DEPFID

DEPARTMENT OF ECONOMIC POLICY, FINANCE AND DEVELOPMENT

UNIVERSITY OF SIENA

DEPFID WORKING PAPERS

Marcello Basili and Maurizio Franzini

Cooperation, reciprocity and self-esteem: A theoretical approach

7 / 2007

DIPARTIMENTO DI POLITICA ECONOMICA, FINANZA E SVILUPPO

UNIVERSITÀ DI SIENA

DEPFID Working Papers - 7/ November 2007

Marcello Basili* and Maurizio Franzini**

Cooperation, reciprocity and self-esteem: A theoretical approach

Abstract

Cooperation occurs even where it is not predicted by economic theory, owing to what is widely recognized as too narrow a conception of self-interest. In particular, relying on plenty of experimental evidence, it has been maintained that agents adopt such a strong reciprocity rules in their behavior as make it worthwhile to punish those who defect or do not act fairly, costly though as this may be. We propose to lay the analytical foundation of such behavior – and more generally to cooperation-proneness – by considering self-esteem. Agents may include self-esteem in their utility (or goal) function and actually produce or destroy self-esteem through their behavior. This amounts to introducing a moral system in individual behavior in such a way as to make it amenable to rational maximization. We also show how the impact of self-esteem on the best contract in Principal-Agents situations and how such impact differs in Moral Hazard and Adverse Selection situations.

KEYWORDS: self-esteem, reciprocity, motivation, incentive, agency

JEL CLASSIFICATION: J41, D64

ADDRESS FOR CORRESPONDENCE: basili@unisi.it

ACKNOWLEDGMENTS: The authors especially wish to thank S. Bowles, M. Chiarolla, H. Gintis and M. Pugno for useful comments

* DEPFID, University of Siena
** University of Rome « La Sapienza »

1. Introduction

Cooperation among genetically unrelated agents is widely observed in behavioral experiments and in everyday life, even when repeated interaction is absent. In most cases economic theory does not contemplate it. Basically, cooperation among strangers is ruled out by the usual assumptions of self-interested behavior. Only repeated interaction may reconcile traditional self-interest with cooperation. We lack an explanation of how cooperation can develop among strangers, in a setting potentially open to free riding and opportunism. Recently, several experiments have expanded our knowledge of important features of cooperative behavior under different circumstances. On the basis of that knowledge, an interesting hypothesis has been proposed: most agents are *strong reciprocators*, i.e. they are ready to punish those who behave opportunistically, even when this is costly to them (Bowles and Gintis 2004, Gintis *et al*. 2003, Gintis 2004). Compared to other possible explanations of cooperation, *strong reciprocity* seems to enjoy the positive feature, at least from an economist's point of view, of demanding a rather weak relaxation of the assumption that agents are self-interested.

The analytical foundations of strong reciprocity are, however, still unclear. In particular, it has not been demonstrated whether such behavior can be derived from a rational process of maximization. The main goal of this paper is to offer a possible explanation of strong reciprocity or, more generally, cooperative behavior as the end result of rational decision-making based on utility maximization.

In our interpretation, a rational foundation for a more cooperative-prone behavior can be provided by the twin assumptions that agents include self-esteem in their utility function and the amount of self-esteem depends on how they behave in social situations. The latter reflects an idea of moral system that is different and not based on reputation effects only. In considering moral values as an important component of individual decision-making, we follow Sen's criticism of the traditional conception of rationality. However, we try to take a step forward by explicitly considering moral values within a maximization process.

Such a model regards cooperation and reciprocity as a possible, not necessary, outcome. This is precisely what experiments and experience suggest we need: not a theory that invariably predicts cooperation but a more general framework that allows for cooperation as a possible outcome.

Moreover the explanation we offer fits well with the observed attempts to induce cooperation through a sort of *gift giving* – as in the famous essay by Akerlof (1982) – and also with the apparent existence of limits to cooperative behavior.

The paper is organized as follows. In Section 2 we introduce and evaluate the strong reciprocity hypothesis. In Section 3 we present our basic model of interaction between utility maximization and moral values, based upon the notion of self-esteem; we also illustrate how such utility function can lead either to cooperative or to more traditional behavior. In Section 4 we analyze more precisely in a Principal-Agent setting the notion of reciprocity and how it relates to self-esteem. In Section 5 the relationship between self-esteem and fairness is investigated, while in the following two sections self-esteem is considered in a Moral Hazard and Adverse Selection setting, to see how it influences the best contract. Concluding remarks end the paper.

2. Strong reciprocity: experimental evidence and theoretical foundations

Many empirical studies document that individuals cooperate in situations in which, according to economic theory, cooperation would not be a rational behavior (Fehr *et al*. 1997, Fehr and Gachter 2000). In particular, there is a large body of evidence about the existence of cooperation in situations involving public goods, common pool resource, ultimatum games and principal-agent interactions (Yamagishi 1986, Ostrom, Walker and Gardner 1992, Fehr and Gächter 2002).

In their attempt to better understand such behavior, and to give it rational foundations, Bowles and Gintis, among others, take a clear stand in favor of strong reciprocity. In their own words: "cooperation is maintained because many humans have a predisposition to punish those who violate group-beneficial norms, even when this reduces their fitness relative to other group members" (Bowles and Gintis 2004). The resulting human behavior is called *strong reciprocity* and is defined as a sort of altruistic behavior that, among others, may confer "group benefits by promoting cooperation, while imposing upon the reciprocator the cost of punishing shirkers" (Bowles and Gintis 2004).

The distinguishing behavior of strong reciprocators is that they punish a defector, when they detect one, even though such behavior is costly to them. Strong reciprocators are considered

altruistic people, in so far as they bear privately the cost of action that is beneficial to the community.

Therefore, individuals – at least some of them and at least in some situations – seem to have a preference for "punishing others". How can this be reconciled with rationality?

Gintis (2004) convincingly argues that it is not possible to offer a theoretical explanation of observed cooperation that fulfils some reasonable conditions[1], while retaining the assumption that agents are strictly self-interested. Indeed, the latter is to be relaxed.

Since we are faced with acts of volition carrying implied costs, reconciling strong reciprocity with rationality is not a trivial task. How and why is "punishing others" evaluated by the agents? Are there limits to the costs one is willing to bear in order to "punish others"? And if so, which are they? We lack a general model for the rational foundations of strong reciprocity or the cooperation-prone behavior that makes it possible to answer specifically these and other questions. More generally, we do not know how to modify traditional utility functions in order for such behavior to become a possible, but not necessary, result.

To this end, it might be a good idea to start from Sen's more recent criticism of the traditional "rational model" of choice, which is structured in three steps (Sen 2002, p. 34). The first is related to a notion of self-centered welfare, whereby "a person's welfare depends only on her own consumptions and other features of the richness of her life". The second criticism concerns what Sen calls self-welfare goal, i.e. the assumption that welfare maximization is the individual's only goal. The last criticism points to self-goal choices, whereby a person's choices are exclusively geared to the pursuit of her own goals.

Sen clearly aims at enriching the traditional model by weakening, in particular, the assumption that people pursue a too-narrowly-defined welfare. But, of the three criticisms he levels against conventional wisdom, the less convincing is precisely the last one, essentially because we are practically left without an operating theory of choice. If people, as Sen argues, are maximizers but care also about things different from their own welfare, how do they solve their maximization problem?

Sen does not say much on this. The solution we propose is largely in line with Sen's approach but departs from it in the assumption that people do maximize their utility function as

[1] These conditions, as Gintis calls them are: Incentive Compatibility, Dynamic Stability, Empirical Relevance, Plausible Informational Requirements and Plausible Discount Factors.

enriched with an endogenously determined "moral" variable. More specifically, individuals are endowed with a "moral system" which transforms their actions into self-esteem. The latter, as determined by such a system, enters their utility function and helps to define their choice within a utility maximizing process. Therefore, self-esteem brings utility but its "amount" is determined also by a "moral system" that lies outside the preference system sustaining the utility function.

In our definition, a moral individual has a high propensity to destroy self-esteem when her actions are not consistent with her moral values. This will be reflected in her final utility, given that self-esteem is positively related to utility. Therefore, her actions, in so far as they destroy her own self-esteem through the "moral value mechanism", are not determined by a too-restricted notion of self-welfare. In this respect, we share Sen's approach. However, the inclusion of self-esteem in the utility function (which could very well be defined as a goal-function) allows us to treat the choice problem as a typical maximization problem and give formal solution to it.

3. Self-esteem, moral values and utility maximization

To explain reciprocity independently of repetition of the game, we assume that individuals may "produce" through their behavior *self-esteem*, which in turn is amenable to cooperation and reciprocity. Self-esteem is created according to different mechanisms in dissimilar situations. We refer to a Principal-Agent framework which is broad enough to encompass many interesting cases. In such a framework, the Agent's self-esteem depends on the effort made (or, more generally, on limiting opportunism) in relation to the compensation obtained by the Principal.

It is common knowledge that the *homo sapiens* species is highly gregarious and there is a huge literature about affiliation among people without family relations. Psychologists, in particular, identify at least four different motivations that prompt human beings to affiliate: "to receive social attention, to obtain emotional support, because they find other people stimulating, and for social comparison" (Leary *et al.* 2003). Motivations are internal and external. Internal or intrinsic motivation is supported by innate psychological needs and there is proof of the strong links between intrinsic motivation and competence as well as satisfaction of the need for both autonomy and relatedness. It is worth bearing in mind "that people will be intrinsically motivated only for activities that hold intrinsic interest for them, activities that have the appeal of novelty,

challenge, or aesthetic value" (Ryan and Deci 2000, 71). On the contrary, external or extrinsic motivation refers to *performance of an activity in order to attain some separable outcome* and "the extrinsically motivated behaviors that are least autonomous are referred to as externally regulated, such behaviors are performed to satisfy an external demand or reward contingency" (Ryan and Deci 2000, p. 71). As a result of the interaction between intrinsic and extrinsic motivation "people can be motivated because they value an activity or because there is strong external coercion. They can be urged into action by an abiding interest or by a bribe. They can behave from a sense of personal commitment to excel or from fear of being surveilled. These contrasts between cases of having internal motivation versus being externally pressured are surely familiar to everyone" (Ryan and Deci 2000, p. 72).

Referring to this psychological literature[2], self-esteem takes into account both extrinsic and intrinsic motivations. It depends positively on the effort and negatively on the compensation, but only for the part of the effort that can be considered a gift, i.e. the part in excess of the price paid for the effort. As the gift gets larger, the agent will suffer a loss of self-esteem, if she refrains from making a greater effort. Thanks to self-esteem, the behavior of the agent can be seen, therefore, as the product of two antagonist forces: *altruism* (the utility of reciprocating a gift) and *self-interest* (the disutility of greater efforts).

In a Principal-Agent framework it is appropriate to assume that self-esteem demands greater efforts as compensation increases: the Agent knows that her Principal seeks greater efforts and is willing to offer a "gift" to that effect. In other settings the appropriate assumptions for self-esteem may be different. For example, in some cases self-esteem may increase by reducing the effort when compensation increases, because the recipient wants to show that her effort was a function of intrinsic motivation, not money. This conception of self-esteem may reinforce – or even supplant - the reputation effects that are considered as the root cause of money crowding out intrinsic motivation for cooperative behavior. Reputation in such models depends on others believing that we behave out of intrinsic motivation and not for money (Benabou and Tirole 2006a, 2006b). However in a Principal-Agent framework – and maybe not only in it – it seems more appropriate to assume that Agents take the exchange ethics as a reference point in establishing what is fair. If I get more money from my principal I owe her a greater effort. If I fail

[2] See for example: Ryan and Connell 1989, Ryan *et al.* 1993, Munir and Jackson 1997, Leary and Baumeister 2000.

to provide this effort I look at myself as a person deserving less esteem. The reputation-for-intrinsic-motivation framework may apply in different situations and is not so general. [3]

In order to understand how self-esteem works in a general setting, let us assume that our Agent gets money (*m*) in exchange for her effort (*e*). Were it not for self-esteem, her indirect utility function would be of the following type:

$$U = U(m, e) \quad \text{and usual assumptions are:}$$
$$U_m > 0; U_{mm} < 0; U_e < 0; U_{ee} < 0; U_{e,m} < 0$$

We can assume that *e* cannot be perfectly monitored, so that the agent is free to choose *e*, given *m*. There is no reason for *e* to increase when *m* changes and, in particular, for *e* to be lower than the minimum possible value it can take. We now let self-esteem (*E*) into the picture.

$$E = E(m, e) \text{ with } E_m < 0;\ E_e > 0$$

In other words: an Agent will increase (decrease) her self-esteem when, given *m*, she provides a higher (lower) effort or when, given *e*, she gets a lower (higher) compensation. The Agent, through her behavior, produces or destroys self-esteem according to the above function. On the other hand, she enjoys utility from self-esteem, which therefore enters her utility function. The E-function represents her "moral system", the utility function her "goal function", borrowing from Sen's terminology. Therefore the Agent faces the following constrained maximization problem[4]:

$$Max\ U = U(m, E, e)$$
$$\text{Subject to } E = E(m, e) \qquad \textbf{(1)}$$

Consistent with what we said before, we assume that *m* is given for the agent and that her control variable is only *e*. Therefore, after substituting the constraint in the goal function, we get the First Order Condition:

[3] See Benabou and Tirole 2006b, Ryan *et al.* 1994 and 1997, Baumeister and Leary 1995, Kim *et al.* 1998.

[4] For a utility function that incorporates identity, based on social categories, as a motivation for behavior, see for example Akerlof and Kranton 2000.

$$U_e + U_E E_e = 0 \qquad \textbf{(2)}$$

the optimum effort, *e**, is the value of *e* which solves

$$U_e = -U_E E_e \qquad \textbf{(3)}$$

The meaning of this condition is clear: a maximizing individual will take her own moral values into account when making a choice. The chosen *e* must be such that it balances the disutility of any additional effort with the utility of additional *E* induced by *e* itself. It can also be formulated as a condition of equality between the Marginal Rate of Substitution between *e* and E (how the individual is ready to trade off lower efforts for higher self-esteem.), on the one hand, and the marginal productivity of *e* on *E*, on the other.

$$\frac{U_e}{U_E} = -E_e \qquad \textbf{(4)}$$

In order for the level of effort satisfying this equation to be a maximum the second order conditions are to be satisfied. This imposes some restrictions on the admissible set of utility functions. Assume that $U(e, E(e,m), m)$ is of class C^2 so that mixed derivatives coincide, in particular $U_{eE} = U_{Ee}$; to guarantee that *e** is a point of maximum it is required that:

$$U_{ee} + U_{eE}E_e + U_{Ee}E_e + U_{EE}E_e^2 + U_E E_{ee} < 0$$

that is,

$$U_{ee} + 2U_{eE}E_e + U_{EE}E_e^2 + U_E E_{ee} < 0 \qquad \textbf{(5)}$$

Assume U is linear in E, i.e. $U(e, E(e,m), m) = \Theta(e,m)E(e,m) + \Gamma(e,m)$, then $U_{EE} = 0$. Also, recall that $U_{ee} < 0$, $U_e = \Theta_e E + \Gamma_e < 0$, $E_e > 0$ and $E_e = -\frac{U_e}{U_E}$ at *e**, therefore $U_E = \Theta > 0$ at *e**; the second order condition at *e** becomes:

$$U_{ee} + 2U_{eE}E_e + U_E E_{ee} < 0 \qquad \textbf{(6)}$$

with

$$U_{ee} + 2U_{eE}E_e + U_E E_{ee} = \frac{-2U_{eE}U_e + U_{EE}^2 E_{ee}}{U_E} = \frac{-2\Theta_e(\Theta_e E + \Gamma_e) + \Theta^2 E_{ee}}{\Theta}$$

Certainly, $U_{ee} + 2U_{eE}E_e + U_E E_{ee} < 0$ at e^* when $2U_{eE}E_e + U_E E_{ee} < 0$, or $2U_{eE}E_e < U_E E_{ee}$, since $U_{ee} < 0$.[5]

If these conditions are fulfilled, equation (4) shows that the choice is the result of both moral and pleasure mechanisms. Moral values dominate the self-esteem producing mechanism while pleasure or welfare mechanisms set the rate at which the two goals can be substituted for each other. It is important to stress that the moral mechanism endogenizes self-esteem, enabling us to understand that a moral individual is not only she who gets pleasure from self-esteem but also – nay especially - she who behaves cooperatively in order to enhance her goal-function. Individuals differ from one another from a moral point of view, because they attach a different marginal utility to self-esteem or because they transform bad behavior into a greater or smaller amount of lost self-esteem. Our model takes both aspects into account. In particular, it shows that the optimum effort level will be higher for any *m* when there are self-esteem effects. These self-esteem effects set an endogenous lower limit at *e*.

To be really general as well as consistent with the experimental results of Principal-Agent situations, the proposed interpretation should be able to include cooperation among the possible outcomes. Cooperation should not be the only possible outcome. This is desirable also from the point of view of the degree of generality of the theory.

In fact, in a much quoted experiment Fehr *et al.* (1997) divided subjects into two sets, employers and employees, and considered their interaction in a Principal-Agent framework. First, they found that many employers offered generous wages and were reciprocated in terms of higher efforts from the employees, which resulted in a greater payoff for both. Secondly, they noticed that there existed, however, a significant difference between the level of effort agreed and the level of effort applied. They observed that this was not the behavior of a small group of fraudulent individuals, because only 26%, i.e. a small minority, of individuals honored their stated commitment.

[5] See also the Appendix for a more detailed case.

Nonetheless, this evidence "is compatible with the notion that the employers are purely self-interested, since their beneficent behavior vis-à-vis their employees was effective in increasing employer profits" (Gintis *et al.* 2003, p. 157). Allowing for the possibility that employers reward and punish employees, Fehr, Gächter, and Kirchsteiger observe an increase by up to 40% of the bet payoff of all subjects.[6] The comment by Gintis *et al.* (2003) is that "the subjects who assume the role of employee conform to internalized standards of reciprocity, even when they know there are no material repercussions from behaving in a self-interested manner. Moreover, subjects who assume the role of employer expect this behavior and are rewarded for acting accordingly. Finally, employers draw upon the internalized norm of rewarding good and punishing bad behavior when they are permitted to punish and employees expect this behavior and adjust their own effort levels accordingly" (Gintis *et al.* 2003, p. 157).

The above situation can be represented in a Principal-Agent framework where: it is in the Principal's interest to induce reciprocal behavior by the Agent, and the Agent may choose to cooperate – even independently of any material punishment – because she is a social being, feels part of a "community" (altruism) and, at least up to a certain extent, will loose self-esteem if she does not cooperate. However – and this is an important point in a rationality-based approach – such a mechanism will not work in every case and regardless of an accurate consideration of the relevant costs and benefits. The loss of self-esteem implied by lack of cooperation is not always high enough to ensure unlimited cooperation. In fact, as recalled above, experiments give support to the idea that there are limits to cooperative behavior.

Our attempt is to show, within a unique theoretical framework, that altruistic individuals do not necessarily choose cooperative behavior. Indeed, it is remarked that "strong reciprocators are inclined to compromise their morality to some extent" (Gintis *et al.* 2003). The approach we suggest seems capable of explaining what determines this willingness to compromise: much depends on the characteristics of the Agents' moral system and how self-esteem enters their utility functions. The next step is, therefore, to explain the conditions under which reciprocity emerges in our model.

[6] Employers punish fraudulent employees (68%), reward employees that over-fulfill their contracts (70%) and reward employees that honor their contracts (Gintis *et al.* 2003).

4. Self-esteem and reciprocity

We can now establish whether and how *m* influences the optimum *e*, this being the crucial condition for reciprocity effects. To accomplish this comparative static exercise we have to compute the second derivative with respect to *m* of the equilibrium condition above. To this end, and to simplify our analysis without loss of generality, we assume that the utility function (but not the *E-function*) is additive in its three variables (*m*, *E*, *e*). Therefore the problem becomes:

$$Max_{e,E}\ U = f_1(m) - f_2(e) + f_3(E)$$
$$\text{subject to}\quad E = E(m,e)$$

The additivity of the utility function allows us to assume that all mixed second derivatives in the utility function are zero. This makes the implicit differentiation of the optimum condition easier (4)**,** yielding:

$$\left[\left(U_{E,E}E_m + U_{E,E}E_e\frac{de}{dm}\right)E_e\right] + \left[U_E\left(E_{e,m} + E_{e,e}\frac{de}{dm}\right)\right] - U_{e,e}\frac{de}{dm} = 0$$

After some manipulation we get:

$$\frac{de}{dm} = -\frac{\left[U_{E,E}E_eE_m\right] + \left[U_EE_{e,m}\right]}{\left[U_{E,E}(E_e)^2\right] + \left[U_EE_{e.e}\right] + \left[U_{e,e}\right]} \qquad \textbf{(7)}$$

To establish whether *e* will change and in what direction, as *m* changes, we need to know the signs of all the relevant derivatives. Some are obvious. However, the general conclusion we can draw is that any result can come out and, more interestingly, the same individual may exhibit a different behavior, depending on some crucial conditions. We show both these results by assuming a further simplified version of the utility function, which fulfills second order conditions and highlights also the specific role that the moral process governing self-esteem plays.

$$U = f_1(m) - f_2(e) + \beta E(e,m) \qquad \textbf{(8)}$$

In this additive function the marginal utility of self-esteem is constant while the direct marginal dis-utility of effort is, as usual, increasing (i.e. $U_{ee} < 0$). As a consequence, the second derivative of *U* with respect to effort (taking account of both direct and indirect effects) will be negative. Therefore this function satisfies the required second order conditions for a maximum.

Since all the mixed second derivatives vanish, equation (7) simplifies to $\frac{de}{dm} = -\frac{\beta E_{em}}{\beta E_{ee} + U_{ee}}$

Hence: $\frac{de}{dm} > 0 \Leftrightarrow \frac{\beta E_{em}}{\beta E_{ee} + U_{ee}} < 0$, that is

$$\left[(\beta E_{em} > 0) \wedge (\beta E_{ee} + U_{ee} < 0)\right] \vee \left[(\beta E_{em} < 0) \wedge (\beta E_{ee} + U_{ee} > 0)\right]$$

we have two cases according to the sign of β. With positive self-esteem effects ($\beta > 0$) reciprocity will take place if:

$$\left[E_{em} > 0\right] \text{ and } [E_{ee} < \frac{-U_{ee}}{\beta}] \text{ or } \left[E_{em} < 0\right] \text{ and } [E_{ee} > \frac{-U_{ee}}{\beta}]$$

Given $U_{ee} < 0$ it is also easy to identify a sufficient condition for reciprocity, i.e.:

$$sign(E_{em}) \neq sign(E_{ee})$$

This very simple conditions makes it clear that reciprocity entirely depends on characteristics of the *E-function*. A sufficient condition for a subject to be a reciprocator is that her moral system is characterized by a process of self-esteem creation such that the marginal variation of *E* with respect to *e* reacts in the opposite way to a change in *m* and in *e*. It is important to stress that the signs of these derivatives may change in accordance with the values of *m* and *e*. Therefore, the same Agent may hold such moral values as call for reciprocation under some conditions, but not always.

This result undermines the usual assumption that cooperative behavior can be inferred from some fixed features of the Agents (their type). We could state, instead, that under very plausible

assumptions, the moral attitudes of the Agent are not enough to predict her behavior under any circumstances. There are no reciprocators, regardless of the other conditions.

5. Fairness and self-esteem

Self-esteem can help to understand how fairness influences the behavior of a rational agent. In particular, we will demonstrate why, also for an agent endowed with moral values, actual behavior may deviate from what is considered fair behavior. If fairness is, at least to some extent, socially determined we have the possibility to understand how social and individual values may interact. Given *m*, we can define:

Maximum effort, e_{max}, as the level of effort that leaves no surplus to the Agent;

Minimum effort, e_{min}, as the level of effort below which the Agent cannot go (for example, monitoring and sanctions for sure).

Fair effort, $e°$: the level of effort that the Agent deems fair in relation to the money she his paid.

Agents differ as to the determination of $e°$. Some could identify it with e_{min}, but in general it will be higher than that. We assume that an Agent will increase (decrease) her self-esteem when she provides an effort *e* greater (lower) than the fair effort, thereby enjoying a lower(higher)-than-fair surplus. A higher-than-fair surplus can be considered a *gift* (*G*). Therefore self-esteem and gift are one the opposite of the other. A reasonable assumption is that as *m* grows so does $e°$ with the result that, at the previous *e*, *E* will be reduced while *G* goes up.

If the Agent is offered a given compensation, and is free to choose her effort, it is as if she were determining her gift. Self-esteem effects ensure that *G* will not be as high as possible. Indeed, self-esteem can be seen as a mechanism setting a ceiling to the acceptable gift by the agent. If the agent's effort falls short of the fairness level, she experiences a loss of self-esteem, which might be small or large – and may have a small or large effect upon her utility. The point is to compare this moral loss with the loss of utility implied by a greater effort. Assuming that fairness will always drive behavior is to assume that the loss of utility due to self-esteem is always greater than the loss of utility implied by a greater effort. This is why self-esteem is a more crucial factor than fairness: it allows for the possibility that, despite her moral values and

her sensitivity to fairness, an Agent may choose a lower-than-fair effort. Indeed, self-esteem is the moral advantage (or disadvantage) for deviating from fairness.

A simplified formulation of the *G* is the following:

$$G_i = m - \theta e_i$$

i.e. given m there will be a gift (positive or negative) for any effort e_i. The equation implies that the fair effort – i.e. the effort level yielding $G = 0$ – is: $e^{\circ} = \frac{m}{\theta}$

In a Principal-Agent setting, the Principal knows that in order to elicit a certain effort on the part of the Agent, he has to offer a higher-than-fair *m* (therefore a gift). The weaker the self-esteem effects, the more *m* must increase. When a cooperation-prone individual – i.e. an individual with positive self-esteem effects - enters a Principal-Agent relationship playing the role of the Agent, the Principal may rationally consider the possibility of turning this proneness to his own advantage, by devising a contract that transforms it into an effective cooperative behavior. In order to achieve this result, the Principal has to bear a cost (much as the gift-type envisaged by Akerlof 1982), in the expectations that the Agent will reciprocate. This may be taken as the cost of an implicit contract based on trust. In this sense, trust, which creates cooperation, is costly and endogenous. It is worth stressing that cooperation-proneness is not the same as effective cooperation. Unlike other approaches, ours draws a clear distinction between propensity to cooperation (that may be understood as a form of altruism) and effective cooperation.

In a previous paper (Basili *et al.* 2004), we developed a model that made it possible to establish the conditions under which a contract based on trust may yield the Principal a higher return than alternative arrangements, like endogenous punishment or auditing. Building on that model we now consider how a cooperation-prone Agent may interfere with the choice of the best contract and how it could make the cooperative solution less costly. Our assumption on the utility function of the Agent and the relevance of self-esteem has, therefore, an impact on traditional Principal-Agent models and may alter the relative benefits of different contractual arrangements.

However, it is also possible that there is no finite gift that will elicit a certain level of effort or that, despite positive self-esteem effects, it is not worthwhile for the Principal to pay the gift required for that effort. Let us see how this works in a simple Moral Hazard case.

6. Self-esteem and Moral Hazard

Let us apply the above analysis to a simple moral hazard case. Typically, in such models the assumption is that the Agent can choose between a low and a high effort level. This has an impact upon our analysis because fairness becomes very important. In fact, the agent will never supply the high effort if the Principal is not paying compensation that includes a positive gift for the high effort. Therefore the problem is trivial when $G_h < 0$. When, on the contrary, this gift is non-negative the agent will attach the following utilities to the two effort levels:

$$U_h = (m - \theta e_h)$$
$$U_l = (m - \theta e_l) - \alpha E(G_h)$$

The expression for U_h is self-evident: the utility depends on the difference between the utility of compensation and the disutility of effort. On the other hand, U_l includes self-esteem effects. More precisely self-esteem here enters the function as a loss, i.e. the loss of self-esteem that the agent experiences when choosing the low effort, even though the Principal is offering a positive gift for the higher effort. The higher G_h, the greater the loss of self-esteem associated with the lower effort, which, in turn, translates into a lower utility through the constant term α.

The high effort will be incentive compatible if:

$$U_h \geq U_l$$
$$(m - \theta e_h) \geq (m - \theta e_l) - \alpha E(G_h), \text{ therefore:}$$
$$E(G_h) \geq \frac{\theta \Delta e}{\alpha}$$

In so far as E decreases with G_h there seems to be a sufficiently high G_h as to make e_h the best choice for the Agent. The marginal cost of differential effort and the marginal utility of self-esteem have a clear effect on the incentive-compatible G_h.: the former makes it greater, the latter smaller. It is also obvious that if $\alpha = 0$ (or $E_G = 0$) there will not be self-esteem effects and there is no finite G_h inducing cooperation. We need binding sanctions, as in the traditional shirking models, to obtain this result. Self-esteem effects may make e_h incentive-compatible regardless of

any explicit sanction, therefore easing moral hazard problems and making efficiency easier to achieve.

It is to be stressed, however, that this may not happen: indeed, self-esteem effects can be too weak. We can distinguish two cases:

i) self-esteem effects cannot compensate for the higher cost of the effort, however high G_h might be. This happens when there is no finite G_h such that $E(G_h) \geq \frac{\theta \Delta e}{\alpha}$;

ii) the required G_h may be too high from the Principal's point of view.

In conclusion, in a Moral Hazard situation self-esteem induced by a positive gift may (though not necessarily) give rise – through reciprocity – to a cooperative solution that would be impossible without such effects, however high the gift. Of course, costly sanctions with monitoring, as in the classical shirking model, could achieve the same result. Indeed, self-esteem can substitute for termination of contracts and provide different foundations for efficiency wages.

7. Self-esteem in Adverse Selection model

We will now see the impact of self-esteem in an adverse selection model. Among the various adverse selection models, that which fits our analysis better is the one analyzed in depth by Laffont and Martimort (2002).

Consider a Principal-Agent model in which the information asymmetry concerns the productivity of the agent, which could be high or low (efficient or inefficient agent), giving rise to low or high marginal costs, respectively. Let θ_H be the constant marginal cost of the efficient agent and θ_L the constant marginal cost of the inefficient agent. Since the principal cannot observe θ, he cannot equalize the marginal value of each agent's production, $S'(q)$, to its marginal cost.

If she were to offer a contract calling for different compensation levels on the basis of the quantity produced and equal to the respective marginal benefit, the efficient agent could simulate being inefficient (producing less) with a view to pocketing the information rent. The utility that the *H-Type* agent gets from making the high or the low efforts are, respectively, the following:

$$U_{H,H} = m_H - \theta_H q_H$$

$$U_{H,L} = m_L - \theta_H q_L$$

Therefore effort *H* will be incentive-compatible if:

$$m_H - \theta_H q_H > m_L - \theta_H q_L$$

Considering that:

$$\theta_h = \theta_l + \Delta\theta$$

and that a rational Principal will pay the reservation price to the *L-type* Agent:

$$U_L = m_L - \theta_L q_L = 0$$

The incentive-compatibility condition becomes:

$$m_H - \theta_H q_H \geq \Delta\theta q_L$$

$\Delta\theta q_L$ represents the information rent the *H-type* can reap and is equal to the difference between the two marginal costs at the low production levels. The principal is forced to pay a gift at least equal to that rent. This makes it impossible to write a first best contract.

Formally, the problem of the principal is that of maximizing profit, or the difference between the value of production and the associated costs. Profit is assumed to be a linear function of the quantity produced *q*. Let:

$S(q_H)$ and $S(q_L)$ be the value of production obtained with the efficient and inefficient agents;

m_H and m_L the compensation of the efficient and inefficient agents, respectively;

θ_H and θ_L the marginal cost of the efficient and inefficient agents;

$\Delta\theta q_L$ the value of the information rent;

v and (*1-v*) the probability to come across an efficient or inefficient Agent, respectively.

Given information asymmetry , the principal's profit maximization problem can be written as follows:

$$\max_{\{q_L, q_H\}} \{v[S(q_H) - m_H]\} + \{(1-v)[S(q_L) - \theta_L q_L]\}$$

such that

(i) $m_H - \theta_H q_H \geq m_L - \theta_H q_L$

(ii) $m_L - \theta_L q_L \geq m_H - \theta_L q_H$

(iii) $U_H \geq 0$

(iv) $U_L \geq 0$

As we know the incentive constraint implies that $m_H - \theta_H q_H \geq \Delta\theta q_L$, therefore the principal problem becomes:

$$\max_{\{q_L, q_H\}} \{v[S(q_H) - \theta_H q_H - \Delta\theta q_L]\} + \{(1-v)[S(q_L) - \theta_L q_L]\}$$

The solution of this problem calls for the same production as first-best for the efficient agent but a reduction with respect to first-best production for the inefficient agent. Indeed its marginal product will be higher than her marginal cost (implying a lower than optimal q_L).

$$S'(q_L) = \theta_L + \frac{v}{1-v}\Delta\theta$$

This result can be interpreted as follows: in order to induce the H-*type* agent not to choose the contract designed for the less efficient agent (and to pocket the information rent) the principal has to pay the *H*-agent a gift equal to the information rent. This makes it worthwhile for the principal to try to reduce that rent, a goal that she can achieve by decreasing q_l. Obviously, there is an optimal level for that reduction.

Let us now introduce self-esteem effects. We will show that they reduce the gift that the Principal has to pay below the information rent. This makes it possible to set a higher q_l and therefore to minimize the deviation from first best.

The presence of self-esteem allows for the Principal's incentive-compatible condition to be written as follows:

$$U_{H,H} = m_H - \theta_H q_H$$

as in the previous case. But

$$U_{H,L} = m_L - \theta_H q_L - \alpha_H G_H$$

where, without loss of generality, we have assumed that: $E = -\alpha G = -\gamma(G)G$, where $\alpha = \gamma(G)$ is a positive real valued monotone increasing function, that is self-esteem can be exchanged with gifts in a way that depends on the characteristic of the Agent and her utility function can be written in the following way:

If the *h* agent chooses the *L-Type* contract, in order to reap the information rent, she will suffer a negative self-esteem effect related to the gift received with reference to the *h* quantity.

Therefore she will not "cheat" if:

$m_H - \theta_H q_H \geq m_L - \theta_H q_L - \alpha_H G_H$

which, after some manipulation, and remembering that U_L=0, boils down to the following:

$G_h \geq \Delta\theta q_l - \alpha_H G_H$

Then: $G_H \geq \dfrac{\Delta\theta q_l}{(1+\alpha_H)}$

It is clear that the self-esteem effect makes the required G_H lower than the information rent, provided that $\alpha_H > 0$.

The higher α_H the lower the required *gift*. The latter will always be smaller than the information rent: self-esteem will bring about cooperation even at a monetary cost. The consequences for the Principal's optimal solution are immediate. After taking account of the constraints (and in particular of the easing of the incentive-compatible constraints) the function to be maximized becomes:

$$\max_{\{q_L, q_H\}} \{v[S(q_H) - \theta_H q_H - \frac{\Delta\theta q_L}{(1+\alpha_H)}]\} + \{(1-v)[S(q_L) - \theta_L q_L]\}$$

The optimal contract implies that there is:

- no distortion, also in this case, with respect to the first-best solution for the efficient agent;
- a downwards distortion, but smaller than before, with respect to the first-best solution for the less efficient Agent, such that: $S'(q_L^A) = \theta_L + \dfrac{v}{1-v}\left[\dfrac{\Delta\theta}{1+\alpha_H}\right]$

This proves that self-esteem effects, in this Adverse Selection model, have an impact on the optimal contract. In particular they imply a smaller gift for inducing cooperation and a slighter deviation from first-best quantities.

8. Concluding remarks

Traditional economic theory is undeniably too pessimistic as to the possibility of cooperation among strangers (Seabright 2004). Genetic relatedness is not the only condition for cooperation

to develop in situations where self-interest would make destructive opportunism the best course of action. A huge bulk of evidence can be invoked to this end. In particular, as Bowles and Gintis have argued, many humans seem to adhere to a strong reciprocity rule of behavior that implies the bearing of a personal cost in order to punish those members of the community who defect from cooperation.

However, the analytical foundation of this type of cooperation-prone behavior, and how it relates to rationality, has not been spelled out yet. In this paper we have advanced our own explanation, referring to Principal-Agent situations, relying on the notion of self-esteem and modeling cooperation-prone agents in terms both of a moral function transforming cooperation into self-esteem and of a utility function which includes self-esteem in its argument.

On the basis of this model we have shown that cooperation may be an outcome, depending both on the characteristics of the Agent and external conditions. This amounts to making a clear distinction between propensity to cooperation, on the one hand, and effective cooperation, on the other – two often muddled concepts. The model also rules out that cooperation or reciprocation depends just on the type of person.

Interestingly enough, our approach is coherent with Sen's most recent criticism of the standard rational model of choice based on the notion of self-centered welfare, that is a system in which a person's welfare depends only on her own consumptions and other features of the richness of her life, welfare maximization is the individual's only goal and an individual's choices are exclusively geared to the pursuit of selfish goals (Sen 2002, p. 34).

However, contrary to Sen, we advocate an operating theory of choice that makes people able to behave as maximizers, particularly with respect to endogenously determined moral variables. More specifically, individuals are endowed with a *moral system* which transforms their actions into self-esteem. The latter, as determined by such system, enters their utility function and contributes to define their choice within a utility maximizing process. Therefore, self-esteem brings about utility but its magnitude is determined by a moral system which lies outside the individual's preference system. Eventually, the inclusion of self-esteem in the utility function (which could very well be defined a goal-function) allows us to treat the choice problem as a typical maximization problem and give formal solution to it.

We have also shown the impact of our hypothesis on the best contract a Principal can offer in a Moral Hazard case and, in a more detailed way, in an Adverse Selection situation. In

particular, we have shown how our model can reduce the inefficiency of asymmetric information and lay the groundwork for a different approach to the best way to elicit cooperation.

References

Akerlof, G. A. (1982), Labour Contracts as Partial Gift Exchange, *Quarterly Journal of Economics* 97, 543-69.

Akerlof, G. A., Kranton R.E. (2000), Economics and Identity, *Quarterly Journal of Economics* 115, 715-753.

Basili, M., Duranti, C., Franzini, M.(2004), Networks, trust and institutional complementarities, *Rivista di Politica Economica* 1-2, 159-180.

Baumeister, R. F., Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117,* 497–529.

Benabou, R., Tirole J. (2006a), Belief in a just world and redistributive politics, *Quarterly Journal of Economics* 121, 699-746.

Benabou, R., Tirole J. (2006b), Incentives and prosocial behavior, *American Economic Review 96,1652-78.*

Bowles, S., Gintis, H. (2004), The evolution of strong reciprocity: cooperation in heterogeneous populations, *Theoretical Population Biology* 65, 17-28.

Fehr, E., Fischbacher, U., Gächter, S. (2002), Strong reciprocity, human cooperation, and the enforcement of social norms, *Human Nature* 13, 1-25.

Fehr, E., Gächter, S., Kirchsteiger, G. (1997), Reciprocity as a contract enforcement device: experimental evidence, *Econometrica* 65, 833–860.

Fehr, E., Gächter, S. (2000), Cooperation and punishment, *American Economic Review* 90, 980–994.

Fehr, E., Gächter, S. (2002), Altruistic punishment in Humans, *Nature* 415, 137–140.

Frey, B.S. (1998). *Not Just for the Money: An Economic Theory of Personal Motivation*, Edward Elgar.

Gintis, H. (2004). Modeling cooperation among self-interested agents: a critique. *mimeo*.

Gintis, H., Bowles, S., Boyd, R., Fehr, E. (2003), Explaining altruistic behavior in humans, *Evolution and Human Behavior* 24, 153-172.

Gintis, H., Bowles, S., Boyd, R., Fehr, E (2005), *Moral Sentiments and Material Interests,* MIT Press, Cambridge.

Kim, Y., Butzel, J. S., Ryan, R. M. (1998), *Interdependence and well-being: A function of culture and relatedness needs.* The International Society for the Study of Personal Relationships, Saratoga Spring, NY.

Laffont, J.J., Martimort, D. (2002), *The theory of Incentives,* University Press, Princeton.

Leary, M. R., Herbst K.C., Crary F. (2003), Finding pleasure in solitary activities: desire for aloneness or disinterest in social contact?, *Personality and Individual Differences* 35, 59-68.

Leary, M.R., Baumeister, R.F. (2000), *The nature and function of self-esteem: sociometer theory*. In M.Zanna (Ed.) *Advances in experimental social psychology*. San Diego Academic Press.

Munir, S. S., Jackson, D. W. (1997), Social support, need for support, and anxiety among women graduate students, *Psychological Reports* 80, 383–386.

Ostrom, E., Walker, J., Gardner, R. (1992), Covenants with and without a sword: self-governance is possible, *American Political Science Review* 86, 404–417.

Ryan, R. M., Connell, J. P. (1989), Perceived locus of causality and internalization, *Journal of Personality and Social Psychology* 57, 749-761.
Ryan, R. M., Rigby, S., King, K. (1993), Two types of religious internalization and their relations to religious orientations and mental health, *Journal of Personality and Social Psychology* 65, 586-596.
Ryan, R. M., Stiller, J., Lynch, J. H. (1994), Representations of relationships to teachers, parents, and friends as predictors of academic motivation and self-esteem, *Journal of Early Adolescence* 14, 226- 249.
Ryan, R. M., Kuhl, J., Deci, E. L. (1997), Nature and autonomy: organizational view of social and neurobiological aspects of self-regulation in behavior and development, *Development and Psychopathology* 9, 701-728.
Ryan, R.M., Deci, E.L., (2000), Self-determination theory and facilitation of intrinsic motivation, social development and well-being, *American Psychologist* 55, 68-78.
Seabright, P. (2004), *The company of strangers. A natural history of economic life*, University Press, Princeton
Sen, A. (2002), *Rationality and freedom*, Belknap Press, Cambridge Ma.
Yamagishi, T. (1986), The provision of a sanctioning system as a public good, *Journal of Personality and Social Psychology* 51, 110–116.

APPENDIX

From second order condition (**6**), at *e** utility is maximized when $2U_{eE}E_e + U_E E_{ee} < 0$, that is:

- if *E* is linear in *e*, then $E_{ee} = 0$, hence $2U_{eE}E_e + U_E E_{ee} < 0$ if $\Theta_e < 0$ at any point;
- if Θ is constant, then $2U_{eE}E_e + U_E E_{ee} < 0$ at *e** when $E_{ee} < 0$ there;
- if Θ is not constant, then $2U_{eE}E_e + U_E E_{ee} < 0$ at *e** if $\Theta_e < 0$ and $E_{ee} < 0$ at ay point.

In general, however, $U_{ee} + 2U_{eE}E_e + U_E E_{ee} < 0$ at e* if and only if

$$\frac{-2\Theta_e(\Theta_e E + \Gamma_e) + \Theta^2 E_{ee}}{\Theta} < -(\Theta_{ee}E + \Gamma_{ee}) \text{ at } e^*$$

but this condition gives rise to many more cases involving first and second order derivatives.

DEPFID WORKING PAPERS

DIPARTIMENTO DI POLITICA ECONOMICA, FINANZA E SVILUPPO
UNIVERSITÀ DI SIENA
PIAZZA S. FRANCESCO 7 I- 53100 SIENA
http://www.depfid.unisi.it/WorkingPapers/
ISSN 1972 - 361X