Is Altruism Bad for Cooperation?

Sung-Ha Hwang
Samuel Bowles

SFI WORKING PAPER: 2008-07-029

SANTA FE INSTITUTE

Is altruism bad for cooperation?

Sung-Ha Hwang* and Samuel Bowles[§]
10 July, 2008

Abstract

Some philosophers and social scientists have stressed the importance for good government of an altruistic citizenry that values the well being of one another. Others have emphasized the need for incentives that induce even the self interested to contribute to the public good. Implicitly most have assumed that these two approaches are complementary or at worst additive. But this need not be the case. Behavioral experiments find that if subjects feel spite towards free riders and enjoy inflicting harm on them, near efficient levels of contributions to a public good may be supported when group members have opportunities to punish low contributors. Cooperation may also be supported if individuals are sufficiently altruistic that they internalize the group benefits that their contributions produce. Using a utility function embodying both spite and altruism we show that unconditional altruism towards other members attenuates the punishment motive and thus may reduce the level of punishment inflicted on defectors, resulting in lower rather than higher levels of contributions. Increases in altruism may also reduce the level of benefits from the public project net of contribution costs and punishment costs. The negative effect of altruism on cooperation and material payoffs is greater the stronger is the reciprocity motive among the members.

JEL codes: D64 (altruism); H41 (public goods)

Keywords: public goods, altruism, spite, reciprocity, punishment, cooperation

---

Affiliations: * Departments of Mathematics and Economics, University of Massachusetts at Amherst; [§]corresponding author, Santa Fe Institute and Dipartimento di Economia Politica, University of Siena. We thank the Behavioral Science Program of the Santa Fe Institute, the U.S. National Science Foundation, the University of Siena and the European Science Foundation for support of this project and Rajiv Sethi, Elisabeth Wood and [acknowledgements to be continued] for comments on an earlier draft.

1. Introduction

Both altruism and reciprocity may motivate individuals to contribute to the provision of a public good. Altruism induces the individual to unconditionally value the payoff of other members, while reciprocity implies a valuation of the others' payoffs that is conditional on their contributions (or other indications of their type). Reciprocators may value the payoffs of low contributors negatively and be motivated to reduce the payoffs of defectors at a cost to themselves, when this option is available. The prospect of punishment for low contributions may induce individuals to contribute more than they otherwise would (Fehr and Gaechter (2000), Anderson and Putterman (2006)).

We explore the possibility that these two motives for contribution – a positive valuation of the payoffs of others and a desire to avoid the punishment induced by a negative valuation of one's payoffs by others – may work at cross purposes. Specifically we show that by attenuating the punishment motive, a general increase in the level of unconditional altruism may reduce rather than increase contributions.

Thus, while one often refers to individuals as being 'cooperative' or 'uncooperative', the motives supporting high levels of cooperation in a group are heterogeneous, and they need not work synergistically. For example, experimental evidence indicates that unconditional altruists contribute more in a public goods game but are significantly less likely to punish low contributors (Carpenter, Bowles, Gintis, and Hwang (2008)).

In the next section we use the ideas of Levine (1998) and Rabin (1993) to explore the joint effects of altruism toward fellow group members and reciprocity-based spite towards low contributors in a public goods game. In section 3 we study the Nash equilibrium levels of punishment and contribution under varying levels of unconditional altruism of the members of a group. We show that because altruism may diminish the motivation to punish low contributors, the relationship between the level of altruism and contributions is non-monotonic, and that under plausible assumptions there exist a range of levels of altruism over which increases in altruism reduce equilibrium levels of contribution. Moreover the range for which altruism is bad for cooperation is larger the more reciprocal are the group members. In the conclusion we suggest some implications for how social preferences may support cooperation despite the sometimes counterproductive effects of increased altruism and the costly nature of punishment.

2. Altruism, reciprocity and cooperation

Consider a community of individuals indexed by $i = 1,...,n$ ($n \geq 3$) who may contribute to a public project by supplying an amount of effort $e_i \in [0, 1]$. The total contributions, $\sum_k e_k$, result in a benefit of $q\sum_k e_k$ which is shared equally among individuals in the community, while each individual experiences the cost of contribution, $1/2\ (e_i^2)$. With the notation of $\phi \equiv q/n$, i's material payoff without the punishment is

(1) $$\pi_i = \phi\sum_k e_k - \frac{1}{2}e_i^2$$

We note that the marginal private benefit of contribution is ϕ and suppose that $1/n < \phi < 1$; $1/n < \phi$ ensures that full contribution, $e_i = 1$, is socially optimal whereas $\phi < 1$ means that in the absence of punishment selfish individuals under-contribute to the public project ($e_i = \phi < 1$).

After contributions have been observed, each individual i can impose a cost on $j \neq i$ with monetary equivalent s_{ij} at cost $c_{ij}(s_{ij}) \equiv c(s_{ij}) = 1/2\ (s_{ij})^2$ to himself. The cost s_{ij} results from public criticism, shunning, ostracism, physical violence, exclusion from desirable side-deals, or another form of harm. Hence $s_i = \sum_{k\neq i} s_{ki}$ is the punishment inflicted upon i by other community members and $c_i = \sum_{k\neq i} c(s_{ik})$ is i's cost of punishing others.

Individual j's standing as a cooperative member of community, b_j, depends on j's level of effort and the contribution that j makes to the group, which we assume is public knowledge. Specifically, we assume

(2) $$b_j = 2e_j - 1$$

So $b_j = -1$ if j contributes nothing, and $b_j = 1$ if j contributes fully. This means that $e_j = 1/2$ is the point at which i evaluates j's cooperative behavior as neither good nor bad. This point could be shifted to any value between 0 and 1, but the added generality is not illuminating.

To model cooperative behavior with social preferences, we say that individual i's utility depends on his own material payoff π_i, the payoff π_k to other individuals $k \neq i$, the cost of punishing others, and the punishment inflicted on i, according to

(3) $$u_i = \pi_i - c_i - s_i + \frac{1}{n-1}\sum_{k \neq i}(a_i + \lambda_i b_k)(\pi_k - s_{ik})$$

where the parameter a_i, $-1 < a_i < 1$, is i's level of unconditional altruism if $a_i > 0$ and unconditional spite if $a_i < 0$ and $0 \leq \lambda_i \leq 1$ is the strength of i's reciprocity motive, valuing j's payoffs more highly if j conforms to i's concept of good behavior, and conversely (The function is similar in spirit to Levine 1998, but i's evaluation of k's type is here based on k's actions, rather than on k's level of altruism). The valuation of others' payoffs is weighted by the inverse of the number of other members so that changes in group size do not alter the importance of an individual's own payoffs relative to the payoffs of others.

Note that an individual punishing a shirker values the punishment per se rather than the benefits likely to accrue to the punisher if the shirker responds positively to the punishment. Members have an intrinsic motivation to punish the shirker, not simply a desire that the shirker should be punished. This means that punishing is 'warm glow' rather than instrumental towards affecting j's behavior (Andreoni, 1995, Anderson and Putterman, 2006). To avoid semantic confusion, note that unconditional altruism and the reciprocity-based spite that motivates punishment of low contributors are both forms of altruism as defined by biologists (assuming that the group benefits associated with the increased contributions induced by punishment outweigh the costs of punishment). Individuals acting according to these motives increase average payoffs in the group but would enhance their own payoffs were they to (respectively) not contribute or forgo punishing low contributors. We use the term altruism for its unconditional variant.

3. Altruism versus cooperation?

We model a two-stage optimization process in which individual i selects an effort level taking account of the effect of this choice on the punishment inflicted on i by other team members. We suppose that individuals in the community are homogenous: $\lambda \equiv \lambda_i$ and $a \equiv a_i$ for all i. To find the punishment inflicted on i, we first determine j's decision concerning the punishment of i depending on i's contribution level:

(4) $$s^*_{ji}(e_i) = \arg\max_{s_{ji}} u_j(e_j, s_{j1}, ..., s_{jn}, s_j) \quad \text{for all } j \neq i$$

With $c(s_{ji}) = 1/2\ (s_{ji})^2$ member j's choice of s^*_{ji} in (4) gives the first order condition for an interior solution as follows.

(5) $$c'(s^*_{ji}) = s^*_{ji} = \frac{1}{n-1}\left[\lambda(1-2e_i) - a\right]$$

or the marginal cost of punishing is equal to the marginal benefit of reducing i's payoffs given j's assessment of i's type, net of the subjective costs of inflicting this punishment on i given j's level of unconditional altruism. When $\lambda = 0$ and $a < 0$, j punishes i, but independent of i's contribution level. If $\lambda = 0$ and $a \geq 0$, no punishment occurs. If $\lambda > 0$ and

(6) $$e_i \geq e_0 \equiv \frac{1}{2\lambda}(\lambda - a)$$

then member j does not punish. Thus j's punishment of i is

(7) $$s^*_{ji}(e_i) = \begin{cases} -\dfrac{2\lambda}{n-1}e_i + \dfrac{\lambda - a}{n-1} & \text{if } e_i < e_0 \\ 0 & \text{if } e_i \geq e_0 \end{cases}$$

Note two things from equations (6) and (7): the level of contribution that i must make to avoid punishment by j is declining in j's level of altruism and if punishment occurs, the marginal reduction in punishment associated with contributing more does not depend on the level of altruism.

From (7) we can find the total punishment inflicted on individual i, $s^*_i(e_i) = \sum_{j \neq i} s^*_{ji}(e_i)$ which is then non-increasing and differentiable when it is positive. Next individual i decides the level of effort by taking account of the effect of his effort choice on the level of punishment he will receive. Thus member i will choose

(8) $$e_i(e_{-i}, a) = \arg\max_{e_i} v(e_i) \equiv u_i(e_i, s_{i1}, ..., s_{in}, s^*_i(e_i))$$

Equation (8) defines member i's best effort response to other's effort levels, $e_i = e_i(e_{-i}, a)$. To find i's best response explicitly we proceed as follows. When there is no punishment of i, an interior solution of $e^{*N}_i(e_{-i}, a)$ for (8) satisfies the following first order condition (recall $b_j = 2e_j - 1$).

(9) $$e^{*N}_i(e_{-i}, a) = \phi + \frac{1}{n-1}\sum_{l \neq i}(a + \lambda b_l)\phi$$

where $e_{-i} = (e_1, .., e_{i-1}, e_{i+1}, ..., e_n)$

Thus when no punishment is inflicted, i's optimal choice of e_i equates the marginal cost of contribution (e_i itself) to the direct benefits to i of contributing to the project, ϕ, plus i's valuation on others' material payoffs. Similarly when i is subject to punishment (hence $e_i < e_0$), i chooses e_i to satisfy the following first order condition :

$$e_i^{*P}(e_{-i}, a) \equiv \phi + \frac{1}{n-1}\sum_{l \neq i}(a + \lambda b_l)\phi - s_i^{*\prime}(e_i) \tag{10}$$

which requires that i take account of the effect of increased contribution in reducing punishment, as well as the marginal costs and benefits of the project expressed in the no-punishment first order condition (9). Since $s_i^{*\prime}(e_i) = -2\lambda$, we see that $e_i^{*P}(e_{-i}, a) > e_i^{*N}(e_{-i}, a)$; punishment supports a higher contribution level. We note that e_i^{*N} and e_i^{*P} are increasing in a. The amount contributed by i will depend on whether punishment is present or not, and this will depend on the level of unconditional altruism of the members of the group. There exist critical values , $\overline{a}$ and $\underline{a}$, such that the best response for member i is following.

$$e_i = \begin{cases} e_i^{*P}(e_{-i}, a) & \text{if } a < \underline{a} \\ \dfrac{1}{2\lambda}(\lambda - a) & \text{if } \underline{a} < a < \overline{a} \\ e_i^{*N}(e_{-i}, a) & \text{if } \overline{a} < a \end{cases} \tag{11}$$

Figure 1 illustrates equation (11).

When altruism is lower than $\underline{a}$, i is subject to punishments by others so the effort level is determined by equation (10) and hence is increasing in a. If the altruism is greater than $\underline{a}$, there is no punishment because $e_i \geq e_0$ and the effort level is determined by equation (9) and as a result is increasing in a. In both of these cases the expected positive effect of altruism occurs because altruism enhances the members' valuation of the external benefits that their contribution allow. However, in the intermediate range of altruism, equation (6) is binding so an increase in altruism *decreases* the equilibrium effort level since altruism lowers the threshold level of effort required to avoid being punished.

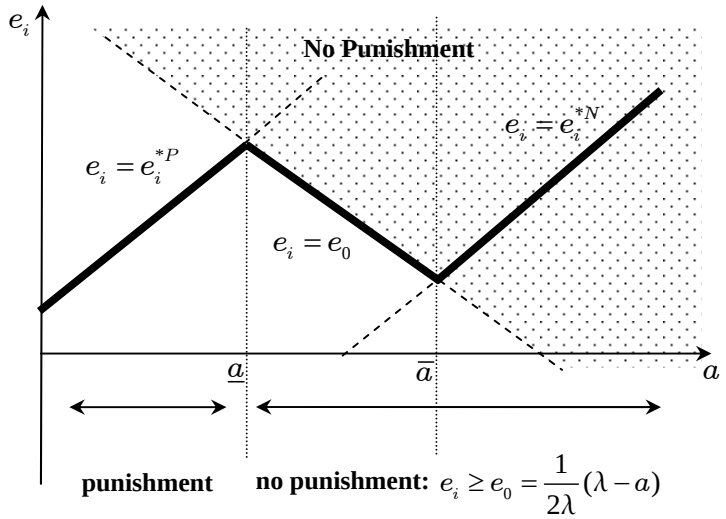

Figure 1. Equilibrium contributions as a function of group member's altruism.

Does the 'altruism bad for cooperation' range $(\underline{a}, \bar{a})$ occur for plausible parameter values? Recall that $e_i = \phi$ is the choice of selfish individuals in the absence of punishment and $e_i = 1/2$ is the critical point around which i's behavior is judged to be good or bad. Thus when the private marginal benefit of contribution, ϕ, is small, so that a selfish individual is motivated to be a bad type (i.e. when $\phi < 1/2$) and members have reciprocal motives ($\lambda > 0$), members would punish others and punishment would induce a higher effort level. So we infer that $\phi < 1/2$ and $\lambda > 0$ are necessary conditions for the existence of an interior equilibrium with positive punishment. And if the reciprocity motive is sufficiently strong among community members that the threshold level of effort to avoid punishment, e_0, reaches 1, an equilibrium with any positive punishment is characterized as full contributions by members. When we exclude cases in which punishment never occurs or in which when it does full contribution is always the result, i.e. $0 < \lambda < 1/4$ and $\phi < 1/2$, we obtain the following proposition.

Proposition 1. We suppose that $0 < \lambda < 1/4$ and $\phi < 1/2$. There exist $\underline{a}$ and $\bar{a}$ such that

$$\frac{de^*}{da} < 0 \text{ for } a \in (\underline{a}, \bar{a})$$

where e^* is a Nash equilibrium. Furthermore, we have

$$\frac{d}{d\lambda}(\bar{a} - \underline{a}) > 0$$

Proof. See appendix. ■

The second part of proposition 1 – that the range over which altruism has a negative effect is increasing in the degree of reciprocity – occurs because the stronger reciprocity motive is, the bigger is the gap between best responses with and without punishment. If $\lambda > 1/4 - \phi/2$ then $\underline{a} < 0$, so contributions are declining in a not only over the range $(0, \bar{a})$ but also over some range of reductions in spite. Note that while increases in altruism for values of a below $\underline{a}$ or above $\bar{a}$ increase the benefits of the public project net of contribution costs and punishment costs, the reverse is true in the 'altruism bad for cooperation' range. Here punishment costs are zero, but increases in altruism reduce contributions to the public good, thus lowering the net benefits.

The mechanism underlying the proposition – that altruism attenuates the motivation to punish low contributors – could be modeled in more general terms. We show in the appendix that if the marginal costs of punishing another group member depend on other group members' levels of altruism, perhaps due to the disapproval one may incur in punishing a fellow group member, then an increase in altruism would reduce the marginal benefits of contribution derived from the resulting reduced punishment. Were this effect large enough, the contributions given by (10) would be declining in a, thereby providing an additional range of values of a for which altruism is bad for cooperation.

4. Discussion

Some philosophers and social scientists have stressed the importance for good government of an altruistic citizenry that values the well being of one another. Others have emphasized the need for incentives that induce even the self interested to contribute to the public good. Implicitly most have assumed that these two approaches are complementary or at worst additive. It is now recognized that this assumption may fail where the presence of monetary or other explicit incentives reduces the salience of altruistic or other public spirited motives (Benabou and Tirole (2003); Benabou and Tirole (2006); Bowles (2008); Falk and Kosfeld (2006); Sliwka (2007); Bowles and Hwang (2008)). But as we have seen, the assumption need not hold even in the absence of such motivational crowding out.

Our results suggest that for a community wishing to sustain high levels of cooperation, seeking to enhance unconditional altruism may be counter-productive. But punishment may also be counter-productive. By definition acts of altruism increase the joint surplus of the community; but punishment is often (as in our model) resource-using. Unless or until levels of contribution sufficient to make punishment rare are achieved, the costs associated with punishment of low contributors may more than offset the gains to cooperation that the punishment allows (Herrmann, Thoni, and Gaechter (2008)). This is particularly true in a case we have not considered, namely when vendetta-like cycles of punishment and counter punishment are allowed. (Hopfensitz and Reuben (2006)).

Nonetheless, cooperation sustained by a combination of altruism and reciprocity-based punishment may be welfare enhancing. This is true in part because punishment is not only an incentive; it is also a signal. The incentive-based response to punishment is enhanced by the feelings of shame that punishment by peers triggers (Bowles and Gintis (2006).) In part for this reason disapproval by peers may induce members to contribute even when it is expressed in non-resource-using ways such as gossip, ridicule or the simple statement that the individual has violated a norm (Masclet, Noussair, Tucker, and Villeval (2003), Barr (2001)).

Appendix

1. Proof of Proposition I

We define the following critical values for a, $\underline{a}$ and $\overline{a}$ which give respectively the values of a for which $e_0 = e_i^{*P}(e_0, \underline{a})$ and $e_0 = e_i^{*N}(e_0, \overline{a})$.

(12) $$\underline{a} = \lambda(1-2\phi-4\lambda)$$

(13) $$\overline{a} = \lambda(1-2\phi)$$

Then $\lambda < 1/4 < (1-\phi)/2$ implies $-\lambda < \underline{a}$. Also since $\phi > 0$, we have $\overline{a} < \lambda$. Thus $-1 < -\lambda < \underline{a} < \overline{a} < \lambda < 1$. We find $\overline{e} = \overline{e}(a)$ such that $\overline{e} = e_i^{*P}(\overline{e}, a)$.

(14) $$\overline{e}(a) = \frac{\phi}{1-2\lambda\phi}a + \frac{\phi(1-\lambda)+2\lambda}{1-2\lambda\phi}$$

By our assumption we have $\overline{e} > 0$. Now if $a < \underline{a}$ then $\overline{e} < 1/(2\lambda)(\lambda - \underline{a}) < e_0 < 1$. Hence when $a < \underline{a}$, $e^* = \overline{e}$ constitutes a Nash equilibrium. Similarly we find $\underline{e} = \underline{e}(a)$ such that $\underline{e} = e_i^{*N}(\underline{e}, a)$.

(15) $$\underline{e}(a) = \frac{\phi}{1-2\lambda\phi}a + \frac{\phi(1-\lambda)}{1-2\lambda\phi}$$

Then for $a > \overline{a} > 0$, $\underline{e} > 1/(2\lambda)(\lambda - \overline{a}) > e_0$. Thus $e^* = \min\{\underline{e}, 1\}$ is a Nash equilibrium. Finally if $\underline{a} < a < \overline{a}$, then $\underline{e} < e_0 < \overline{e}$ thus $e^* = 1/(2\lambda)(\lambda - a)$ becomes a Nash equilibrium. From this the proposition 1 follows. We summarize this result.

(16) $$e^* = \begin{cases} \overline{e}(a) & \text{if } a < \underline{a} \\ \dfrac{1}{2\lambda}(\lambda - a) & \text{if } \underline{a} < a < \overline{a} \\ \min\{\underline{e}(a), 1\} & \text{if } \overline{a} < a \end{cases}$$

The second part of proposition 1 follows from

(17) $$\overline{a} - \underline{a} = 4\lambda^2$$

■

2. Altruism may increase the cost of punishing others

In the model we consider only the effects of altruism on the benefits of punishing others. But as we observed (p.7) the costs may also be effected. We consider the following specification of the cost function.

(18) $$c(s_{ji}) = \frac{1}{2}(a_j)^{\kappa}(s_{ji})^2$$

As before, we take $a \equiv a_j$ for all j, and κ is a positive constant representing the fact that in a more altruistic population the cost of punishing others may be greater either due to one's own distaste for harming others or because of the disapproval of others. Thus the marginal cost of punishing is

(19) $$c'(s_{ji}) = a^{\kappa} s_{ji}$$

and the effect of greater contribution on the amount of punishment received is

(20) $$s_i^{*\prime}(e_i) = -\frac{2\lambda}{a^{\kappa}}$$

From (20) we see that over the region of altruism such that punishment is positive, greater altruism in the population means that contributing more is associated with a lesser (in absolute value) reduction in the extent of punishment an individual may expect. If this effect is sufficiently large, (14) need not increase in a. When positive punishment occurs at equilibrium we can find the effect of altruism on contribution as follows.

(21) $$\frac{de^*}{da} = \frac{\phi - 2a^{-(\kappa+1)}\kappa\lambda}{(1-2\lambda\phi)}$$

Since $1-2\lambda\phi > 0$ by our assumptions we see that for sufficiently large κ and positive λ we may have $de^*/da < 0$. This negative effect of altruism on contribution is greater, the more reciprocal are the members of the population.

References

Andreoni, James. 1995. “Warm- Glow versus Cold-Prickle: The effects of Positive Negative Framing on Cooperation in Experiments.” *Quarterly Journal of Economics*, CX:1, pp.1-21.

Anderson, Christopher and Louis Putterman. 2006. "Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism." *Games and Economic Behavior*, 54, pp. 1-24.

Barr, Abigail. 2001. "Social dilemmas, shame-based sanctions, and shamelessness: experimental results from rural Zimbabwe." Centre for the Study of African Economies Working Paper WPS/2001.11: Oxford University.

Benabou, Roland and Jean Tirole. 2003. "Intrinsic and extrinsic motivation." *Review of Economic Studies*, 70, pp. 489-520.

Benabou, Roland and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96:5, pp. 1652-78.

Bowles, Samuel. 2008. "Policies designed for self interested citizens may undermine "the moral sentiments:" evidence from experiments." *Science*, 320:5883 (June 20).

Bowles, Samuel and Herbert Gintis. 2006. "Social Emotions," in *The Economy as a Complex Evolving System III: Essays in Honor of Kenneth Arrow*. Steven Durlauf and Lawrence Blume eds. Oxford: Oxford University Press.

Bowles, Samuel and Sung-Ha Hwang. 2008. "Social Preferences and Public Economics: Mechanism design when preferences depend on incentives." *Journal of Public Economics*, Vol 92:8-9, pp. 1811-20.

Carpenter, Jeffrey, Samuel Bowles, Herbert Gintis, and Sung-Ha Hwang. 2008. "Strong Reciprocity and Team Production: Theory and Evidence." SFI Working Paper.

Falk, Armin and Michael Kosfeld. 2006. "The Hidden Costs of Control." American Economic Review, 96:5, pp. 1611-30.

Fehr, Ernst and Simon Gaechter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, 90:4, pp. 980-94.

Herrmann, Benedikt, Christian Thoni, and Simon Gaechter. 2008. "Antisocial Punishment Across Societies." *Science*, 319: 7 March 2008, pp. 1362-67.

Hopfensitz, Astrid and Ernesto Reuben. 2006. "The importance of emotions for the effectiveness of social punishment." *Tinbergen Institute Working Paper 05-0571* (http://www.tinbergen.nl/discussionpapers/05075.pdf).

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1:3, pp. 593-622.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. 2003. "Monetary and Non-monetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review*, 93:1, pp. 366-80.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83:5, pp. 1281-302.

Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review*, 97:3, pp. 999-1012.