed. Intuitively, an individual is better off foregoing a small reward and pooling with the socially desirable types who ask for none at all.

Other related papers include Bodner and Prelec (2003) and Bénabou and Tirole (2004) on self-signaling, Akerlof and Kranton (2000) on identity, Brekke et al. (2003) on moral motivation, Denrell (1998) on credibility and compensation, Veblen (1899), Leibenstein (1950) and Pesendorfer (1995) on ostentatious consumptions as signaling devices, and Bernheim (1994) on

actions designed to signal conformity of tastes with others.[7] As in Pesendorfer (1994), we build on the idea that sponsors can exploit agents' signaling concerns to their advantage. Finally, our paper also ties in to the large literature on gifts and donations, such as Andreoni (1993) Glazer and Konrad (1996), Harbaugh (1998) and Prendergast and Stole (2001).

The paper is organized as follows. Section I presents the model and Section II a first, intuitive illustration of the *image-spoiling* effect of rewards. Section III then uses a Normal signal-extraction specification to demonstrate the *crowding-out* phenomenon, as well as related forms of the *overjustification effect.* Section IV centers on the phenomenon of *social norms,* explaining how multiple standards of socially or personally "acceptable" behavior arise from the interplay of honor and stigma, and more generally identifying situations that make individual participation decisions *strategic complements or substitutes.* Sections V and VI then turn to the design of contracts, exploring respectively the case of a monopoly sponsor –including issues of *confidentiality and disclosure*– and that of *sponsor competition,* which is shown to cause rewards to bid down rather than up, and to potentially reduce social welfare. Section VII concludes.

I The model

We study the behavior of agents (either a single individual or a large population) who must decide on their participation in some prosocial activity: provision of a public good, contributing to a worthy cause, engaging in a friendly or reciprocating action, refraining from imposing negative externalities on others, etc. Each thus select a participation level a from some choice set $A \subset \mathbb{R}$ that may be discrete (voting, blood donation) or continuous (time or money volunteered, fuel efficiency of car purchased). Contributing entails a utility cost $C(a)$ and yields a monetary or material reward ya. The incentive rate $y \gtrless 0$ may reflect a proportional subsidy or tax on a faced by agents in this economy, or the fact that participation requires a monetary contribution.[8] It is set by a principal or "sponsor", and for now we assume that agents take it as given.

Letting v_a and v_y denote a typical agent's intrinsic valuations for his own prosocial actions and for money (consumption of market goods), participation at level a yields a direct net benefit

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

An individual's preference type or "identity" $\mathbf{v} \equiv (v_a, v_y) \in \mathbb{R}^2$ is drawn from a continuous distri-

[7] Our work is also technically related to a small literature on signals that convey diverging news about different underlying characteristics. Thus Araujo et al. (2004) develop a model to reflect the evidence that the General Educational Development (GED) dipoma signals both high cognitive skills and low non-cognitive skills.

[8] The latter case corresponds for instance to situations where a contribution of $-y > 0$ dollars feeds a hungry children, funds a artistic performances, etc. A higher "reward", meaning a lower $-y$, may for example reflect a higher matching rate by a sponsor or government.

bution 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 decides how to act but not observable by others.

a) Social signaling

In addition to these direct payoffs, actions also carry reputational costs or benefits. In the *social signaling* interpretation of the model these arise from interactions with the rest of society –family, friends, colleagues, compatriots. The value of reputation may then be instrumental (allowing the individual to match with more desirable partners, as in Denrell (1998), Gintis et al. (2001) or Seabright (2002)), or purely hedonic (social esteem as a consumption good). Let x denote the probability that the individual's choice of a is observed by others, and $\hat{f}(\mathbf{v}|a, y)$ the posterior distribution of $\mathbf{v}$ conditional on (a, y). For simplicity, we assume that the (continuation) value of reputation is a linear functional of $\hat{f}$ –that is, only expectations matter– with type-independent coefficients. The reputational payoff from choosing a is thus[9]

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

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). Defining $\mu_a \equiv x\gamma_a$ and $\mu_y \equiv x\gamma_y$, the overall utility of an agent with preferences $\mathbf{v} \equiv (v_a, v_y)$ and reputational concerns $\boldsymbol{\mu} \equiv (\mu_a, \mu_y)$ is thus

$$(v_a + v_y y)\, a - C(a) + \mu_a E(v_a|a, y) - \mu_y E(v_y|a, y), \tag{3}$$

which he will maximize over $a \in A$.[10] In the basic version of the model, $\boldsymbol{\mu}$ is taken to be common to all agents, and thus public knowledge. Later on we shall allow for unobserved heterogeneity in image-consciouness, with $\boldsymbol{\mu}$ distributed independently of $\mathbf{v}$ according to a density $m(\boldsymbol{\mu})$.

b) Self-signaling and identity

The model also admits an important reinterpretation in terms of *self-signaling.* Suppose that at the time he makes his decision, the individual engages in a self-assessment, or receives some external signal about his type: "How important is it for me to contribute to the public good? How much do I care about the money that I would then receive or forfeit? What are my

[9] 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 linear valuation of reputation also avoids building into agent's preferences either information-aversion (concave functional of $\hat{f}$) or information-loving (convex functional of $\hat{f}$. Naturally, the expectations in (2) are also conditional on all publicly available information in addition to y.

[10] Agents may also care about the aggregate provision of the public good, but as long as it enters their preferences separably it will not affect their decisions or inferences, so we leave it out of (3). One could also easily incorporate a form of "reciprocity" whereby a higher aggregate contribution $\bar{a}$ by other raises the individual's intrinsic desire to contribute –e.g., replace $v_a a$ by $v_a(1+\lambda\bar{a})a$. While such a complementarity in payoffs is often invoked to explain why people contribute more when they know that others do, we will show in Section IV that it is in fact not required: in our information-based model, social norms of participation emerge endogenously.

values?". This self-assessment or signal, however, may not be perfectly recalled or "accessible" later on –in fact, there will be strong incentives to remember it in a self-serving way. Actions, by contrast, are much easier to quantify, record and remember than their underlying motivation, making it rational for an agent to define himself partly through his past choices: "I am the kind of person who behaves in this way".[11] Suppose therefore that the signal motivating the participation decision is forgotten with probability x, and that later on the agent cares about his self-image –that is, derives utility from his own beliefs concerning his type. This may reflect a hedonic motive (people enjoy feeling generous or disinterested, e.g. Akerlof and Dickens (1987) or Köszegi (2000)), an instrumental purpose (providing motivation to undertake and persevere in long-term tasks or social relationships, e.g. Carrillo and Mariotti (2000) or Bénabou and Tirole (2002)), or both. 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, it is clear that the model is formally equivalent to the social-signaling one.

II The image-spoiling effect of rewards: basic intuitions

Should people be rewarded for performing good deeds or will this taint these acts, raising doubts as to the true underlying motivation? We provide here a first, intuitive illustration of this idea, based on the observation that the presence of rewards *changes the pool of participants.*

We assume here, as in much of the paper, that the participation decision is binary: $A = \{0, 1\}$; we then normalize $C(0) = 0$ and denote $C(1) = c_a$. An individual therefore participates if

$$v_a - c_a + v_y y + R(y) \geq 0, \tag{4}$$

where

$$R(y) \equiv \mu_a \left[E\left(v_a|1, y\right) - E\left(v_a|0, y\right)\right] - \mu_y \left[E\left(v_y|1, y\right) - E\left(v_y|0, y\right)\right] \tag{5}$$

is the net reputational gain (or loss) from participating, given that there is a monetary incentive of y to do so. Note that $R(y)$ is an equilibrium variable, but is type-independent.

The impact of incentives on participation is illustrated in Figure 1, for the case where v_a and v_y are independent variables, $f\left(v_a, v_y\right) = g\left(v_a\right) h\left(v_y\right)$, while μ_a and μ_y are fixed. Consider first the case in which no reward is offered, $y = 0$. An agent then participates if and only if $v_a \geq c_a - R(0) \equiv v_a^*$, so nothing is learned about v_y and the participation rule is defined by a threshold for intrinsic motivation. To determine this value, let us define, for all v_a,

$$\mathcal{M}^+\left(v_a\right) \equiv E\left(v_a \left|\tilde{v}_a \geq v_a\right.\right) = \frac{\int_{v_a}^{\infty} v g\left(v\right) dv}{\int_{v_a}^{\infty} g\left(v\right) dv}, \tag{6}$$

[11] For a model of, and psychological references on, the links between imperfect recall and self-signaling, see Bénabou and Tirole (2004) or Battaglini et al. (2002). On self-signaling in a dual-self or dual-utility (decision utility plus "diagnostic" utility) model, see Bodner and Prelec (2003).

$$\mathcal{M}^{-}(v_a) \equiv E(\tilde{v}_a | \tilde{v}_a \leq v_a) = \frac{\int_{-\infty}^{v_a} v g(v)\, dv}{\int_{-\infty}^{v_a} g(v)\, dv}. \tag{7}$$

The first expression governs the *"honor"* conferred by participation, which is the difference between $\mathcal{M}^{+}(v_a)$ and the unconditional expectation $\bar{v}_a$. The second one governs the *"stigma"* from abstention, which is $\mathcal{M}^{-}(v_a) - \bar{v}_a$. Since both are nondecreasing functions of the (potential) cutoff v_a, their difference $\mathcal{M}^{+}(v_a) - \mathcal{M}^{-}(v_a)$, which defines the net reputational effect of participation, may in general increase or decrease with v_a.[12] The cutoff for unpaid participation (when interior) is thus defined as the solution to

$$\Psi(v_a^*) \equiv v_a^* + \mu_a \left[\mathcal{M}^{+}(v_a^*) - \mathcal{M}^{-}(v_a^*)\right] = c_a, \tag{8}$$

assuming for the moment that Ψ is increasing, ensuring uniqueness (multiplicity will be considered later on).

Let us now introduce a reward $y > 0$; the reasoning would work in reverse for a fine or punishment $y < 0$. As seen from (4) and illustrated in Figure 1, the indifference locus is now a line with slope $-1/y$. Ignoring, *in a first step,* changes in inference, participation expands, as types in the hatched area $(A + B)$ are drawn in. There are, however, two reputational effects:

– The new participants have a lower valuation v_a for the public good than those who participate in the absence of reward. The honor from choosing $a = 1$ thus declines, but so does the stigma from choosing $a = 0$.[13]

– The new participants are greedy types (tend to be interested in money): $E(v_y|1, y) > E(v_y|1, 0)$, which is always an adverse reputational effect.

If the overall impact on reputation is negative, $R(y) < R(0)$ as drawn in Figure 1 and ensured by Proposition 1 below, the reward *attracts* some new participants (more greedy agents in area B) but *repels* some existing ones (more public-spirited agents in area C).[14] Overall, participation may increase or decrease, depending on the weights given to B and C by the distribution f.

The next sections will identify situations in which net crowding out does occur. Clearly, a necessary condition is that greater incentives depress the reputational value of participation, $R(y)$. Using a simple reasoning based on Figure 1, we can already find conditions under which this (weaker) effect occurs when a reward is introduced, starting from a no-reward situation.

[12] For example, we have: $\mathcal{M}^{+} - \mathcal{M}^{-} \equiv 1/2$ for $g(v_a) \equiv 1$ on $[0, 1]$; more generally, if $g(v_a) = (\alpha + 1) v_a^{\alpha}$ on $[0, 1]$, with $\alpha > -1$, then $\mathcal{M}^{+}(v_a) - \mathcal{M}^{-}(v_a) = [(1 + \alpha)/(2 + \alpha)]\left[(1 - v_a)/\left(1 - v_a^{1+\alpha}\right)\right]$. This function is increasing in v_a when $\alpha > 0$ (e.g., $\alpha = 1$) and decreasing when $-1 < \alpha < 0$. (e.g., $\alpha = -1/2$); see Proposition 7.

[13] Indeed, from the independence of the two variables, $E(v_a \mid v_a + v_y y + R(0) \geq c_a) = E_{v_y}(E_{v_a}(v_a \mid v_a \geq -v_y y + c_a - R(0))) \leq E_{v_y}(E_{v_a}(v_a \mid v_a \geq c_a - R(0))) = E_{v_a}(v_a \mid v_a \geq v_a^*)$, under the very plausible assumption that v_y is bounded below by zero.

[14] This is in line with 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.

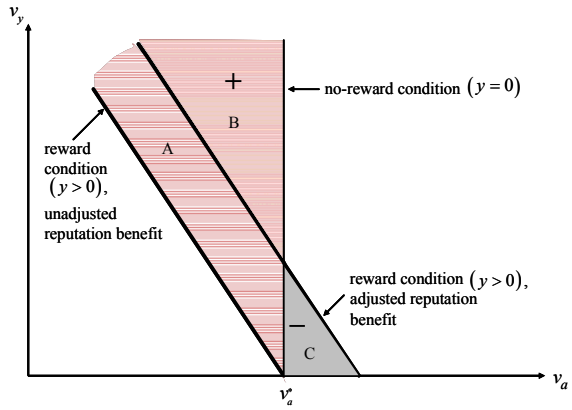

Figure 1: the effects of rewards on the pool of participants

Proposition 1 *Assume that $\Psi' \geq 0$, where Ψ was defined in (8), and that the lower bound of agents' valuation for money is $v_y^- = 0$. Then, if $\mu_y = 0$, or if v_a and v_y are independent or negatively affiliated, the introduction of a reward lowers the net reputational value of participation: $R(y) < R(0)$, for all $y > 0$.*

Proof: see the appendix.

Negative affiliation implies that the two posteriors about v_a and v_y tend to be updated in opposite directions, implying that agents who contribute only in response to external incentives $y > 0$ must pay a "double dividend" in terms of lost reputation. The condition $\Psi' \geq 0$ ensures the uniqueness of the equilibrium (at least when $\mu_y = 0$ or v_y is known), and more generally rules out the kind of strong complementarity between agents' participation decisions that will be extensively studied in Section IV.

III The overjustification effect and crowding out

We provide in this section a more explicit model of the image-spoiling effect of rewards, in which we also allow for variations in individuals' degrees of image consciousness. The main result is that the presence of stronger extrinsic incentives causes observers (or, in retrospect, the individual himself) to attribute a smaller role to intrinsic motivation in explaining behavior; this in turn, leads to a *decreasing supply curve,* under conditions that we identify. Formally, we show how the "overjustification effect" discussed by psychologists can be understood as a simple signal-extraction problem in which rewards *amplify the noise,* and thus crowd out reputational motivation.

We allow here actions to vary continuously: $a \in A = \mathbb{R}$, with a strictly convex cost $C(a)$; for technical reasons we also assume that as $|a| \to +\infty$, $C(a)$ is asymptotically equivalent to

a polynomial in a. The overall utility of an agent with preferences $\mathbf{v} \equiv (v_a, v_y)$ and (self) reputational concerns $\boldsymbol{\mu} \equiv (\mu_a, -\mu_y)$ is still given by (3), but now *both* $\mathbf{v}$ and $\boldsymbol{\mu}$ may vary across individuals and are private information. Observers (or the individual in retrospect) only know the joint distribution, which is defined by

$$\begin{pmatrix} v_a \\ v_y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{v}_a \\ \bar{v}_y \end{pmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ay} \\ \sigma_{ay} & \sigma_y^2 \end{bmatrix} \right), \quad \bar{v}_a \gtreqless 0, \quad \bar{v}_y > 0, \tag{9}$$

$$\begin{pmatrix} \mu_a \\ \mu_y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{\mu}_a \\ \bar{\mu}_y \end{pmatrix}, \begin{bmatrix} \omega_a^2 & \omega_{ay} \\ \omega_{ay} & \omega_y^2 \end{bmatrix} \right), \quad \bar{\mu}_a \geq 0, \quad \bar{\mu}_y \geq 0 \tag{10}$$

and the simplifying assumption that these two random variables are independently distributed.[15]

A Individual and aggregate responses to incentives

For an agent with type $(\mathbf{v}, \boldsymbol{\mu})$, the optimal choice of a when the subsidy rate is y is determined by the first-order condition :

$$C'(a) = v_a + v_y y + R(a, y), \tag{11}$$

where the last term captures the *marginal* reputational value of contributing at level a :[16]

$$R(a, y) \equiv \mu_a \frac{\partial E\left(v_a | a, y\right)}{\partial a} - \mu_y \frac{\partial E\left(v_y | a, y\right)}{\partial a}.$$

Consider now the inference problem of an external or (retrospective) internal observer. From (11), the agent's choice of a reveals the *sum his three motivations* to contribute (at the margin): *intrinsic, extrinsic, and reputational.* Moreover, since $\boldsymbol{\mu}$ is uncorrelated with $\mathbf{v}$ and the functions $E\left(v_a | a, y\right)$ and $E\left(v_y | a, y\right)$ are known in equilibrium, $R(a, y)$ is independent of $\mathbf{v}$, conditionally on a. Reputational motivation thus acts as a heteroskedastic normal shock, with mean

$$\bar{R}(a, y) \equiv \bar{\mu}_a \frac{\partial E\left(v_a | a, y\right)}{\partial a} - \bar{\mu}_y \frac{\partial E\left(v_y | a, y\right)}{\partial a} \tag{12}$$

and variance

$$\Omega(a, y)^2 \equiv \begin{pmatrix} \frac{\partial E(v_a|a,y)}{\partial a} & -\frac{\partial E(v_y|a,y)}{\partial a} \end{pmatrix} \begin{bmatrix} \omega_a^2 & \omega_{ay} \\ \omega_{ay} & \omega_y^2 \end{bmatrix} \begin{pmatrix} \frac{\partial E(v_a|a,y)}{\partial a} \\ -\frac{\partial E(v_y|a,y)}{\partial a} \end{pmatrix}. \tag{13}$$

[15] As is often the case, normality yields great tractability at the cost of allowing certain variables to take implausible negative values. By choosing the means large enough, however, one can make the probability of such realizations arbitrarily small; but (9)-(10) should really be interpreted as local approximations, consistent with the linearity of preferences assumed throughout the paper.

[16] In addition to (11), the second-order condition $\partial R\left(a; y\right) / \partial a \leq C''(a)$ must also hold. We focus throughout this section on equilibria where the two components of reputation are twice-differentiable functions of a.

Standard signal-extraction results for normal random variables then yield:

$$E\left(v_a|a,y\right) = \bar{v}_a + \rho(a,y)\cdot\left(C'(a) - \bar{v}_a - y\cdot\bar{v}_y - \bar{R}(a,y)\right), \tag{14}$$

$$E\left(v_y|a,y\right) = \bar{v}_y + \chi(a,y)\cdot\left(C'(a) - \bar{v}_a - y\cdot\bar{v}_y - \bar{R}(a,y)\right). \tag{15}$$

where

$$\rho(a,y) \equiv \frac{\sigma_a^2 + y\sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2 + \Omega(a,y)^2}, \tag{16}$$

$$\chi(a,y) \equiv \frac{y\sigma_y^2 + \sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2 + \Omega(a,y)^2}. \tag{17}$$

Quite 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 $C'(a)$, *net of the average* extrinsic and reputational incentives to contribute at that level.

An equilibrium is thus defined as a solution to the system of nonlinear differential equations (14)-(15) in $E\left(v_a|a,y\right)$ and $E\left(v_y|a,y\right)$, such that the second-order condition also holds everywhere. This is in general a complicated problem, but we are able to solve it in an important class of cases, namely that where the variance of the reputational incentive, $\Omega(a,y)^2$, is independent of a. This means that either:

– *all agents have the same reputational concerns,* so that the covariance matrix of $\boldsymbol{\mu}$ in (10) is zero; or,

– *the cost function $C(a)$ is quadratic,* leading to expectations $E\left(v_a|a,y\right)$ and $E\left(v_y|a,y\right)$ and a differential system that is linear in a.

These two cases are solved in Propositions 2 and 3 respectively.

Proposition 2 *Let all agents have the same valuation $(\bar{\mu}_a, \bar{\mu}_y)$ for reputation, and denote $\mu(y) \equiv \bar{\mu}_a\rho(y) - \bar{\mu}_y\chi(y)$, where $\rho(y)$ and $\chi(y)$ are defined by (16)-(17) with $\Omega \equiv 0$. The optimal action a for an agent with type (v_a, v_y) is an increasing function of $v_a + v_y y$, given by*

$$\Gamma(a,\mu(y)) \equiv \int_0^{+\infty} C'(a-\mu(y)z)\ e^{-z}dz = v_a + y\cdot v_y. \tag{18}$$

The marginal reputations from contributing are $\partial E\left(v_a|a,y\right)/\partial a = \rho(y)\ \partial\Gamma(a,\mu(y))/\partial a$ and $\partial E\left(v_y|a,y\right)/\partial a = \chi(y)\partial\Gamma(a,\mu(y))/\partial a$, with a net value of $\bar{R}(a,y) = \mu(y)\partial\Gamma(a,\mu(y))/\partial a$.

Proof: see the appendix.

Rewriting (18) in the more familiar form of the first-order condition

$$C'(a) - \mu(y)\int_0^{+\infty} C''(a-\mu(y)z)\ e^{-z}dz = v_a + y\cdot v_y \tag{19}$$

makes clear that $\Gamma(a, \mu(y))$ can be interpreted as a *reputation-adjusted marginal cost* of contributing at the level a, and allows us to analyze the effects of incentives on individual's behaviors and inferences. On one hand, a higher y increases agents' non-reputational motivation to contribute, $v_a + y \cdot v_y$. On the other hand, it tends to reduce the value of a marginal contribution along both dimensions of social or self image. In the benchmark case of no correlation ($\sigma_{ay} = 0$), for instance,

$$\rho(y) = \frac{1}{1 + y^2\sigma_y^2/\sigma_a^2} \qquad \text{and} \qquad \chi(y) = \frac{y\sigma_y^2/\sigma_a^2}{1 + y^2\sigma_y^2/\sigma_a^2}\ , \tag{20}$$

so a higher y acts very much like an increase in the *noise-to-signal ratio* $\theta \equiv \sigma_y/\sigma_a$, leading observers who seek to parse out the agent's motives to decrease the weight given to social orientation, $\rho(y)$, and increase its counterpart for greediness, $\chi(y)$.[17] Thus $\mu(y)$ falls, and this lowers the reputational incentive to act prosocially, $\bar{R}(a, y)$, which is the second term in (19). Simply put, extrinsic motivation crowds out reputational motivation.

A positive correlation between v_a and v_y tends to amplify the decline in $\rho(y)$, thereby accentuating the overjustification effect; a negative σ_{ay} works in reverse. Intuitively, the more v_a and v_y tend to move together, the less observing a higher contribution $a = \Gamma^{-1}(v_a + v_y y)$ represents good news about the agent's intrinsic valuation v_a; and the larger is y, the stronger is this "discounting" effect. 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 $1/(1 + \theta y)$.[18]

The results in Proposition 2 can be made more specific when the cost function is quadratic: with $C(a) = ka^2/2$, individual supply (18) becomes

$$a = \frac{v_a + y \cdot v_y}{k} + \bar{\mu}_a \rho(y) - \bar{\mu}_y \chi(y). \tag{21}$$

The aggregate supply curve $\bar{a}(y)$, obtained by summing over all agents, thus has slope

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

clearly reflecting the crowding out (or in) of reputational motivation by extrinsic incentives.

Corollary 1 ***(small incentives).*** *Let $C(a) = ka^2/2$, and assume that all agents have the same reputation concerns ($\Omega \equiv 0$). Small rewards or punishments are counterproductive, $\bar{a}'(0) < 0$, whenever*

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

[17] More specifically, $y\chi(y) = 1 - \rho(y)$ rises with y everywhere, but the same is true of $\chi(y)$ only up to $y = 1/\theta$. For what follows, note that that the function $\mu\partial\Gamma(a, \mu)/\partial a = C'(a) - \Gamma(a, \mu)$ is increasing in μ.

[18] Recall that $\theta \equiv \sigma_y/\sigma_a$. The effect of σ_{ay} on the slope of $\chi(y)$ is somewhat more complex, as it depends on σ_{ay}^2. The exact formulas for $\rho'(y)$ and $\chi'(y)$ are given in the proof of Corollary 1.

Proof: see the appendix.

This case is interesting because some of the experimental evidence on crowding out finds that these effects occur mainly for relatively small rewards (see, e.g., Gneezy and Rusticchini (2000 a,b)). Note also that the above condition may hold even when only one reputational concern is operative ($\mu_a = 0$ or $\mu_y = 0$), as well as when v_a and v_y are either positively or negatively correlated (the latter case requires $\mu_y > 0$). Sufficiently large incentives, on the other hand, always increase supply: as $|y| \to +\infty$, $\rho(y)$ and $\chi(y)$ both tend to zero, so behavior is dominated by the direct, non-reputational effect.

Is crowding out only a "small-stakes" phenomenon? The answer is negative: some of the studies reviewed earlier find such an effect even with fairly substantial incentives (e.g., Fehr and Gächter (2000), Bohnet et al. (2001)), and our model shows that it can in fact be intermediate-sized rewards that have paradoxical effects.

Corollary 2 ***(intermediate incentives).*** *Let $C(a) = ka^2/2$ and $\Omega \equiv 0$. Let v_a and v_y be uncorrelated, and let $\theta \equiv \sigma_y/\sigma_a$. Incentives are counterproductive, $\bar{a}'(y) < 0$, whenever*

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

Consequently: (i) if $\bar{\mu}_y\theta^2 < \bar{v}_y/k$, then for all $\bar{\mu}_a$ above some threshold $\mu_a^* > 0$ there exist an interval of prices $[y_1, y_2]$ with $0 < y_1 < y_2$, such that $\bar{a}(y)$ is decreasing on $[y_1, y_2]$ and increasing everywhere else on $\mathbb{R}$. As $\bar{\mu}_a$ rises, y_1 decreases and y_2 increases, so that $[y_1, y_2]$ widens; conversely, this is interval becomes empty as μ_a falls below μ_a^*.
(ii) if $\bar{\mu}_y\theta^2 > \bar{v}_y/k$, there exists an interval $[y_1', y_1]$ with $y_1' < 0 < y_1$ such that, for all $\bar{\mu}_a \geq 0$, $\bar{a}(y)$ is decreasing on $[y_1', y_1]$ and increasing everywhere else on $\mathbb{R}$.

Proof: see the appendix.

Since a number of experimental studies seem to find a "discontinuity at zero" in subjects' response to rewards,[19] it may be worth commenting on the fact that in our model individual behavior is always continuous, whether or not crowding out occurs. This reflects the fact that agents make very fine inferences based on the precise size of the reward and optimal (Bayesian) filtering rules, which admittedly represents greater cognitive sophistication than most individuals actually have, or find worthwhile to apply in one-shot experimental situations. If people use in fact a more "coarse" filter, such as broadly classifying situations into "unrewarded", "small reward" and "large reward" categories, the crowding-out results arising from the mechanism we identify will also result in a discontinuities, particularly at the origin.

[19] See, e.g., Gneezy and Rusticchini (2003). One should note, however, that these studies typically involve only at most two or three data points (reward rates) to the right or left of zero, making it hard to identify exactly where the supply curve bottoms out.

B Prominence, memorability and publicity

Intuition suggests that participation in good causes will be enhanced if it is more conspicuous, namely if the contribution is more likely to be noticed by others (social signaling) or the individual more likely to be reminded of it (self-signaling). Indeed, public authorities and other sponsors make heavy use of both public displays and private mementos conveying honor or stigma. 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 "diplomas" rather than significant cash prizes to professors delivering special lectures, etc.[20] Conversely, the ancient practice of the pillory has been updated in the form of televised arrests and, in some states and towns, 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 effects also play an important role in boosting contributions, as they create a rehearsal mechanism: if acquaintances all contribute to a cause, one is constantly reminded of one's generosity, or lack thereof. Relatedly, people 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.

Formally, prominence can be modeled as an increase in x, translating into a homothetic increase in $\left(\mu_a, \mu_y\right)$. Our model then confirms the above intuitions, but also delivers some important, and empirically relevant, caveats. First, when the existing structure of rewards and punishments is such that the (marginal) reputational return to participation is actually negative, an increase in visibility naturally tends to reduce rather than increase participation. Second, and more subtly, when there is heterogeneity in image-consciousness, giving increased prominence or scrutiny to individual's decisions may backfire (at least partially), as *good actions come to be suspected of being image-motivated.* Our model thus brings to light *a new form of the overjustification effect,* linked not to material incentives but to image-related ones.

We analyze these issues using the quadratic-cost specification, which allows the model to be solved when image concerns as well as goods valuations differ across agents.

Proposition 3 *Let the cost of contributing be quadratic, $C(a) = ka^2/2$.*[21] *Each agent's optimal action a is then defined by*

$$a = \frac{v_a + y \cdot v_y}{k} + \mu_a \rho(y) - \mu_y \chi(y), \tag{23}$$

[20] Potters et al. (2001, 2002) explain (experimentally) charities' frequent strategy of publicly announcing "leadership" contributions, and the higher yields achieved when donors act sequentially rather than simultaneously, by a signaling effect about the quality of the public good. They show in particular that this explanation works better than one based on "reciprocating" the generosity of early donors. A complementary explanation could be donors' desire to signal, socially and to themselves, how generous and public-spirited they are.

[21] We also focus attention here on equilibria where the expectations $E\left(v_a|a,y\right)$ and $E\left(v_y|a,y\right)$ are linear in a. While this requires that costs be quadratic, and conversely quadratic costs naturally lead to such a solution for (14)-(15), even with $C(k) = ka^2/2$ one cannot rule out a priori the existence of other, nonlinear, solutions.

where $\rho(y)$ and $\chi(y)$ are obtained from (16)-(17) with $\Omega(a,y) \equiv \Omega(y)$ given by the fixed-point equation:

$$\Omega(y)^2/k^2 \equiv \omega_a^2\rho(y)^2 - 2\omega_{ay}\rho(y)\chi(y) + \omega_y^2\chi(y)^2. \tag{24}$$

The marginal reputations from contribution are $\partial E\left(v_a|a,y\right)/\partial a = \rho(y)k$ and $\partial E\left(v_y|a,y\right)/\partial a = \chi(y)k$, with a net value of $R(y) = (\mu_a\rho(y) - \mu_y\chi(y))k$.

Proof: see the appendix.

These results demonstrate in particular how a greater variability of image concerns, $\Omega(y)^2 = Var\left(R(y)\right)$, 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 it itself endogenous, however, as reputation-motivated agents take account of how their collective behavior affects observers' signal-extraction-problem; this is reflected in the fixed-point equation.[22]

Proposition 3 also allows us to formalize the idea, mentioned above, that increased prominence gives rise to an offsetting overjustification effect. Indeed, let all reputational weights $\boldsymbol{\mu} = (\mu_a, \mu_y)$'s be scaled up by some factor x –not necessarily a probability: x could also reflect the number of future periods during which the "record" of the agent's behavior will be kept, or the number of people who will hear about it. The material incentive y remains constant. Aggregate supply is now

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

where the dependence on x indicates that all the covariance terms $\left(\omega_a^2, \omega_{ay}, \omega_y^2\right)$ in the fixed-point equation (24) defining $\Omega(y,x)$ are now multiplied by x^2. As a result, $\Omega(y,x)$ is increasing in x,[23] and $\rho(y,x)$ and $\chi(y,x)$ consequently decrease with x. Intuitively, a greater visibility of actions and the rewards attached to them has two offsetting effects on the reputational incentive to invest:

– 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 (μ_a is large relative to μ_y) this increases the incentive to act in a prosocial manner, whereas for those most concerned about not appearing greedy (μ_y is large relative to μ_a) it has the reverse effect.[24]

[22] We show in the proof of Proposition 3 that there always exists a solution to (24), and that it is unique when $\omega_{ay} = 0$. When $\omega_{ay} \neq 0$ there might be multiple equilibria, with different degrees of informativeness. Since the general theme of multiplicity is investigated in Section IV.B, we do not pursue it here.

[23] This is clear from (24) in the case where $\omega_{ay} = 0$, but we show in the appendix that it holds more generally in any stable equilibrium.

[24] We are focussing here, for illustrative purposes only, on the "natural" case where ρ and χ are both positive, which occurs unless σ_{ay} is very negative; see (16)-(17).

– a dampening effect, as *reputation* along both dimensions *becomes less sensitive* to the individual's behavior, which observers increasingly ascribe to image concerns. Formally, $\rho(y,x)$ and $\chi(y,x)$ take lower values.

This tradeoff implies that publicity for good or bad actions may be of somewhat limited effectiveness, even when it is relatively cheap to provide. Consider for instance the case where agents are concerned only about their reputation with respect to v_a (more generally, $\omega_y = 0$); as x becomes large, the fixed-point equation (24) yields

$$\rho(y,x) \approx \left(\frac{\sigma_a^2 + y\sigma_{ay}}{k^2\omega_a^2}\right)^{1/3} x^{-2/3}. \tag{25}$$

Since the aggregate social benefit from publicity $\bar{\mu}_a x\rho(y,x)$ grows with x only as $x^{1/3}$, it will be optimal to provide only a finite level even when x can be increased at a constant marginal cost, or even a marginal cost that declines slower than $x^{-2/3}$.[25]

A policy by the government or other sponsors to increase the public visibility of individuals' pro- or anti-social behavior is thus, in sense, self-limiting. In Section V.B.2 we shall demonstrate how a similar phenomenon may lead individuals who have the option to publicly disclose their good deeds to refrain from doing so, for fear of appearing driven by personal vanity or a quest for social image.

IV Honor, stigma, and social norms

In this section we derive further results on the interactions between the intrinsic, extrinsic and reputational motives for prosocial behavior. We first present an alternative mechanism through which rewards may backfire, due the presence of costs of participation; we also examine contributions in kind versus cash. We next turn to the issue of *norms,* explaining how multiple standards of socially or personally "acceptable" behavior can be sustained, and what characteristics of the "market" facilitate or impede their emergence.[26] We revert from here on to the case of discrete actions, $A = \{0,1\}$, in which the notions of honor and stigma are most sharply apparent.

A Signal reversal and crowding out

We analyze here settings in which as the reward grows, participation may come to be interpreted as a signal of greed rather than one of high-mindedness, creating a crowding-out effect.

a) Participation involves an income opportunity cost

[25] On the other hand there cannot be complete crowding out, namely $x\rho(y,x)$ actually decreasing with x : otherwise, by (16) and (24) $\rho(y,x)$ would be increasing in x, a contradiction.

[26] The source of both sets of results differs from the "signal-garbling" effect of the previous section, as there is no uncertainty on v_a in the first case and on v_y in the second.

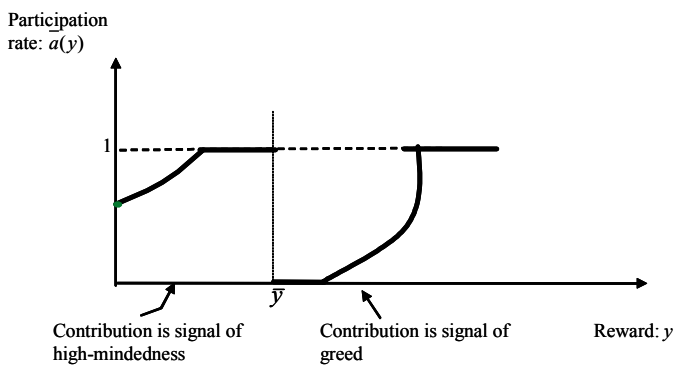

Figure 2: crowding out through signal reversal ($v_y \backsim U[0,1]$, $-\bar{y} < c_a - v_a - \mu_y/2 < 0 < c_a - v_a + \mu_y/2$)

Suppose that v_a is known, while v_y is not. Furthermore, contributing –volunteering, helping, voting, behaving honestly– entails an opportunity cost (foregone earnings or opportunistic gains) with monetary value $\bar{y}$. An agent then participates if and only if

$$v_a - c_a + v_y (y - \bar{y}) + R(y) \geq 0, \tag{26}$$

where $R(y) = -\mu_y [E(v_y|1, y) - E(v_y|0, y)]$. We are thus led to consider two cases:

– *Low-reward condition*: for $y < \bar{y}$, greedy types do not contribute; only those with v_y *below* a cutoff v_y^* do, where v_y^* (when interior) is given by

$$v_a - c_a + v_y^* (y - \bar{y}) + \mu_y \left[\mathcal{M}^+ \left(v_y^*\right) - \mathcal{M}^- \left(v_y^*\right)\right] = 0.$$

– *High-reward condition*: for $y > \bar{y}$, participation is a signal of greed; only those with v_y *above* a cutoff v_y^* do it, where v_y^* (when interior) is now given by:

$$v_a - c_a + v_y^* (y - \bar{y}) - \mu_y \left[\mathcal{M}^+ \left(v_y^*\right) - \mathcal{M}^- \left(v_y^*\right)\right] = 0.$$

Figure 2 describes the participation rate as a function of the reward for a uniform density of v_y on $[0,1]$, under which $\mathcal{M}^+ - \mathcal{M}^- \equiv 1/2$. Participation jumps down at $y = \bar{y}$ (crowding out): at that point, it becomes a "profitable" activity and therefore switches from being a source of reputational gain to a source of loss. The supply curve rises again with y beyond this threshold, and a high enough reward ultimately attracts the whole population.

Proposition 4 ***(signal reversal).*** *Opportunity costs may induce signal reversal and a concomitant crowding out of participation by monetary incentives.*

Remark (perfect negative correlation). Suppose now that v_a and v_y are unknown but perfectly

negatively correlated, so that a public-spirited individual is also disinterested: $v_a = \kappa - \bar{y}v_y$. Denoting the slope as $\bar{y}$ facilitates the comparison with the previous case. Indeed, given a reputational benefit $R(y)$,an agent participates if and only if

$$v_a - c_a + v_y y + R(y) = \kappa - c_a + v_y (y - \bar{y}) + R(y) \geq 0, \tag{27}$$

so that this case is mathematically identical to the opportunity-cost model just analyzed.

b) Cash donations versus volunteering

Let agents now differ not only in their valuation v_y for money, but also in their opportunity costs ($\bar{y}_i$ for individual i). We assume that the same reward y is offered to everyone, as the sponsor cannot discriminate and offer different payments to different people. The target audience for signaling (self, family, friends, colleagues,...), by contrast, knows the individual's opportunity cost. We focus, for simplicity, on the case where $\bar{y}_i$ and v_y are independent; in practice they may be correlated, for instance through the individual's level of income. The nature of the participation decision is depicted in Figure 3.

Proposition 5 ***(volunteering).***
(i) Individuals with a high opportunity cost of time may volunteer more than others with the same valuation for the public good but a lower opportunity cost of their time.
(ii) (Wrong currency). When faced with the choice of whether to contribute in cash or in kind (volunteering), an individual with a low hourly wage may contribute in cash, while another with a higher hourly wage may volunteer time.

Proof: (i) Take two individuals, i and j, such that $\bar{y}_i < y < \bar{y}_j$. Participating is a signal of greed for i and of lack of greed for j. So from Figure 2 we know that individual j will participate more than individual i (i.e., for a wider range of v_y's), at least if $\bar{y}_i$ and $\bar{y}_j$ are close enough to y.

(ii) Individuals are now given the choice between contributing y in cash and volunteering $b = 1$ unit of time for the cause, with opportunity cost $\bar{y}_i$ for agent i; the payoffs are $v_a - c_a - v_y y - \mu_y E(v_y | a = 1, y)$ and $v_a - c_a - v_y \bar{y}_i - \mu_y E(v_y | b = 1, y)$ respectively. If $\bar{y}_i > y$, then as the value for money grows the individual prefers first contributing in kind, then in cash, and last not contributing at all.[27] Similarly, for $\bar{y}_j$ just below y, individuals with wage $\bar{y}_j$ contribute cash. ■

c) Choice of currency

Some rewards are non-monetary and therefore would seem to be highly inefficient.[28] Our basic model already captures an obvious first rationale for such rewards: to the extent that these

[27] The second interval is degenerate (no one in group i contributes in cash) if and only if $v_y^i (\bar{y}_i - y) \leq \mu_y \left[v_y^i - \mathcal{M}^- \left(v_y^i\right)\right]$, where $v_a - c_a - v_y^i \bar{y}_i + \mu_y \left[\mathcal{M}^+ \left(v_y^i\right) - \mathcal{M}^- \left(v_y^i\right)\right] = 0$ defines the cut-off v_y^i for participation in the group with wage $\bar{y}_i$.

[28] Other arguments have been proposed recently to explain the existence of non-monetary transfers in other contexts. Prendergast and Stole (2001) explain why gifts are usually in kind rather than in cash with a model in

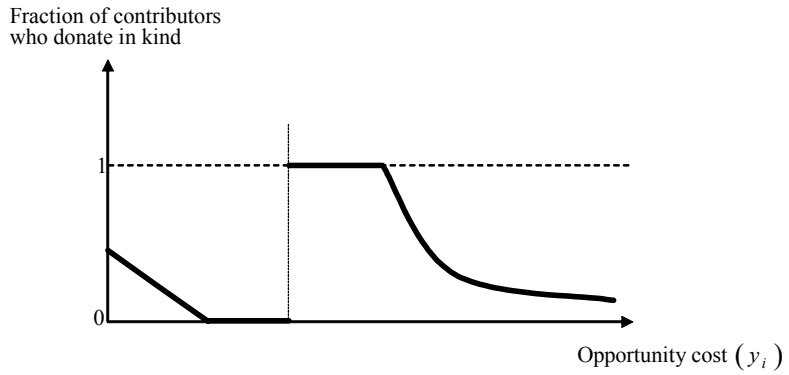

Figure 3: contributing in kind and in cash

are tied to a particular deed, they are more memorable for the self and more revealing to others than money, which is highly fungible. Proposition 5 provides a second motive: to signal their commitment, individuals with a high opportunity cost of their time may do voluntary work, while those with a lower opportunity cost may contribute cash.

"Inefficient currencies" may also be desirable for sponsors because they are valued differently by different types. For example, a teaching award or a nice dinner with students is worth more to a committed teacher. The opportunity to attend special events at the opera or mingle with artists has a higher value to a true friend of the arts (see Busraschi and Cornelli (2002) for related evidence). Screening agents through the use of such non-monetary rewards can be particularly valuable when the sponsor is in search of high-quality participation (see Section V.A).

B 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.[29]

which donors differ in their degree of certainty about what the recipient prefers. Provided that donors suffer when giving a bad gift, those with high certainty can signal themselves by taking the risk of giving in kind. In Friebel and Guriev (2002), firms that are local monopsonists in the labor market pay their workers in kind in order to prevent the latter from accumulating enough money and moving (or threatening to move).

[29] The socially determined nature of norms is heavily emphasized by psychologists (e.g. Batson (1998)) and sociologists. Multiple equilibria also arise from (very different) reputational concerns in Bernheim's (1994) model of conformity.

We show here how complementarities 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 aggregate donations, 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. For simplicity we focus here on the case where v_y is known ($v_y \equiv 1$), while v_a is distributed on some interval $[v_a^-, v_a^+]$. An agent then participates if

$$v_a + y - c_a + \mu_a \left[E\left(v_a|1, y\right) - E\left(v_a|0, y\right)\right] \geq 0,$$

meaning that v_a is above a threshold defined by comparing the net cost of participation, $c_a - y$, with the same function $\Psi(v_a)$ as in Section II,

$$\Psi(v_a) \equiv v_a + \mu_a \left[\mathcal{M}^+\left(v_a\right) - \mathcal{M}^-\left(v_a\right)\right] \equiv v_a + \mathcal{R}\left(v_a\right), \quad \text{for all } v_a \in [0,1]. \tag{28}$$

The following properties follow directly.

Proposition 6 *The equilibrium set depends on the monotonicity properties of* $\Psi\left(v\right) = v + \mathcal{R}\left(v\right)$.
i) When Ψ *is increasing, there is a unique equilibrium. If* $c_a - y \in (\Psi(v_a^-), \Psi(v_a^+))$, *the cutoff is interior and defined by* $\Psi\left(v_a^*\right) = c_a - y$; *if* $c_a - y < \Psi(v_a^-)$, *it is* $v_a^* = v_a^-$ *(full participation) and if* $c_a - y > \Psi(v_a^+)$, *it is* $v_a^* = v_a^+$ *(no participation)*
(ii) When Ψ *is decreasing and* $c_a - y \in (\Psi(v_a^+), \Psi(v_a^-))$, *there are three equilibria:* $v_a^* = v_a^-$ *(full participation),* $v_a^* = v_a^+$ *(no participation) and the interior cutoff defined by* $\Psi\left(v_a^*\right) = c_a - y$; *this last equilibrium, however, is unstable (in the usual tâtonnement sense). If* Ψ *is decreasing and* $c_a - y \notin (\Psi(v_a^+), \Psi(v_a^-))$ *there is unique equilibrium, located at a boundary.*
(iii) When Ψ *is non-monotonic, there exists a range of values of* $c_a - y$ *for which there are at least two stable equilibria, of which one at least is interior.*

We provide two examples.

a) *Upward sloping supply.* Let v_a be uniform on $[0,1]$. Then $\Psi(v_a) \equiv v_a + \mu_a/2$, so the supply curve $\bar{a}\left(y\right) \equiv \Pr\left(v_a \geq v_a^*(y)\right)$ is a familiar, upward-sloping one, illustrated in Figure 4a.

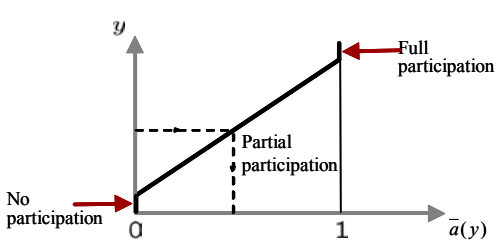

Figure 4a: unique equilbrium

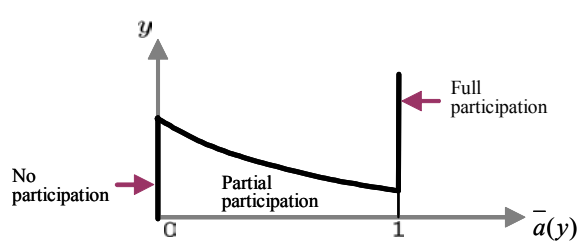

Figure 4b: multiple equilbria

b) *Multiple equilibria.* Let v_a be distributed on $[0,1]$ density $g(v_a) = 2v_a$. Then $\Psi(v_a) \equiv v_a + (2\mu_a/3)(1+v_a)^{-1}$ is decreasing for $\mu_a > 6$, resulting in three equilibria as in Figure 4b. For $\mu_a \in (3/2, 6)$, Ψ is hump-shaped, making the high-participation equilibrium interior.

As explained below, the general intuition for these results is that $\Psi' < 0$, or, equivalently, $\mathcal{R}' < -1$, corresponds to a (strong) form of strategic complementarity.

C Sources of strategic complementarity

In what follows, we maintain the assumption that v_y is known but v_a is not.

Definition *Participation decisions exhibit strategic complementarities if* $\mathcal{R}'(v_a) < 0$ *for all* v_a.

When $\mathcal{R}' = \mu_a(\mathcal{M}^+ - \mathcal{M}^-)' < 0$, a wider participation ($dv_a < 0$) worsens the pool of abstainers more than that of contributors, so that the *stigma* from abstention $\mathcal{M}^-(v_a) - \bar{v}_a$ *rises faster than the honor* from participation $\mathcal{M}^+(v_a) - \bar{v}_a$ fades. When $\mathcal{R}' < -1$, the increase in the net 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 4b. When $\mathcal{R}' \in (-1, 0)$ 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, $\mathcal{R}' > 0$.

Equipped with these results and intuitions, we now investigate the main factors that make strategic complementarity –and thus the existence of socially determined norms– more likely.

C.1 Increasing density

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. This intuition is confirmed by the (more general) next two propositions.

Proposition 7 ***(Jewitt 2004).*** *If the distribution of* v_a *has a density which (on its support) is (a) decreasing, (b) increasing, (c) unimodal, then* $\mathcal{M}^+(v_a) - \mathcal{M}^-(v_a)$ *is respectively (a) increasing, (b) decreasing, (c) quasi convex.*

Proposition 7 provides a sufficient condition for the monotonicity of $\mathcal{M}^+ - \mathcal{M}^-$. What ultimately matters for uniqueness or multiplicity and the slope of of the supply curve, however, is the behavior of $\Psi(v_a) = v_a + \mu_a[\mathcal{M}^+(v_a) - \mathcal{M}^-(v_a)]$. So the second requirement for multiple equilibria is that reputational concerns be strong enough: μ_a must be sufficiently high.

The following proposition provides a sufficient condition (weaker than (a) above) for supply to be uniquely defined and increasing in price. No simple analogue is available for the converse case.

Proposition 8 *If the distribution of v_a has a log-concave density g, or more generally a log-concave distribution function G, then for all $\mu_a \in [0,1]$ the supply function is everywhere upward-sloping.*

Proof: We can write:

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

and observe that both $\mathcal{M}^+$and $\mathcal{M}^-$ are increasing functions, while

$$v_a - \mathcal{M}^-(v_a) = v_a - E(\tilde{v}_a | \tilde{v}_a < v_a) = \frac{\int_{-\infty}^{v_a} G(v)\, dv}{G(v_a)}$$

is also increasing in v_a, provided the integral of G is log-concave. As shown for instance in Caplin and Nalebuff (1991), log-concavity is preserved by integration over convex sets, so it suffices that G itself be log-concave (g/G decreasing, which is weaker than g decreasing as in Proposition 7(a) above). In turn, a sufficient condition for this is that g be log-concave. ■

C.2 Excuses and forced participation

We have so far assumed that observers (future "self", other agents) know for sure that the individual had an opportunity to contribute. This is often not the case. On one hand, the individual may have faced (unobserved or imperfectly remembered) circumstances that *precluded participation:* not being informed, having to deal with some emergency, etc. By lessening the stigma from abstention, such excuses will tend to inhibit the emergence of strategic complementarities. Conversely, with some probability a participating agent may have done the right thing for reasons other than public-mindedness: the opportunity cost could have been unusually low, or strong social pressure or extrinsic incentives may have induced *forced behavior.* By tarnishing the "distinction" from performing the prosocial action, this will facilitate the emergence of strategic complementarities. We look at these two possibilities in sequence.

- *Involuntary non-participation*

Suppose that with probability $\delta \in [0,1]$, an individual is unable to participate. For a given potential cutoff v_a the information conveyed by participation is unchanged, while that conveyed by non-participation becomes less damaging. Specifically,

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

Note that if $(\mathcal{M}^P - \mathcal{M}^{NP})'(v_a;\delta) > 0$, this expression is also positive for all $\delta' > \delta$, since

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

and the last term is clearly decreasing in v_a.

- *Forced participation*

Conversely, suppose that with probability $\delta \in [0,1]$, an individual is forced to contribute. The information conveyed by non-participation is unchanged, but the positive signal conveyed by participation is dulled:

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

It can similarly be shown that if $\left(\mathcal{M}^{P}-\mathcal{M}^{NP}\right)'\left(v_a;\delta\right)<0$ it is also negative for all $\delta'>\delta$.

Proposition 9 *An increase in the probability of (unobserved) forced participation facilitates the emergence of strategic complementarities, whereas an increase in the probability of (unobserved) involuntary non-participation inhibits it.*

A related set of factors involve the *observability* to others and *memorability* to himself of the agent's actions. If participation is observed but non-participation can go undetected (or be forgotten) with some probability δ or, conversely, if antisocial behavior is detected for sure while a good deed may go unnoticed with some probability δ', this will lead to effects qualitatively similar to those just analyzed. We already discussed in Section IV.A the role of peers and kin in raising the visibility and memorability of both pro and anti-social deeds. Combining this observation with the above results makes it easy to understand the emergence of group-specific norms of social responsibility.

V Equilibrium contracts

A Sponsor's choice of reward

Sponsors, whether private or public, derive benefits from agents' participation but face resource costs in offering rewards. Let B denote the ratio between the monetary value of the benefit that participation by each agent confers to a sponsor, and the latter's opportunity cost of funds. A monopoly sponsor thus solves (possibly after an equilibrium selection, as in Figure 4b)

$$\max_{y}\left\{(B-y)\,\bar{a}(y)\right\},$$

where $\bar{a}(y)=\int_{\mathbb{R}^4}a(\mathbf{v},\boldsymbol{\mu};\ y)f\left(\mathbf{v}\right)m(\boldsymbol{\mu})\ d\mathbf{v}\ d\boldsymbol{\mu}$ is the aggregate supply response to an offer of y, computed in previous sections under alternative assumptions on the distribution of agents'

types $(\mathbf{v}, \boldsymbol{\mu})$. Naturally, rewards that lead to net crowding out, $\bar{a}'(y) < 0$, are never optimal for the sponsor. A more surprising result is the following one.

Proposition 10 *Let v_a be unknown, $v_y \equiv 1$, and assume that $\Psi' > 0$. A monopoly sponsor may offer contributors a reward that is too high from the point of view of social welfare.*

Proof: see the appendix.

The normalization $v_y = 1$ allows us here to add up the sponsor's profit and individuals' surpluses to obtain aggregate welfare, while $\Psi' > 0$ yields a unique cutoff $v_a^*(y)$ and upward-sloping supply function $\bar{a}(y) = 1 - G(v_a^*(y))$. To understand the intuition for the result, consider the effect of a marginal decline in the reward from the monopolist's preferred level. By definition, this has a negligible (first-order) effect on his profit, so the welfare impact falls on the agents. On the one hand, the $\bar{a}(y)$ inframarginal contributors receive lower rewards. On the other hand, $g(v_a^*(y)(-v_a^{*\prime}(y)) = g(v_a^*(y))/\Psi'(v_a^*(y)$ individuals at the margin stop contributing, and in doing so they bring up the reputation (the average v_a) of both contributors and non-contributors. By the martingale property, these gains corresponds to the marginal agents' reputational loss, which is $\mu_a \left[\mathcal{M}^+\left(v_a^*\right) - \mathcal{M}^-\left(v_a^*\right)\right]$ per person. Thus, when

$$\frac{\mu_a g\left(v_a^*\right)}{1 - G\left(v_a^*\right)} \left(\frac{\mathcal{M}^+\left(v_a^*\right) - \mathcal{M}^-\left(v_a^*\right)}{\Psi'\left(v_a^*\right)} \right) > 1, \tag{29}$$

where v_a^* is evaluated at the monopolist's optimum, the reputational externality dominates the transfer effect, and offering less generous rewards would increase aggregate welfare.

- *Quality of participation.*

Sponsors often care about "high-quality" participation, not just total enrollment. This arises when actual participation is an open-ended contract, subject to adverse selection or moral hazard. Thus, one argument for low pay to the military or no pay for volunteers is that one wants to select, respectively, patriots and do-gooders, rather than people whose main loyalty is to money. Implicitly, there is an opportunity cost in "recruiting" an agent —either the use of some complementary capital, or the risk that the might end up harming the principal's objectives (e.g., mercenaries finding out that the enemy pays better). Similarly, it is often argued that not paying people for blood reduces the fraction of donors with hepatitis.

For instance, if we introduce a hidden action (beyond $a \in A$, which is observed) whose marginal cost to the individual decreases with v_a, we obtain, in reduced form, a benefit for the sponsor $B\left(v_a\right)$ with $B' > 0$ (more generally, B could depend on all of $(\mathbf{v}, \boldsymbol{\mu})$). The theory is then the same, with the sponsor now solving:

$$\max_y \left\{ \int_{\mathbb{R}^4} \left(B\left(v_a\right) - y\right) a(\mathbf{v}, \boldsymbol{\mu};\ y) f\left(\mathbf{v}\right) m(\boldsymbol{\mu})\ d\mathbf{v}\ d\boldsymbol{\mu} \right\}.$$

Building on this basic setup, we will now examine a number of issues concerning the nature of equilibrium contracts and resulting participation rates. Until Section V.B, we shall assume that there is no variability in $\boldsymbol{\mu}$.

- *Menus.*

We have assumed that the sponsor makes a single offer. There is clearly no point in offering multiple rewards if the audience the agent signals to does not observe which one is selected, or if side-contracting between the sponsor and the agent is feasible. Ignoring this, would a menu be desirable for the sponsor?

Proposition 11 ***(menus)***. *(i) Suppose that an individual's marginal valuation v_a is unknown, whereas v_y is known and common to all agents. Then, menus cannot benefit the sponsor.*
(ii) Suppose that an individual's marginal utility for money v_y is unknown, whereas v_a is known and common to all agents. Assume also that the distribution of v_y satisfies the monotone hazard rate property: $h(v_y)/(1-H(v_y))$ is increasing in v_y. Then a monopoly sponsor finds it optimal to separate the types who contribute, using a menu with a continuum of rewards.

Proof: (i) Suppose that v_a is unknown and consider any menu of rewards $\mathcal{Y}$. Conditionally on contributing, the agent chooses $y \in \mathcal{Y}$ so as to maximize $v_a - c_a + v_y y + \mu_a E\left(v_a | a, y\right)$. The optimal choice is independent of v_a, so the sponsor cannot screen the agent.

(ii) The proof that it is always optimal for the sponsor to separate types when v_y is unknown is provided in the appendix. Consider therefore a fully separating menu $\mathcal{Y} = \{Y\left(\tilde{v}_y\right)\}$, where $\tilde{v}_y$ is the agent's announcement of his type and Y some strictly monotonic function. The agent then chooses $\tilde{v}_y$ so as to maximize $v_a - c_a + v_y Y(\tilde{v}_y) - \mu_y \tilde{v}_y$. Taking derivatives and using the truthtelling condition ($\tilde{v}_y = v_y$ in equilibrium) easily yields

$$Y(v_y) = \mu_y \log v_y + \text{constant}. \tag{30}$$

Intuitively, a greedy agent has a high marginal rate of substitution between money and reputation, so it makes sense for the sponsor to reward such types with money (associated with a calamitous image) and a less greedy ones with a better reputation (and less money). ■

- *Should the fee remain confidential?*

We have assumed that the fee y is public. What would confidentiality imply for sponsors and agents? To examine this question, we use the same model with unknown v_a's, and $v_y \equiv 1$, and assume $\mathcal{R}' > -1$ (or $\Psi' > 0$, see Proposition 6) to avoid a multiplicity of equilibrium participation rates. We consider a sponsor who can commit to one of two policies: *confidentiality* (C), under which only the agent knows the level of y offered (but participation is publicly observable), or *public disclosure* (D). In both cases we assume that the sponsor's objective function is quasiconcave in y.

Confidentiality. Under confidentiality, the target audience rationally expects a fee and a cutoff $\left(y^C, v_a^C\right)$ satisfying $v_a^C - c_a + y^C + \mathcal{R}\left(v_a^C\right) \equiv 0$. If the sponsor secretly deviates and offers y, agents with $v_a \geq c_a - y - \mathcal{R}\left(v_a^C\right)$ will contribute, since the reputational impact of action or inaction remains unchanged. The sponsor therefore faces the *ex-post* supply curve

$$\bar{a}_C(y) = 1 - G\left(c_a - y - \mathcal{R}\left(v_a^C\right)\right), \tag{31}$$

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 y that is selected, agents with v_a above the cutoff $v_a^*(y)$ defined by $v_a^*(y) - c_a + y + \mathcal{R}\left(v_a^*(y)\right) \equiv 0$ contribute, so the sponsor faces the *ex-ante* supply curve

$$\bar{a}_D(y) = 1 - G\left(c_a - y - \mathcal{R}\left(v_a^*(y)\right)\right) \tag{32}$$

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 12 *(i) It is optimal for the sponsor to publicly disclose the fee.*
(ii) With strategic complements $(\mathcal{R}' < 0)$ *the sponsor offers a higher fee and elicits a higher participation under disclosure than under public confidentiality. The reverse holds for strategic substitutes* $(\mathcal{R}' > 0)$.
(iii) The optimal reward under disclosure y^D *is immune to secret renegotiation between the sponsor and the agent when* $\mathcal{R}' < 0$. *By contrast, when* $\mathcal{R}' > 0$, *the equilibrium reward when secret renegotiation is feasible is* y^C.

Proof: see the appendix.

Intuitively, 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 participants would not agree to a secret lowering of the fee, so this effect is renegociation-proof. Strategic substitutability has the converse effect on supply and thus leads to lower announced fees; but in this case both the sponsor and the participants would agree to increase them ex-post, if they could find ways of doing it secretly.

B Active signaling by the agent

It has so far been assumed that agents take their environment (reward, prominence of their actions, etc.) as given, and just decide on their level of participation. Yet there are many situations in which an individual plays a more active role in shaping his signaling problem.

First, even if the reward is set by the sponsor, the agent may be able to refuse all or part of it. Second, he may make his contribution more or less conspicuous.

B.1 Turning down rewards

An agent may be eager to participate in order to demonstrate his high-mindedness, but concerned that the reputational benefit 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 the crowding-out argument?

Note first that the issue may just not arise if give-backs are not observable by those to whom the agent is trying to signal, or if the sponsor can reward the agent secretly. As shown earlier, when $\mathcal{R}' < 0$ the principal and the agent may indeed collude ex-post to raise the reward above what was publicly announced.

Suppose now that the realized transfer from the sponsor to the agent is effectively observed by others. 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 all or any part of the reward, or even giving money to the sponsor: the loss of monetary income, $v_y\,(y - y')$, and the net reputational benefit, $R(y') - R(y)$, are both type independent.

Proposition 13 *The equilibria studied in Sections II.B and II.C are still equilibria of the enlarged game in which the individual can turn down part or all of the reward, if either transfers from the sponsor are secret or if the uncertainty is about v_a.*[30]

By contrast, when the uncertainty is (also) about v_y, turning down the reward or part of it could be used to signal the absence of greed. Yet even in this case, as we shall now see, it may be that, in equilibrium, all agents either just accept y or do not participate, but *never turn down some of the reward.* The intuition is that doing so could lead the audience to question the 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 III.B, and will recur again when examining private disclosure in the next subsection.

[30] It can also be verified that these equilibria satisfy the Cho-Kreps (1987) Never-a-Weak-Best-Response (NWBR) criterion. According to this criterion, one eliminates a type $\mathbf{v}$ as a possible source of an off-the-equilibrium-path offer y' if the set of inferences that make this type weakly better off than his equilibrium outcome is strictly included in the union of similar sets for other types $\mathbf{v}'$.

We thus return to the case where two-dimensional uncertainty about $\mathbf{v} = (v_a, v_y)$ combines with uncertainty about agents' degree of image-consciousness $\boldsymbol{\mu} = (\mu_a, \mu_y)$. The reputational gross benefit when choosing action an $a \in \{0, 1\}$ and a reward $y' \leq y$ is thus

$$\mu_a E\left(v_a | a, y'\right) - \mu_y E\left(v_y | a, y'\right).$$

Because the insight we seek to capture is very intuitive, yet its analysis is rather technical, we shall limit ourselves to a straightforward illustration. Suppose that $(\mu_a, \mu_y) = z(\gamma_a, \gamma_y)$, where (γ_a, γ_y) is fixed and thus known to the audience, whereas the preference parameter z is independently distributed from (v_a, v_y) and takes one of two extreme values: the agent is either *image indifferent* ($z = 0$) or *image driven* ($z = +\infty$). Image-indifferent individuals participate if and only if $v_a - c_a + v_y y \geq 0$, whereas image-driven ones have lexicographic preferences: they first maximize their reputation, then, for a given reputation, choose the action that maximizes their current payoff. Finally, 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, for example).

Clearly, participating and turning down the reward (or part of it) is a strictly dominated strategy for image-indifferent individuals. As to image-driven individuals, they put no weight on current payoffs relative to reputational ones, and therefore all pool on the action(s) that yield(s) the highest reputation. 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 reputations would correspond to the prior mean of the distribution of (v_a, v_y).[31] 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 $v_a - c_a + v_y y \geq 0$.

Proposition 14 *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 noting 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. In practice, one often has a negative reaction when discovering that an acquaintance is highly image conscious. There are two possible reasons. First, one may start questioning the motivation behind the person's good deeds; this is the route taken here, as in Section III and in the next subsection. Second, one may have an intrinsic distaste for such a trait –e.g., vanity. We are agnostic as to whether such a direct opprobrium is warranted. After all, someone who

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

is highly image-conscious may be more reliable, with reputational concerns helping to discipline his behavior; on the other hand, such a person may spend his life "pandering", that is, doing and telling others just what they want to see or hear.[32]

B.2 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 issue we shall assume that if the 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, at a cost d –either a resource cost or a goodwill cost, as "showing off" may hurt others' self-esteem or make him look inconsiderate.

We again allow people to be heterogenous in two respects: first, they differ in their valuation v_a for the public good, whereas $v_y \equiv 1$ is known. Second, a fraction θ have a high value for reputation γ_a^H and a fraction $1-\theta$ a lower value γ_a^L. In the absence of disclosure, the unit reputational gains are, as earlier, $\mu_a^L \equiv x\gamma_a^L$ and $\mu_a^H \equiv x\gamma_a^H$.

The timing is as follows: i) the sponsor sets the reward y; ii) each individual chooses whether to participate; iii) in case he does, he can disclose it; iv) if he participated but did not disclose, others learn of it with probability x. Throughout this section we will assume that all relevant supply curves are uniquely defined and upward sloping ($\Psi' > 0$, for the relevant Ψ), so as to avoid equilibrium multiplicity. We examine and compare disclosure in two situations:

(a) Symmetric information about image consciousness

We first consider the case in which $\gamma_a^H = \gamma_a^L = \gamma_a$ is known, while v_a is not. In addition to serving as a natural benchmark it is also interesting in its own right, as we shall see that there are strategic complementarities in disclosure itself. An *equilibrium with disclosure* is defined by a cutoff v_a^D and the following equations:[33]

$$\Psi_D(v_a^D) \equiv v_a^D + \gamma_a \left[\mathcal{M}^+\left(v_a^D\right) - \mathcal{M}^-\left(v_a^D\right)\right] = c_a + d - y \tag{33}$$

and

$$\gamma_a (1-x) \left[\mathcal{M}^+\left(v_a^D\right) - \mathcal{M}^-\left(v_a^D\right)\right] \geq d. \tag{34}$$

An *equilibrium without disclosure* is defined by a cutoff v_a^N and the following equations:

[32] In the same way that politicians with strong re-election concerns follow policies that they know to have detrimental consequences, but are popular with the electorate (see Maskin and Tirole 2004).

[33] To obtain (33), we assume that in the off-the-equilibrium-path event in which the agent does not disclose but is found out to have contributed, the audience attributes to him an expected type $\mathcal{M}^+\left(v_a^D\right)$; this is for example what would happen if the disclosure technology were not perfect, so that the audience learned about participation only with probability $1-\varepsilon$, for ε small.

$$\Psi_N(v_a^N) \equiv v_a^N + \gamma_a x \left[\mathcal{M}^+\left(v_a^N\right) - E(v_a \,|\, \phi, v_a^N)\right] = c_a - y, \tag{35}$$

and

$$\gamma_a(1-x)[\mathcal{M}^+(v_a^N) - E(v_a \,|\, \phi, v_a^N)] \leq d, \tag{36}$$

where

$$E(v_a|\phi, v_a^N) = \frac{\int_0^{v_a^N} v\, g(v)\, dv + (1-x)\int_{v_a^N}^{\infty} v\, g(v)\, dv}{G(v_a^N) + (1-x)[1 - G(v_a^N)].} > \mathcal{M}^-\left(v_a^N\right).$$

Proposition 15 ***(disclosure when image-consciousness is known)****. Let the functions Ψ_D and Ψ_N in (33)-(34) be increasing. When γ_a is common knowledge,*

(i)There exists γ_a^* and γ_a^{**}, with $0 < \gamma_a^* < \gamma_a^{**}$, such that for $\gamma_a < \gamma_a^*$, the agent never discloses his contribution, for $\gamma_a > \gamma_a^{**}$ he always discloses, and for $\gamma_a^* \leq \gamma_a \leq \gamma_a^{**}$ there exist multiple norms: both disclosure and non-disclosure are equilibrium behaviors.

(ii) Where multiple norms coexist, there is more participation in the disclosure equilibrium.

Proof: (i) Because Ψ_D is increasing, (33) implies that v_a^D is a decreasing function of γ_a. Rewriting (34) as $(1-x)\left(d + c_a - v_a^D - y\right) \geq d$ then shows that disclosure is an equilibrium behavior when γ_a exceeds some threshold. A similar reasoning applies for non-disclosure equilibria.

Next, let v_a^* and v_a^{**} denote the two valuation cutoffs for the reputation types γ_a^* and γ_a^{**} respectively. Using equations (33) through (36), with (34) and (36) satisfied with equality at $\gamma_a = \gamma_a^*$ and γ_a^{**} respectively, one obtains $v_a^* = v_a^{**}$ and

$$\gamma_a^*\left[\mathcal{M}^+(v_a^*) - \mathcal{M}^-(v_a^*)\right] = \gamma_a^{**}[\mathcal{M}^+(v_a^{**}) - E(v_a \mid \phi, v_a^N)],$$

and so $\gamma_a^{**} > \gamma_a^*$.

(ii) This results from the fact that $\mathcal{M}^+(v_a^D) - \mathcal{M}^-(v_a^D) > x[\mathcal{M}^+(v_a^D) - E(v_a \mid \phi, v_a^N)]$. ■

Intuitively, the absence of information about an agent's contribution carries a *lower stigma* if contributors do not disclose than if they do, which reduces the incentive to disclose. Hence, the existence of multiple norms. Furthermore, disclosure encourages participation, through both the increased probability that good deeds will not go unnoticed and the higher stigma attached to the absence of information.

(b) Asymmetric information about image-consciousness

Let us now assume that γ_a, like v_a, is private information, and show that even if there is no social opprobrium on image-consciousness this may reduce disclosure, which now itself carries a stigma. The idea is that since the people most prone to let others know about their good deeds are those with a high concern for self-image, disclosure of a prosocial act makes it more likely that the act was motivated less by genuine public-mindedness (a high v_a) than by image-seeking

(a high γ_a). Formally, suppose that it is an equilibrium under symmetric information for type γ_a^L not to disclose and for type γ_a^H to disclose; we will show that asymmetric information about γ_a may lead to neither type disclosing.[34]

Let $\hat{v}_a^L$ and $\hat{v}_a^H$ denote the valuation cutoffs under *symmetric* information associated (in the equilibrium under consideration) to γ_a^L and γ_a^H respectively. It must then be that $\hat{v}_a^H < \hat{v}_a^L$, for two reasons: type γ_a^L does not disclose, and furthermore he has a lower reputational gain.

Is such separation (with respect to γ_a) still an equilibrium behavior under asymmetric information? In a separating equilibrium, type γ_a^H participates if and only if v_a exceeds some cutoff v_a^H, and type γ_a^L participates if and only if v_a is above some v_a^L. Therefore, 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 the updated reputation in the absence of information is

$$E\left(v_a \,\middle|\, \phi, v_a^H, v_a^L\right) \equiv \frac{\theta \int_0^{v_a^H} v\, g(v)\, dv + (1-\theta)\left[\int_0^{v_a^L} v\, g(v)\, dv + (1-x)\int_{v_a^L}^{\infty} v g(v) dv\right]}{\theta G(v_a^H) + (1-\theta)[G(v_a^L) + (1-x)(1-G(v_a^L)]}. \tag{37}$$

Proposition 16 *Under asymmetric information about the extent of image-consciousness:*
(i) In a separating equilibrium where the more image-conscious type discloses while the less image-conscious one does 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: $v_a^H < v_a^L$.
(ii) Asymmetric information about the extent of image-consciousness may reduce disclosure: for some range of values d, *the* γ_a^H *type no longer discloses when* γ_a *is unobservable.*

Proof: see the appendix.

VI "Holier than Thou" sponsor competition

While sponsors of prosocial behavior may be monopolists for various reasons (legal, technical, first-mover and visibility advantages),[35] there is often competition. Local government agencies and charities compete for volunteering time, NGO's and foundations compete for donations, universities compete for speakers and visiting committee members, and so forth. Last but not least, religions compete for believers. We therefore now investigate the impact of market structure on equilibrium rewards and welfare.

[34] In a very different context, Sadowski (2004) presents a model in which an "overeagerness" to engage in costly signaling conveys bad news about the agent (revealing that he has few outside opportunities), so that high-ability types may even end up signaling less than low-ability ones.

[35] For instance, blood collection is often centralized.

A The reversal of Bertrand competition

The first insight is that the standard logic according to which Bertrand competition leads to "undercutting", or more precisely overbidding in our context where sponsors purchase a service, may well be reversed: when a rival offers y, an offer of $y-\varepsilon$ (and *not* $y+\varepsilon$) attracts everyone away from him. Indeed, the monetary loss for the agents is small, whereas the reputational gain from joining the group that selects $y-\varepsilon$ rather than the definitely more greedy group that opts for y is substantial.[36] This mechanism, in turn, explains why sponsor competition does not lead to a situation where volunteers are well compensated for their time, or most of the monies given to charities and the arts dissipated on perks or special events for the donors. Because rewards tend to be *bid down, not up,* sponsors retain a significant share of the surplus.

We first verify in Proposition 17 below that (under a reasonable assumption) if two offers –made by two distinct sponsors or a single one– are close to each other, the higher offer attracts no one. Because uncertainty about v_a only does not allow any segmentation of demand, we assume here that v_a is known, while v_y is not.

Lemma 1 *(monotonicity by intervals): Let v_a be fixed. If n offers, $y^1 < y^2 < \cdots < y^n$, are accepted in equilibrium with positive probability, then there exist $0 \leq v_y^1 < v_y^2 < \cdots < v_y^n$ such that types in $\left[0, v_y^1\right)$ do not participate, while types in $\left[v_y^i\ ,\ v_y^{i+1}\right)$ select reward y^i.*

Proof: An agent's payoff is $v_a - c_a + v_y y^i - \mu_y E(v_y|1, y^i)$ when selecting y^i, versus $-\mu_y E(v_y|0)$ when choosing not to participate. Standard revealed preference implies that a higher v_y must choose a (weakly) higher y^i; hence the property of monotonicity by intervals. ■

In view of Lemma 1, it seems reasonable to require monotonicity of beliefs off-the-equilibrium path as well:

Assumption A: Types self-select monotonically, both on and off the equilibrium path: if $\left\{y^1, \cdots, y^m\right\}$ denotes the ordered set of offers faced by agents, there exist $0 \leq v_y^1 \leq v_y^2 < \cdots \leq v_y^{m+1}$ such that non-participation (respectively, acceptance of y^i) leads to posterior beliefs equal to the prior truncated to the (possibly degenerate) interval $\left[0, v_y^1\right]$ (respectively, $\left[v_y^i, v_y^{i+1}\right]$).

Proposition 17 *(i) For any $\eta > 0$, there exist $\varepsilon > 0$ such that if offer y^i attracts a fraction of agents at least equal to η, then any offer $y^j \in \left(y^i, y^i + \varepsilon\right]$ attracts no one.*

(ii) Symmetrically, under Assumption A, for all ε there exists $\lambda(\varepsilon)$, with $\lim_{\varepsilon \to 0} \lambda(\varepsilon) = 0$, such that no offer in $y^j \in \left(y^i,\ y^i + \varepsilon\right]$ can attract a market share greater than $\lambda(\varepsilon)$ (regardless of whether offer y^i attracts a positive market share).

[36] A related idea applies to matching marketplaces: a marketplace charging a price slightly below that of a competing marketplace may attract a substantially less attractive clientele (Damiano and Li (2003)).

Proof: To establish (i), let v_y^{i+1} denote the upper bound of the equilibrium "clientele" of offer y^i. Then $-\mu_y E(v_y|1, y^i) = -\mu_y \left(v_y^{i+1} - \eta'\right)$ for some η'. Suppose that offer y^j attracts types in $[v_y^j, v_y^{j+1}]$, where $v_y^j < v_y^{j+1}$ and $v_y^j \geq v_y^{i+1}$ by monotonicity. Then,

$$v_y^j \left(y^j - y^i\right) > \mu_y \left[v_y^j - \left(v_y^{i+1} - \eta'\right)\right],$$

which is impossible for $y^j - y^i$ small enough. The proof of (ii) is similar to that of (i). ■

Let us now turn to the competitive determination of rewards. The market structure could be one where there is a continuum of potential entrants facing a fixed cost $k > 0$ for operating in the market, or a finite number of large competitors (who will offer menus of rewards). We will further assume that, as in the monopoly case, the payoff to a sponsor from offering y is equal to the product of $B - y$ and the number of contributors attracted by this offer.[37]

To show that Bertrand competition in such markets need not bring all accepted equilibrium rewards to the sponsors' reservation level (B), we derive an equilibrium in which participating agents separate. Specifically, if the support of v_y is $\left[v_y^-, v_y^+\right]$, types $v_y \in \left[v_y^*, v_y^+\right]$ choose reward $y = Y(v_y)$, with $Y' > 0$ and $Y\left(v_y^+\right) = B$, while types $v_y \in \left[v_y^-, v_y^*\right]$ do not participate. The function Y must satisfy incentive compatibility:

$$Y(v_y) = \arg\max_{\tilde{y}} \left\{v_a - c_a + v_y \tilde{y} - \mu_y Y^{-1}(\tilde{y})\right\}.$$

Together with the boundary condition, this yields:

$$Y(v_y) = B + \mu_y \log\left(v_y / v_y^+\right). \tag{38}$$

Last, the cutoff type v_y^*, if interior, must be indifferent as to participation:

$$v_a - c_a + v_y^*[B + \mu_y \log\left(v_y^* / v_y^+\right)] - \mu_y \left[v_y^* - \mathcal{M}^-\left(v_y^*\right)\right] = 0.$$

Thus, a perfectly competitive market features a *continuous range of rewards* offered and accepted in equilibrium, $[B - \mu_y \log\left(v_y^+ / v_y^*\right), B]$. Depending on the posited market structure, these may be offered separately by a continuum of sponsors (with a mass $n(v_y) = (B - Y(v_y))g(v_y)/k$ offering reward v_y), or as a single menu offered by each of a finite number of large sponsors.

[37] This requires some justification, for if sponsors were motivated solely by social welfare they would be indifferent as to who attracts contributors, and the analysis would boil down to that of the monopoly case. For sponsor competition to be meaningful, foundations trying to raise money for different types of medical research, museums courting donors of paintings, or universities trying to attract speakers must put some weight on their own "local" welfare –whether because they care about different public goods, or because their administrators have career concerns in addition to philanthropic goals. This is what our specification captures.

B Welfare-reducing competition

A perhaps even more surprising result is that *competition may reduce welfare,* when sponsors can screen contributors in inefficient ways. It formalizes in particular the idea of religions and sects competing on orthodoxy, asceticism, and other costly requirements for membership (e.g., Berman (2000)). Suppose that there are no entry costs ($k = 0$) and that

a) the uncertainty is about v_a (we normalize again $v_y \equiv 1$), which equals v_a^H with probability ρ or v_a^L with probability $1 - \rho$, where $v_a^H > v_a^L$;

b) the non-monetary cost of contributing is c_a, unless the sponsor demands a "sacrifice" (which it is able to verify). The cost then becomes c_a^H for the high type and c_a^L for the low type, where

$$c_a^L > c_a^H > c_a. \tag{39}$$

A sacrifice is a pure deadweight loss, whose only benefit for the sponsor is to help screen the agent's motivation. The assumption that $c_a^H > c_a^L$ reflects the idea that such a sacrifice is less costly to a more motivated agent. For expositional simplicity, we will assume that c_a^L is so large that the low type is not willing to sacrifice.

Proposition 18 *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 total welfare.*

Proof: see the appendix.

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 by requiring costly sacrifices has two effects:

(a) it inflicts a deadweight loss $c_a^H - c_a$ 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), however, he must also compensate the low type for his reputational loss. Since reputational benefits are linear and beliefs form a martingale, these losses 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 monopoly sponsor serving the whole market will never find it profitable to require sacrifices.

Things are quite different under free entry. First, since v_y is known, an agent's choice of financial reward has no reputational consequence; therefore, price competition will drive *all sponsors to offer* B. Second, by requiring a costly sacrifice, entrants can now attract the high types away from competitors who impose no such requirement, leaving low-type (or their sponsors) with the resulting reputational loss. This "cream-skimming" leads inevitably to an equilibrium where all active sponsors offer a reward of B, with a proportion ρ of them requiring an inefficient sacrifice and serving the high-types, while the remaining $1-\rho$ require only the normal level of contribution c_a, and serve only 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 agents gain: total participation remains unchanged (both types still behave prosocially), the same is true of average reputation (by the martingale property), and rewards are pure transfers. There is now, however, a deadweight loss of $\rho(c_a^H - c_a)$, corresponding to the wasteful sacrifices made by the high-types to separate. Therefore, *competition unambiguously reduces welfare.* ■

VII 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 visibility of 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 yields 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. Our results can be organized into four main themes.

– *Rewards and punishments.* The presence of extrinsic incentives casts suspicion on the reason why prosocial actions are performed, acting like an increase in the noise-to-signal ratio or even reversing the sign of the signal. This "spoiling effect" depresses the reputational motive for good behavior, and the resulting crowding out can be so large that greater incentives actually reduce total supply. Sponsors may respond to contributors' desire to appear intrinsically motivated rather than greedy by publicly announcing low rewards, but then find it profitable to

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

offer higher ones in private, creating a commitment problem. Alternatively, contributors could themselves consider turning down (all or part of) the rewards that are offered; they may refrain from doing so, however, for fear that it would signal a high degree of high image-consciousness and thereby cast another form of doubt on the true motivation for their contribution.

– *Publicity and disclosure.* Prominence and memorability of contributions strengthen signaling concerns and thus generally encourage prosocial behavior. When individuals are heterogeneous in their image concerns, however, greater prominence also acts like an increase in the noise-to signal-ratio: good actions come to be suspected of being image-motivated, which severely limits the effectiveness of such policies. Similarly, individuals who have the option to publicly disclose their good deeds may refrain from doing so, for fear of appearing driven by personal vanity or a quest for social image.

– *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 values (the density is increasing), and when unobserved circumstances that could prevent someone from contributing (excuses) are more rare than those that make it inevitable or unusually easy. The second case applies in the reverse circumstances. When setting and publicizing their rewards, 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 competing sponsors), however, their chosen policies will generally be inefficient. Thus, even a monopoly sponsor may offer rewards that are too generous from the point of view of social welfare.

– *Competition.* In the "market" for prosocial contributions, sponsors will be lead to offer agents competing opportunities for reputationally motivated sacrifices. Thus, in price competition the best way to steal a customer away from a rival may be to offer a little less: locally, individual supply curves are again decreasing. As result, rewards will tend be bid down rather than up, leaving sponsors with a significant share of the surplus even under Bertrand competition. The same "holier than thou" form of emulation can even cause sponsor competition to reduce social welfare, by leading agents to engage in more inefficient sacrifices than they would have under a monopoly.

Appendix

Proof of Proposition 1

As observed earlier, if introducing a reward y reduces participation, then it must necessarily be that $R(y) < R(0)$. So let us assume that participation increases.

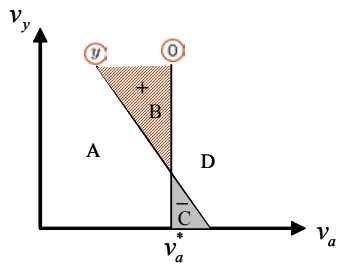

Figure 5a

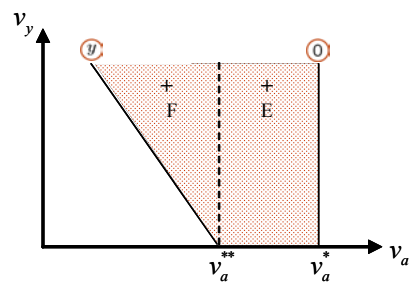

Figure 5b

(a) Suppose first that the reward y attracts new participants (area B in Figure 5a) and induces some former ones to quit (area C). Let α_A, α_B, α_C, and α_D denote the weights on each area. An increase in participation implies that $\alpha_B \geq \alpha_C$. Now, $E(v_a|a=1)$ changes,

$$\text{from} \quad \frac{\alpha_D E(v_a|D) + \alpha_C E(v_a|C)}{\alpha_D + \alpha_C} \quad \text{to} \quad \frac{\alpha_D E(v_a|D) + \alpha_B E(v_a|B)}{\alpha_D + \alpha_B},$$

which is smaller since $E(v_a|B) < \min\{E(v_a|C)\ ,\ E(v_a|D)\}$ and $\alpha_B \geq \alpha_C$.

The impact on the reputation about v_y is irrelevant when $\mu_y = 0$. More generally, $E(v_y|a=1)$ changes

$$\text{from} \quad \frac{\alpha_D E(v_y|D) + \alpha_C E(v_y|C)}{\alpha_D + \alpha_C} \quad \text{to} \quad \frac{\alpha_D E(v_y|D) + \alpha_B E(v_y|B)}{\alpha_D + \alpha_B}.$$

Given that $E(v_y|B) > E(v_y|C)$ and $\alpha_B \geq \alpha_C$, this represents an increase unless $E(v_y|D) > E(v_y|B)$, or

$$E(v_y|v_a \geq K\ ,\ v_y y + v_a \geq J) > E(v_y|v_a \leq K\ ,\ v_y y + v_a \geq J) \tag{A.1}$$

for some K,J, which is ruled out by the negative affiliation (or independence) of v_a and v_y.[39]

(b) Suppose next that, as shown in Figure 5b, the introduction of the reward attracts new types in areas E and F and does not induce any defection. Individuals with the minimal valuation for

[39] Indeed, if two random variables $(-Y)$ and X are positively affiliated, then $E_Y(-Y \mid -Y < x' - T, X = x)$ is increasing in both x and x', for all T. Hence, $E_Y(Y \mid Y + x > T, X = x) = E_Y(Y \mid Y + X > T, X = x)$ is decreasing in x. Consequently, averaging this expectation over x's larger than any given K must give a smaller number than averaging it over x's smaller than K

$$E_X(E_Y(Y \mid Y + X > T, X)|X > K) < E_X(E_Y(Y \mid Y + X > T, X) \mid X < K)],$$

which means that $E[(Y|Y + X > T, X > K) < E(Y|Y + X > T, X < K)$.

money $v_y^- = 0$ participate for $v_a \geq v_a^{**} \equiv c_a - R(y)$, and by assumption $v_a^{**} < v_a^* \equiv c_a - R(0)$. From part (a) of the proof, we know that new equilibrium reputation $R(y)$ of participants is less than the one –which we denote as R– that would obtain if only those in area E had joined in as a result of the reward y being offered. Let us now evaluate

$$\begin{aligned} R(0) - R &= \mu_a \left[E\left(v_a|v_a \geq v_a^*\right) - E\left(v_a|v_a < v_a^*\right)\right] - \mu_a \left[E\left(v_a|v_a \geq v_a^{**}\right) - E\left(v_a|v_a < v_a^{**}\right)\right] \\ &\quad -\mu_y \left[E\left(v_y|v_a \geq v_a^*\right) - E\left(v_y|v_a < v_a^*\right)\right] - \mu_y \left[E\left(v_y|v_a \geq v_a^{**}\right) - E\left(v_y|v_a < v_a^{**}\right)\right]. \end{aligned}$$

The (weakly) negative affiliation between v_a and v_y implies that $E\left(v_y|v_a \geq X\right)$ is nonincreasing in X, whereas $E\left(v_y|v_a < X\right)$ is nondecreasing; therefore,

$$E\left(v_a|v_a \geq v_a^*\right) - E\left(v_a|v_a \geq v_a^{**}\right) \leq 0 \leq E\left(v_y|v_a < v_a^*\right) - E\left(v_y|v_a < v_a^{**}\right),$$

hence

$$\begin{aligned} R(0) - R &\geq \mu_a \left[E\left(v_a|v_a \geq v_a^*\right) - E\left(v_a|v_a < v_a^*\right)\right] - \mu_a \left[E\left(v_a|v_a \geq v_a^{**}\right) - E\left(v_a|v_a < v_a^{**}\right)\right] \\ &= \mu_a \left[\mathcal{M}^+(v_a^*) - \mathcal{M}^-(v_a^*)\right] - \mu_a \left[\mathcal{M}^+(v_a^{**}) - \mathcal{M}^-(v_a^{**})\right]. \end{aligned}$$

This, in turn, implies that

$$\begin{aligned} \Psi(v_a^{**}) - \Psi(v_a^*) &\equiv v_a^{**} - v_a^* + \mu_a \left[\mathcal{M}^+(v_a^{**}) - \mathcal{M}^-(v_a^{**})\right] - \mu_a \left[\mathcal{M}^+(v_a^*) - \mathcal{M}^-(v_a^*)\right] \\ &\geq v_a^{**} - v_a^* + R(0) - R \geq v_a^{**} - v_a^* + R(0) - R(y) \equiv 0, \end{aligned}$$

which contradicts the fact that Ψ is increasing. Therefore, Figure 5b cannot represent an equilibrium, and thus $R(y) < R(0)$ necessarily. ■

Proof of Proposition 2.

With $\Omega(a, y) = 0$, the regression coefficients in (16)-(17) are independent of a and are therefore functions of y only, so that the differential system (14)-(15) *becomes linear.* 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.

As only marginal values matter for decisions (and therefore also inference), let us differentiate (14)-(15) with respect to a, yielding:

$$\frac{dE\left(v_a|a\right)}{da} = \rho\left[C''(a) - \bar{R}'(a)\right] = \rho\left(C''(a) - \mu_a \frac{dE\left(v_a|a\right)}{da} + \mu_y \frac{dE\left(v_y|a\right)}{da}\right), \tag{A.2}$$

$$\frac{dE\left(v_y|a\right)}{da} = \chi\left[C''(a) - \bar{R}'(a)\right] = \chi\left(C''(a) - \mu_a \frac{dE\left(v_a|a\right)}{da} + \mu_y \frac{dE\left(v_y|a\right)}{da}\right). \tag{A.3}$$

This implies that $dE\left(v_a|a\right)/da = \rho h(a)$ and $dE\left(v_y|a\right)/da = \chi h(a)$, where $h(a) \equiv C''(a) - \bar{R}'(a)$

is a solution to the linear differential equation

$$h(a) + \mu h'(a) = C''(a), \tag{A.4}$$

where $\mu \equiv \bar{\mu}_a \rho - \bar{\mu}_y \chi$ may be positive or negative. Define the function:

$$\hat{h}(a) \equiv \int_0^{+\infty} C''(a - \mu z)\ e^{-z} dz, \tag{A.5}$$

noting that the integral is convergent due to the fact that, whatever the sign of μ, $C''(a - \mu z)$ has a polynomial approximation as $z \to +\infty$. This function has the following properties:

$$C'(a) - \mu \hat{h}(a) = \int_0^{+\infty} C'(a - \mu z)\ e^{-z} dz, \tag{A.6}$$

$$C''(a) - \mu \hat{h}'(a) = \int_0^{+\infty} C''(a - \mu z)\ e^{-z} dz = \hat{h}(a). \tag{A.7}$$

The first equation is obtained from integration by parts of (A.5). The second, obtained by differentiating the first, shows that $\hat{h}$ is a solution to the differential equation (A.4). The generic solution is therefore $h(a) = \hat{h}(a) + \kappa e^{-a/\mu}$, where κ is a constant of integration. These additional solutions, however, are not linked to the problem's economic "fundamentals" (e.g., the cost function $C(\cdot)$). Moreover, unless $\kappa = 0$, $|h(a)|$ will tend to $+\infty$ as a tends to $+\infty$ (if $\mu < 0$) or $-\infty$ (if $\mu > 0$), meaning that a marginal increase in contribution will have an arbitrarily large effect on reputation and utility. Excluding such counterintuitive, "bubble-like" solutions leaves $\kappa = 0$ and $h = \hat{h}$ as the only economically sensible solution.

Finally, using (18), the first-order condition (11) takes the form:

$$v_a + y \cdot v_y = C'(a) - \mu h(a) = \int_0^{+\infty} C'(a - \mu z)\ e^{-z} dz.$$

Replacing $\mu = \mu(y)$ everywhere yields (18), (19), and the other results. ■

Proof of Corollaries 1 and 2

From (16) and (17) we have

$$\rho'(y) = -\frac{2y\sigma_a^2\sigma_y^2 + \sigma_{ay}\left(\sigma_a^2 + y^2\sigma_y^2\right)}{\left(\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2\right)^2}, \tag{A.8}$$

$$\chi'(y) = \frac{\sigma_y^2\left(\sigma_a^2 - y^2\sigma_y^2\right) - 2\sigma_{ay}\left(y\sigma_y^2 + \sigma_{ay}\right)}{\left(\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2\right)^2}. \tag{A.9}$$

Substituting into (22) immediately yields Corollary 1 in the case $y = 0$, and the condition given

in Corollary 2 when $\sigma_{ay} = 0$. This last inequality can be rewritten as

$$Q(y) = (\bar{v}_y/k)\left(1 + y^2\theta^2\right)^2 + \bar{\mu}_y^2\theta^2 y^2 < 2\bar{\mu}_a\theta^2 y + \bar{\mu}_y\theta^2 \equiv L(y). \tag{A.10}$$

The left hand side is a second order polynomial in y^2; it is necessarily convex and symmetric over all of $\mathbb{R}$, and takes 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:

i) if $L(0) \geq Q(0)$, then for any $\bar{\mu}_a > 0$, $L(y)$ intersects $Q(y)$ once on $\mathbb{R}_+^*$ and once on $\mathbb{R}_+^-$. Let $y_1' < 0 < y_1$ denote these two points.

ii) if $L(0) < Q(0)$ then there exists a unique $\mu_a^* > 0$ for which $L(y)$ has a (single) tangency point $y^* > 0$ with $Q(y)$. For all $\bar{\mu}_a < \mu_a^*$, $Q(y) > L(y)$ 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 on $\mathbb{R}_+^*$, at points that we denote $0 < y_1 < y_2$.

The properties, together with the linearity of L in $\bar{\mu}_a y$, yield the desired results. ■

Proof of Proposition 3

It is straightforward to verify that the solution described in the proposition is a solution to the general problem with $C'(a) = ka$, and in fact the only possible one for which $\partial E\left(v_a|a,y\right)/\partial a$ and $\partial E\left(v_y|a,y\right)/\partial a$, or equivalently $\rho(a,y)$ and $\chi(a,y)$, are independent of a. Indeed, one can replicate the proof of Proposition 2 but now with $C'' \equiv k$, which implies that $\hat{h}(a) = k$. The only difference is the presence of the term $\Omega(y)^2 = k^2\ Var[R(y)]$ in the denominator of ρ and χ (see (16)-(17)), leading to the fixed-point equation defining $\Omega(y)$:

$$\Omega^2 = k^2 Var\left[\mu_a\left(\frac{\sigma_a^2 + y\sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2 + \Omega^2}\right) - \mu_y\left(\frac{y\sigma_y^2 + \sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2 + \Omega^2}\right)\right] \equiv Z(\Omega^2) \tag{A.11}$$

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. This implies that 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 small enough. Finally, multiplying all the $\left(\mu_a, \mu_y\right)$'s by a common "publicity factor" x is the same, from the point of view of inference, as multiplying k^2 by x, and therefore it has (in any stable equilibrium) the above-mentioned dampening effects on $\rho(y)$ and $\chi(y)$. ■

Proof of Proposition 10

In what follows, we shall denote $v_a^*(y)$ simply as v_a^*. Social welfare is equal to $W = \pi + \bar{U}$, where $\pi(y) = [1 - G(v_a^*)]\,[B - y]$ is the sponsor's payoff and $\bar{U}$ denotes individuals' average

(ex-ante) utility:

$$\begin{aligned}\bar{U}(y) &= \mu_a \mathcal{M}^- (v_a^*) G(v_a^*) + \int_{v_a^*}^{v_a^+} \left[v_a - c_a + y + \mu_a \mathcal{M}^+ (v_a) \right] dG(v_a) \\ &= \mu_a \overline{v}_a + \int_{v_a^*}^{v_a^+} [v_a - c_a + y] \, dG(v_a) . \end{aligned} \tag{A.12}$$

by the martingale property of reputations. At the optimal reward y^*, $d\pi/dy = 0$ so

$$1 - G(v_a^*) = (B - y) \, g(v_a^*) \left(-\frac{dv_a^*}{dy} \right), \tag{A.13}$$

and

$$\frac{dW}{dy} = \frac{d\pi}{dy} + \frac{d\bar{U}}{dy} = \frac{\partial \bar{U}}{\partial y} + \frac{\partial \bar{U}}{\partial v_a^*} \cdot \frac{dv_a^*}{dy}.$$

Using (A.12) together with $\Psi(v_a^*) \equiv c_a - y$ and $dv_a^*/dy = 1/\Psi'(v_a^*)$ to evaluate this expression establishes that $dW/dy \,|_{y=y^*} < 0$ if and only if condition (29) is satisfied.

In the uniform case ($v_a \sim U[0,1]$), condition (29) becomes $\mu_a/2 > 1 - v_a^*$; substituting in the appropriate values shows that a marginal decrease in the reward from $y^* = (B + c_a - 1 - \mu_a/2)/2$ increases welfare if and only if $B - c_a < (\mu_a/2) - 1$. This inequality is consistent with the conditions that $0 < v_a^* < 1$ respectively, which are

$$-1 - \frac{\mu_a}{2} < B - c_a < 1 - \frac{\mu_a}{2}. \blacksquare$$

Proof of Proposition 11, part (ii) For any menu designed by the sponsor, revealed preference implies that an individual with a higher v_y selects a (weakly) higher reward. Suppose that all types $v_y \geq w$ contribute. We shall keep w as fixed (and therefore also the utility from non-participation $U(w) = -\mu_y \mathcal{M}^-(w)$ and show that an optimal menu in this class must involve separation on $[w, v_y^+]$, where (v_y^-, v_y^+) denotes the (interior of) the support of v_y.

By the revelation principle, a menu can be characterized by allocation of income $Y(v_y)$, reputation $r(v_y)$ and resulting overall utility $U(v_y) = v_a - c_a + v_y Y(v_y) + r(v_y)$ for all types $v_y \geq w$. By the martingale property of reputations, the average reputation of participants, $E\left(r(v) | v \in [w, v_y^+)\right)$ is invariant across all such menus, and equal to the average participating type, type $E\left(v | v \in [w, v_y^+)\right)$. The following Lemma characterizes the (cumulative) allocation of reputation up to any $v_y \leq v_y^+$.

Lemma 2 *For any feasible menu such that the cutoff for participation is w,*

$$V(v_y) \equiv E\left(r(v) | v \in [w, v_y)\right) \geq E\left(v | v \in [w, v_y)\right), \tag{A.14}$$

with equality if and only if the menu is separating ($r(v) = v$) on $[w, v_y]$.

Proof: If $Y(v) > Y(v_y)$ for all $v > v_y$ then there is no pooling of types above v_y with types below it, so by the martingale property (A.14) holds with equality. Suppose therefore that $z \equiv \sup\{v \geq v_y | Y(v) = Y(v_y)\} > v_y$ and that $t \equiv \inf\{v \geq w | Y(v) = Y(v_y)\} < v_y$. Then Y is constant on $[t, w]$ implying that all types in this interval must also receive the same reputation $r(v_y) = E(v | v \in [t, z))$. Hence:

$$E(r(v)|v \in [t, v_y)) = E(v|v \in [t, z)) > E(v|v \in [t, v_y)). \tag{A.15}$$

Next, denoting by H the cumulative distribution of v_y, we can write

$$V(v_y) = \left(\frac{H(t) - H(w)}{H(v_y) - H(w)}\right) E(r(v)|v \in [w, t)) + \left(\frac{H(v_y) - H(t)}{H(v_y) - H(w}\right) E(r(v)|v \in [t, v_y)),$$

By definition of t there is no pooling between $[w, t]$ and $[t, v_y^+]$, so $E(r(v)|v \in [w, t)) = E(v|v \in [w, t))$. Together with (A.15), this establishes the claim. ■

We now consider the problem of the sponsor designing an optimal contract, which can be written as

$$\min_{\{Y(\cdot),\ r(\cdot),\ U(\cdot)\}} \left\{ \int_w^{v_y^+} Y(v_y)\ h(v_y)\ dv_y \right\}, \quad \text{subject to:}$$
$$U(v_y) = v_a - c_a + v_y Y(v_y) + r(v_y),$$
$$\dot{U}(v_y) = Y(v_y),$$
$$V(v_y) \geq \int_w^{v_y} vh(v)\ dv,$$
$$\dot{V}(v_y) = r(v_y)h(v_y)$$

The first and last constraints follow directly from the definitions of U and V; the third one expresses incentive-compatibility ($v_y\dot{Y}(v_y) + \dot{r}(v_y) = 0$), while the fourth corresponds to (A.14). Using the first constraint to eliminate $Y(v_y)$ and denoting as $\zeta(v_y), \theta(v_y)h(v_y)$ and $\xi(v_y)$ the multipliers associated to the subsequent constraints, the Hamiltonian is

$$\mathcal{H} = h\left[(r + v_a - c_a - U)/v_y + \theta V\right] - \zeta\left[(r + v_a - c_a - U)/v_y\right] + \xi rh, \tag{A.16}$$

and the optimality conditions are

$$\frac{\partial \mathcal{H}}{\partial r} = \frac{h}{v_y} - \frac{\zeta}{v_y} + \xi h = 0, \tag{A.17}$$

$$\dot{\zeta} + \frac{\partial \mathcal{H}}{\partial U} = \dot{\zeta} - \frac{h}{v_y} + \frac{\zeta}{v_y} = 0, \tag{A.18}$$

$$\dot{\xi} + \frac{\partial \mathcal{H}}{\partial V} = \dot{\xi} + h\theta = 0, \tag{A.19}$$

Combining (A.17) with the transversality condition $\zeta(v_y^+) = 0$ yields

$$\zeta(v_y)v_y = -\left[1 - H(v_y)\right]. \tag{A.20}$$

Suppose now that the constraint $V(v_y) \geq \int_w^{v_y} vh(v)\ dv$ is not binding on some interval $[t, z]$. This implies that $\theta \equiv 0$ on this interval, so by (A.19) ξ must be constant. But (A.19) and (A.20) then imply that

$$\xi(v_y) = \frac{\zeta(v_y)}{h(v_y)v_y} - \frac{1}{v_y} = -\left(\frac{1 - H(v_y)}{h(v_y)v_y^2} + \frac{1}{v_y}\right).$$

By the monotone hazard rate the right-hand side is strictly increasing in v_y, contradicting the fact that ξ must be constant. Therefore, it must be that $V(v_y) = \int_w^{v_y} vh(v)\ dv$ for all $v_y \geq w$, which by Lemma 2 means that the menu is separating on $[w, v_y)$. ■

Proof of Proposition 12

(i) A sponsor with the ability to credibly commit to (or verifiably disclose) the terms of the contract he offers can always replicate the equilibrium choice of one without commitment, and therefore he can do at least as well. Since we show below that he in fact chooses a different fee (as long as $\mathcal{R}' \neq 0$), he must in fact do strictly better.

(ii) Consider the case where there is disclosure. If the sponsor still chooses $y = y^C$, the reservation value and the level of supply that result remain the same as in the confidentiality equilibrium: since $v_a^C - c_a + y^C + \mathcal{R}\left(v_a^C\right) \equiv 0$ 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:

$$\bar{a}_C'(y^C) = g\left(c_a - y^C - \mathcal{R}\left(v_a^C\right)\right), \tag{A.21}$$

whereas

$$\begin{aligned} \bar{a}_D'(y^C) &= g\left(c_a - y^C - \mathcal{R}\left(v_a^C\right)\right)\left[1 + \mathcal{R}'\left(v_a^*(y)\right) v_a^{*\prime}(y\right] \\ &= g\left(c_a - y^C - \mathcal{R}\left(v_a^C\right)\right)\left[1 + \mathcal{R}'\left(v_a^*(y)\right)\right]^{-1}, \end{aligned} \tag{A.22}$$

where the second equation follows from the definition of $v_a^*(y)$. Therefore, if $\mathcal{R}' < 0$ we have $\bar{a}_D'(y^C) > \bar{a}_C'(y^C)$, implying that the optimal price under disclosure is strictly above y^C, since

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

and π_D was assumed to be quasiconcave. Hence $y^D > y^C$, resulting in a higher supply $\bar{a}_D(y^D) > \bar{a}_D(y^C) = \bar{a}_C(y^C)$. The same reasoning works in reverse when $\mathcal{R}' > 0$.

(iii) Suppose that the sponsor can secretly offer a reward y different from the announced y^D. If $\mathcal{R}' < 0$, then $y^D > y^C$ as we have just seen, so ex-post he would like to secretly offer the lower payment y^C; the agent, however, can insist on receiving y^D. By contrast, when $\mathcal{R}' > 0$,

$y^D < y^C$ so it is optimal for the sponsor to secretly offer an increase $y^C - y^D$ to the agent, who will happily accept. Anticipating this collusive renegotiation, the audience properly expects that the actual fee will be y^C and not y^D. ■

Proof of Proposition 16

i) In a separating equilibrium, the individual's cutoff is assessed to be v_a^H in case of disclosure, and v_a^L in its absence. The image-conscious type therefore discloses after contributing if: and only if

$$\gamma_a^H \left[\mathcal{M}^+(v_a^H) - x\mathcal{M}^+(v_a^L) - (1-x)\, E\left(v_a \left|\phi, v_a^H, v_a^L\right.\right)\right] \geq d, \tag{A.23}$$

where $E\left(v_a \left|\phi, v_a^H, v_a^L\right.\right)$ is given by (37). The cutoffs v_a^H and v_a^L are respectively defined by

$$v_a^H + \gamma_a^H \left[\mathcal{M}^+(v_a^H) - E\left(v_a \left|\phi, v_a^H, v_a^L\right.\right)\right] = c_a + d - y \tag{A.24}$$

and

$$v_a^L + \gamma_a^L x \left[\mathcal{M}^+(v_a^L) - E\left(v_a \left|\phi, v_a^H, v_a^L\right.\right)\right] = c_a - y, \tag{A.25}$$

assuming as usual that the Ψ-type functions on the left-hand side are increasing in each case. These two inequalities, together with type v_a^H's willingness to disclose, imply that:

$$v_a^H < v_a^L. \tag{A.26}$$

ii) First, recall that under symmetric information about γ_a, the cutoff $\widehat{v}_a^H$ is given by

$$\widehat{v}_a^H + \gamma_a^H \left[\mathcal{M}^+\left(\widehat{v}_a^H\right) - \mathcal{M}^-\left(\widehat{v}_a^H\right)\right] = c_a + d - y, \tag{A.27}$$

while disclosure occurs only if

$$\gamma_a^H (1-x) \left[\mathcal{M}^+(\widehat{v}_a^H) - \mathcal{M}^-(\widehat{v}_a^H)\right] \geq d. \tag{A.28}$$

Let us now demonstrate part (ii) by way of an example: suppose that $x = 0$ (or more generally that x is not too large). Then (A.23) and (A.28) reduce to:

$$\mathcal{R}_D^H(v_a^H) \equiv \gamma_a^H \left(\frac{\theta G(v_a^H)\left[\mathcal{M}^+(v_a^H) - \mathcal{M}^-(v_a^H)\right] + (1-\theta)[\mathcal{M}^+(v_a^H) - \bar{v}_a]}{\theta G(v_a^H) + (1-\theta)}\right) \geq d$$

and

$$\widehat{\mathcal{R}}_D^H(\widehat{v}_a^H) \equiv \gamma_a^H \left[\mathcal{M}^+(\widehat{v}_a^H) - \mathcal{M}^-(\widehat{v}_a^H)\right] \geq d,$$

respectively. Note that $\widehat{\mathcal{R}}_D^H(v_a) > R_D^H(v_a)$ for all v_a. Making the now standard assumption $1 + \left(\widehat{\mathcal{R}}_D^H\right)' > 0$, and using the fact that

$$v_a^H - c_a + y + \mathcal{R}_D^H(v_a^H) - d = \widehat{v}_a^H - c_a + y + \widehat{\mathcal{R}}_D^H(\widehat{v}_a^H) - d = 0,$$

we obtain $\widehat{v}_a^H < v_a^H$, and therefore $\mathcal{R}_D^H(v_a^H) < \widehat{\mathcal{R}}_D^H(\widehat{v}_a^H)$. Therefore, for $R_D^H(v_a^H) < d < \widehat{\mathcal{R}}_D^H(\widehat{v}_a^H)$, disclosure by γ_a^H types will no longer occur under asymmetric information about γ_a. ■

Proof of Proposition 18

(i) 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 then sets y so as to make the low type indifferent:

$$y = c_a - v_a^L - \mu_a \left(\bar{v}_a - v_a^L\right),$$

where $\bar{v}_a \equiv \rho v_a^H + (1-\rho) v_a^L$ is the prior mean. The sponsor's payoff is then:

$$\pi_1 \equiv B - y = B - c_a + v_a^L + \mu_a \left(\bar{v}_a - v_a^L\right). \tag{A.29}$$

Suppose now that the high type is asked to sacrifice. Rewards are then y^L and y^H, respectively, where $y^L = c_a - v_a^L$ and (from incentive compatibility)

$$y^H = y^L + c_a^H - c_a - \mu_a \left(v_a^H - v_a^L\right).$$

The sponsor's payoff is then

$$\pi_2 = B - \rho y^H - (1-\rho) y^L = \pi_1 - \rho \left(c_a^H - c_a\right) < \pi_1, \tag{A.30}$$

hence the first result.

(ii) As mentioned earlier, under free entry all sponsors offer, and all contributors accept, a reward equal to B. Moreover, it is now an equilibrium for the high type to separate from the low type by choosing to sacrifice (opting for a sponsor who imposes such a requirement), if and only if

$$c_a^H - c_a \leq \mu_a \left(v_a^H - v_a^L\right). \tag{A.31}$$

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 $\mu_a v_a^L + v_a^L - c_a + B$, while the high type's payoff increases by at least $v_a^L - c_a + B$, 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, J. (1993) "An Experimental Test of the Public-Goods Crowding-Out Hypothesis," *American Economic Review,* 83(5), 1317-27.

Akerlof, G. and Dickens, W. (1982) "The Economic Consequences of Cognitive Dissonance," *American Economic Review,* 72(3): 307–319.

Akerlof, G., and R. Kranton (2000) "Economics and Identity," *Quarterly Journal of Economics*, 115: 716–753.

Araujo, A., D. Gottlieb, , and H. Moreira (2004) "A Model of Mixed Signals with Applications to Countersignaling and the GED," Getulio Vargas Foundation mimeo.

Batson, D. (1998) "Altruism and Prosocial Behavior," chapter 23 in D. Gilbert, S. Fiske, and G. Lindzey eds., *Handbook of Social Psychology*, vol. II,. McGraw Hill, 282–316.

Battaglini, M., R. Bénabou, and Tirole, J. (2002) "Self-Control in Peer Groups," CEPR Discussion Paper No. 3149, January. Forthcoming in the *Journal of Economic Theory.*

Bénabou, R., and J. Tirole (2004) "Willpower and Personal Rules," *Journal of Political Economy,* 112(4), 848-887.

—— (2002) "Self Confidence and Personal Motivation," *Quarterly Journal of Economics*, 117(3): 871–915.

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

Bem, D. J. (1972) "Self-perception Theory", in L. Berkowitz , ed., *Advances in Experimental Social Psychology*, Vol. 6, 1-62. New York: Academic Press.

Berman, E. (2000) "Sect, Subsidy and Sacrifice: An Economist's View of Ultra-Orthodox Jews," *Quarterly Journal of Economics,* August, 905-953.

Bernheim, D. (1994) "A Theory of Conformity," *Journal of Political Economy*, 102: 842–877.

Bodner, R. and D. Prelec (2003) "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.

Bohnet, I., Huck, S., and B.S. Frey (2001) "More Order with Less Law: On Contract Enforcement and Crowding," *American Political Science Review*, 95: 131–134.

Brekke, K. A., S. Kverndokk and K. Nyborg (2003) "An Economic Model of Moral Motivation," *Journal of Public Economics,* 87(9-10), 1967-83.

Buraschi, A. and Cornelli, F. (2002) "Donations," CEPR Discussion Paper No. 3488, August.

Caplin, A., and B. Nalebuff (1991) "Aggregation and Social Choice," *Econometrica*, 1-24.

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

Cho, I.K., and D. Kreps (1987) "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102: 179–221.

Damiano, E., and H. Li (2003) "Competing Match Making," mimeo, University of Toronto.

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

Deci, E. (1975) *Intrinsic Motivation*, New York: Plenum.

Deci, E. and Ryan, R. (1985) *Intrinsic Motivation and Self-Determination in Human Behavior* (New York: Plenum Press).

Denrell, J. (1998) "Essays on the Economic Effects of Vanity and Career Concerns," Stockholm Institute of International Economics Press.

Fehr, Klein, A. and K. Schmidt (2001) "Fairness, Incentives, and Contractual Incompleteness", University of Munich, mimeo.

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

Festinger, L. and Carlsmith, J. (1959) "Cognitive Consequences of Forced Compliance," *Journal of Abnormal and Social Psychology,* 58, 203–210.

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

Frey, B. (1997) *Not Just for the Money. An Economic Theory of Personal Motivation*, Edward Elgar, Cheltenham.

Frey, B., and L. Götte (1999) "Does Pay Motivate Volunteers?" Unpublished manuscript, Institute for Empirical Economic Research. University of Zurich.

Frey, B., and R. Jegen (2001) "Motivation Crowding Theory: A Survey of Empirical Evidence," *Journal of Economic Surveys*, 15(5): 589– 611.

Friebel, G., and S. Guriev (2002) "Should I Stay or Can I Go?: Worker Attachment in Russia," CEPR DP2368.

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

Gintis, H., Smith, E., and S. Bowles (2001) "Costly Signaling and Cooperation," *Journal of Theoretical Biology*, 213: 103–119.

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

Gneezy, U. and A. Rustichini (2000a) "A Fine is a Price," *Journal of Legal Studies*, 29(1), Part 1, 1-17.

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

— (2003) "The W effect of Rewards," University of Chicago Graduate School of Management, mimeo.

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

Jewitt, I. (2004) "Notes on the Shape of Distributions," Oxford University mimeo, July.

Johansson-Stenman, O., and Svedsäter, H. (2003) "Honestly, Why Are You Driving a BMW?" Göteborg University mimeo. Forthcoming in the *Journal of Economic Behavior and Organization.*

Kahneman, D. and Knetsch, J. (1992) "Valuing Public Goods: the Purchase of Moral Satisfaction," *Journal of Environmental Economics and Management,* 22, 57-70.

Köszegi (2000) "Ego Utility, Overconfidence, and Task Choice," U.C. Berkeley mimeo.

Lamont, M. (2000) *"The Dignity of Working Men."* New York: Russel Sage Foundation Press.

Lazear, E. (2000a) "Performance Pay and Productivity," *American Economic Review*, 90: 1346–1361.

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

Leibenstein, H. (1950) "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand," *Quarterly Journal of Economics*, 64: 183–207.

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

Lerner, M. (1980) *The Belief in a Just World: A Fundamental Delusion.* New York: Plenum.

Maskin, E., and J. Tirole (2004) "The Politician and the Judge : Accountability in Government," *American Economic Review*, 94 .

Murningham, K., Oesch J. and M. Pillutla (2001) "Player Types and Self-Impression Management in Dictatorship Games: Two Experiments," *Games and Economic Behavior* 37, 388-394.

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

Pillutla, M. and Murningham, K. (2003) "Fairness in Bargaining," *Social Justice Research,* September.

Potters, J. Sefton M. and L. Vesterlund. (2001) "Why Announce Leadership Contributions? An Experimental Study of the Signaling and Reciprocity Hypotheses," University of Pittsburgh mimeo.

Prendergast, C. (1999) "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37: 7–63.

Prendergast, C., and L. Stole (2001) "The Non-Monetary Nature of Gifts," *European Economic Review*, 45: 1793–1811.

Sadowski, P. (2004) "Overeagerness," Princeton University mimeo.

Seabright, P. (2002) "Continuous Preferences Can Cause Discontinuous Choices: an Application to the Impact of Incentives on Altruism," mimeo, IDEI.

Smith, A. (1759) *The Theory of Moral Sentiments.*

Titmuss, R. (1970) *The Gift Relationship*, London: Allen and Unwin.

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

Veblen, T. (1899/1934) *The Theory of the Leisure Class.* New York, Modern Library.