CONTRIBUTI DI RICERCA CRENOS

CRENoS
CENTRO RICERCHE
ECONOMICHE NORD SUD
Università di Cagliari
Università di Sassari

HOW TO INCENTIVE WHO? INTRA-PERSONAL AND INTER-PERSONAL MECHANISMS

Vittorio Pelligra

WORKING PAPERS

CUEC

2004/04

Vittorio Pelligra[1]
Department of Economics
Università di Cagliari

HOW TO INCENTIVE WHO? INTRA-PERSONAL AND INTER-PERSONAL MECHANISMS

Abstract
The paper focuses on the working of incentives both in parametric and strategic situations. It challenges some of the basic assumptions of the traditional model of economic agent which is usually assumed as self-interested and consequentialist. Psychological researches have stressed the descriptive limitations of that model and pointed out the relevance of other behavioral principles. Intrinsic motivations, reciprocity and trust being the most prominent among them. The paper analyses two different kinds of incentive mechanisms, namely, intra-personal and inter-personal and presents the results of an experiment that emphasize the empirical relevance of the latter. Besides providing a more descriptively adequate picture of interactive agency, such mechanisms have important normative implications that are discussed in the closing section.

JEL Classification: M52, C7, C91, D23.
Keywords: Incentives, reciprocity, trust, crowding-out, institutional design.

Acknowledgements: while writing this paper I have benefited from discussions with many people. I thank especially Michael Bacharach, Robert Frank, Benedetto Gui, Shaun Hargreaves-Heap, Chris Starmer, Robert Sugden, Tullio Usai and Stefano Zamagni. Giuseppe Argiolas, Luigino Bruni, Luca Crivelli, Bruno S. Frey and Margit Osterloht have read and commented on an earlier version of the paper. I wish to thank them all, while retaining the whole responsibility for the final result.

February 2004

---

[1]Vittorio Pelligra, V.le S.Ignazio 17, 09123 Cagliari – Italy. Tel. +39 070 6753319. E-mail: pelligra@unica.it

"One simple possibility is that economic models that ignore social psychology are incomplete descriptions of incentives in organizations. A more troubling possibility is that management practices based on economic models may dampen (or even destroy) non-economic realities such as intrinsic motivations and social relations"

R. Gibbons, (1998, p. 130)

1. Introduction

In the last decade or so, it has come to be widely accepted that "incentives are the essence of economics" (Predergast, 1999, p.7). There emerged a huge literature that analyses all the relevant aspects of how incentives work and how incentives provision systems have to be designed to foster efficiency. Nearly the totality of such a strand of studies is based on the assumption that agents are classical individual utility maximizers. By "classical individual utility maximiser", here I mean an agent who is actuated by the desire to achieve the preferred among the outcomes her actions could lead to, and, as corollary of that assumption, that material rewards play a dominant role in shaping agent's preference orderings. Although such an assumption is the most commonly used when modeling economic facts, a growing body of evidence has began, in the recent years, to be accumulated that casts doubts about its descriptive accuracy, both in parametric and strategic situations. Is becoming more and more clear that actual people, when confronted with certain classes of decisions problems, systematically do not conform to the predictions implied by traditional Expected Utility Theory and Game Theory. It is, nevertheless, possible to rationalize and explain those anomalies when we consider an enlarged version of rationality, where the maximization of material utility is no longer the only motive to action, and agents are not only interested in the outcomes of their choices but also in the processes that lead to those outcomes.

Here I shall explore ways to complete that picture of human

agency focusing my attention, as Gibbons suggests in the opening quotation, on the role of intrinsic motivations and social relations.

Some of the required amendments turn out to be compatible with the basic framework of Rational Choice Theory, while others require new concepts to be developed. In both cases, the introduction in the motivational structure of economic agents, of those psychologically grounded elements, enhances the predictive power of economic theory allowing us to understand at a deeper level the working of incentives in the decision-making process and their many important consequences on the normative side of institutional design.

The paper is organized as follow: first, the basic tenets of the classical agency theory are described (2); secondly, examples are presented that show the importance of *intra*-personal motivational mechanisms (3); thirdly, two of the core assumptions of the classical theory are critically analyzed (4); fourth, the role of *inter*-personal mechanisms is discussed (5), and a theoretical framework based on social approval, reciprocity and trust, is provided, that accounts for the empirical evidence discussed earlier (6). The implications for the institutional design are drawn (7). The conclusions close the paper (8).

2. How incentives are supposed to work… in theory

Agency theory assumes two kinds of subjects, the principal and the agent. It is usually hold that principals have some interest that cannot be pursued without the participation of the agent(s). Two main facts characterize the principal-agent relationship: firstly, their interests are conflicting, and secondly, agent's actions or characteristics are only imperfectly observable by the principal. Let me frame for simplicity, such a relation as a relation between employer and employee. The employer aims at maximizing profit which positively depends on the employee's effort. The employee is effort adverse and the level of effort actually performed cannot be directly observed by the principal. A wage, which constitutes a

cost for the employer and a source of utility for the employee, has to be provided to the employee to persuade her to perform some level of effort. In this sense, an incentive provision system (as implied by a contract, for instance) is a device designed to align employer's and employee's conflicting objectives. The wage provided by the employer must satisfy some requirements: from the employer's viewpoint, it has to be at least as higher as her reservation utility (participation constraint); and, given that, the employee will perform an effort which maximizes her net utility, the difference between the utility derived from the wage and the disutility from work has to be positive (incentive compatibility constraint).

Underlying such a classical theory there are three assumptions according to which:

1) the higher the paid wage, the higher the effort exerted;
2) decisions are path-independent;
3) given the asymmetry in the information structure, each time it will be possible, the agent will behave opportunistically.

Most of the recent developments in agency theory[2] aim at finding optimal compensation schemes capable to reduce the risk of opportunism while contextually making the contract attractive enough, to be accepted by the agent. Some of the contributions, for instance, seek to devise instruments to overcome the imperfect observability of agent's effort linking it to different observable signals. Others introduce dynamic considerations to extend the relationship over time. Such a move enables theorists to account for the strategic role of reputation. Others focus on the problems emerging from free-riding in teamwork. Although the lively debate ongoing in this area, it seems, nevertheless, that the three assumptions above mentioned have remained mostly unchallenged.

In what follows I shall discuss some phenomena that are at

[2] See Prendergast (1999) and Gibbons (1998) for complete surveys.

odds with those assumptions. In particular, in the next section I shall describe patterns of behavior that disconfirm the positive relationship, asserted by assumption 1), between material incentives and performance and that refute the consequentialist orientation of the classical agency theory, as described by assumption 2). This latter fact will shed light upon the simplistic nature of assumption 3).

3. How incentives seem to work… in practice: Intra-personal motivations.

It is usually believed that assumption 1) is a general law of human behavior. So general and well grounded that it has gained almost the status of an axiom. While, on the one hand, it is true that the assumption has received some degree of empirical validation, many, in fact, have found a strict correlation, on the aggregate, between increase in wage and increase in productivity, on the other hand, such a support can be variously interpreted. Firstly, it has to be noticed that the correlation between pay and productivity, which is the main empirical finding, may have a twofold explanation: first, the more you pay a subject, the more she will perform, as the assumption maintains, and secondly, the more you offer for a job, the higher the probability to attract skilled workers with higher productivity. Thus, the observed correlation may be explained both by the "incentive matter-argument" and the "selection-argument". Below I shall present examples that directly address these two arguments. Let me start with the latter.

This argument entails that increasing the monetary reward for a given task one is able to attract subjects better suited for that task. Consider the following examples.

In a seminal study on gift-giving, sociologist Richard Titmuss (1970), found that, despite its voluntary basis, the blood donation system adopted in England was more efficient (in terms of volume, quality and temporal availability of blood received), when compared with the remunerated system used in the United States, in those

years. Paying for giving blood leads to a reduced quality and quantity of blood supply. Trying to increase the supply of blood, Americans allowed commercial blood banks to pay donors for the blood they gave. But contrary to what they would have expected the result turned out to be worse along all the dimensions, when compared to the donor system.

In the study carried-out by Barkema (1995) two groups of managers subject to two different regimes of monitoring are compared. Group A is left with a larger degree of discretion, while group B is strictly monitored. The underlying idea is that as the monitoring becomes more stringent it will be easier to observe each agents' effort, and reward it accordingly. That strict correlation between effort and reward should exert a positive effect on the level of effort itself. However, Barkema reports a puzzling result, as in fact, the effect of such different regimes is that group B, that more strictly monitored, performs less than Group A, which on the contrary is not monitored.

A third example highlights the same counterproductive effect. Gneezy and Rustichini (2000) run an experiment using parents who have their children in a kindergarten. They analyzed for 20 weeks how subjects reacted to the introduction of a fine for those parents who were late at picking their children causing, this way, problems to the staff. The fine was intended, by imposing an additional cost, to reduce the number of latecomer parents, which, on the contrary, surprisingly increased. The increased number remained stable even after the fine was removed.

These examples highlight the problematic nature both of the "selection-argument" and "incentive matter-argument". As we will see Titmuss' explanation emphasizes the role of the differences in the kind of reasons that move *different* agents (i.e. different people react in different ways to the same material incentive), the other two examples, however, go further focusing on the difference in the kind of reasons that move the *same* agent (i.e. the same agent's performance may be positively or negatively affected by the same material incentive).

That latter phenomenon may be explained by considering the

so-called "crowding-out effect" (Frey, 1997; Frey – Oberholzer-Gee, 1997). In certain cases the subjects' willingness to perform a given action is decreased (instead of increased, as the intuition would suggest) by the prospect of a monetary or material reward. Titmuss stressed the importance of being aware that while there is people that perform certain tasks because of some form of prospective material reward (extrinsic motivation), others may have intrinsic reasons to perform the same task. That is important because introducing monetary reward will tend to select only the former subjects, as the blood donation example shows. The crowding-out theory affirms that the *same* person may have both extrinsic and intrinsic reasons, and that when one tries to increase, through material rewards, subjects' willingness to perform certain classes of actions ruled by intrinsic motivations, the underlying motivation is transformed, from intrinsic to extrinsic, and the overall result is a decrease in the agent' willingness to undertake those classes of actions.

4. A critique of the "Selection-argument" and the "Incentives Matter-argument"

The examples described above stress how the two aspects underlying assumption 1), I have dubbed "the Selection-argument" and the "Incentives Matter-argument", posses, in given circumstances, a limited empirical validity.

Consider Titmuss' explanation. His argument is built around the different kinds of motivations that underlie the same action (blood donation). In the case of the voluntary donor, the motive is altruistic and other-regarding, based on intrinsic reasons; while in the case of the remunerated donor, the motive is materially self-interested, based on extrinsic motivations. According to Titmuss, this leads to a self-selection of potential donors, intrinsically motivated in one case, and subjects more interested in the material, extrinsic reward, on the other; people subject to a stronger temptation of opportunism, of being, for instance, less truthful about the risk of serum hepatitis. What happened is that the

intrinsic motivated potential donors were displaced by the introduction of a monetary reward in the American commercial blood banks. Such a self-selection strongly affected the quality of the blood actually given.

While Titmuss stresses the risk of adverse selection implied by the use of material incentives, the other two examples show how the introduction of a monetary reward may discourage the *same* person to perform the very action the incentive was intended to encourage. The reasons behind such a phenomenon are many[3]; among them, particularly relevant are those concerning subject's self-determination and self-esteem (Frey, 1997, ch. I). Imposing an external system of material incentives produces in the subjects the impression of being controlled and of loosing the control of the situation (Rotter, 1966), so that, the *locus* of motivation shifts from internal to external. An external intervention, in the same way, may bring the message that subject's individual responsibility (and therefore also the potential merit) is not acknowledged and that her intrinsic motivation is rejected. "An intrinsically motivated person is denied the chance to display his or her own interest and involvement in an activity, when someone else offers a reward" (Frey, 1997, p.47). As a result of an underestimated responsibility, the subject experiences an impairment of her self-esteem, that, as a consequence, reduces her willingness to perform the given action. Furthermore, the way the subject perceives the external intervention plays a crucial role in determining the crowding-out or crowding-in effect. In fact, such an intervention can be seen either as *controlling* or as *supporting*, subjects' behavior. In the latter case, we observe a strengthening of subjects' other-regarding attitudes (crowding-in), in the former case, because of the impaired self-determination and self-esteem, we observe its weakening (crowding-out).

The examples described in the previous sections have usually been analyzed in term of the crowding-out effect. While this

[3] See Frey (1997), for a complete review.

explanation accounts for the so-called hidden cost of rewards in term of self-esteem and supportive or controlling interventions, this explanatory strategy cannot account for another class of behavioral anomalies that relates to assumption 2) and that refers to the relational aspects of the agents' motivational structure.

In the following section I shall discuss some examples of such anomalous behavior and then I will try to provide some elements for a unifying framework with which both classes of violations could be accounted for.

5. How incentives seem to work… in practice: Interpersonal Motivations

The examples I have been discussing so far show how incentives may be ineffective in achieving their aim because of the coexistence of both intrinsic and extrinsic motivations is often neglected and incentive systems are designed "as if" only extrinsic ones existed. Considering different sources of personal motivation is just the first step towards a more descriptively accurate picture of economic agency. A further step should include those sources of motivations which are relational, that is, that arises within an interpersonal relationships.

Jack Hirshleifer has argued some years ago that - "perhaps the grossest flaw in the economist's traditional view of human being is illustrated by the attention we devote to his *man-thing* activities as opposed to *man-man* activities (…) Economist have been studying only a chapter of the book of economic life" (1978, p.336).

Here I shall discuss examples that stress the inadequacy of such one-sided view of economic interaction, in particular I shall focus on the limitations of a consequentialist model of agency that assumes agents who are exclusively interested in the outcomes their actions lead to whose decisions are, thus, *path-independent*. Next I shall describe the role of relational incentives as trust and reciprocity.

Using the so-called "gift-exchange game" (figure 1), Fehr and Gatcher (1997) show how, contrary to the standard economic

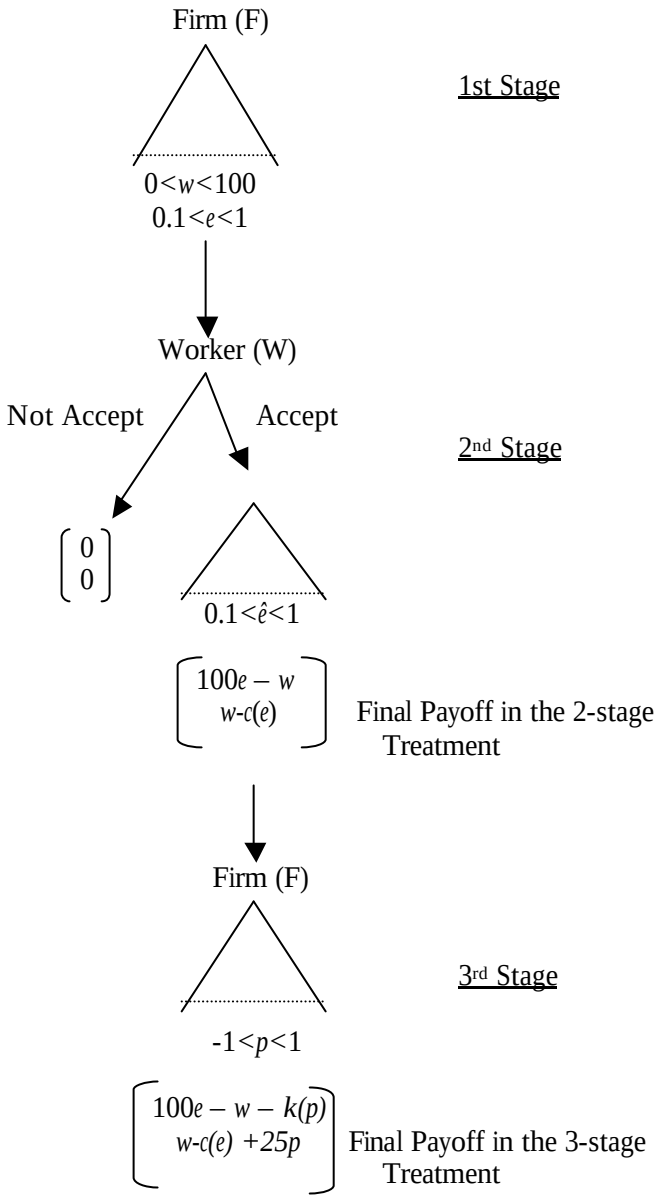
prediction, entering exchange relationships regulated by not-fully enforceable contracts could be efficient in terms of utility's maximization. The game is framed as a labor relationship, but the underlying logic can be applied to a wider range of contractual relations. Two different treatments are considered, the 2-stage game and a 3-stage version of the same game. The first stage is equal in both games and consists of a posted offer from six principals stating the offered wage and the required effort. Each principals offers only one contract per period. All the offers are communicated to the eight agents in a different room.

Agents are then required to make a choice (accept/not accept) among the offers to complete the first stage. In the second stage, agents have to determine an effort level which is associated, in a predetermined way, to a certain cost which is known also to the principals. Efforts and consequent costs, are chosen privately and communicated by an experimenter to the chosen principal who do not know agent's identity.

This kind of relationship between agents and principal has been so designed as to reflect the typical incompleteness of the majority of labor contracts, involving a great deal of discretion in agents' determination of their effort and a principal's relatively small ability in enforcing the desired level; only the minimum level of effort is, in fact, efficiently enforceable.

After decisions about wage and effort are made, payoffs are calculated and assigned to players: principal's payoff (profit) is positively affected by agents' effort and negatively by the wage being paid and agent's payoff (utility) are proportional to wage and negatively affected by the level of effort. If the offered wage is higher than a certain threshold value and the effort expressed equals the minimum enforceable level, then the principal faces the risk of a severe loss. That is why backward induction suggests that the principal should offer only the minimum wage and the agent should exert the minimum level of effort.

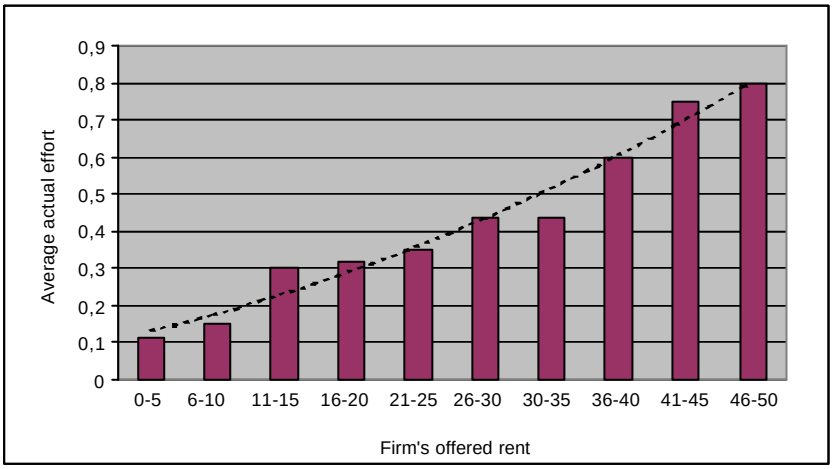
Figure 1: “The Gift Exchange Game” (Fehr and Gatcher, 1997)



The 3-stage treatment is similar to the 2-stage one; the third additional stage, gives the principal the possibility of a (positive or negative) reciprocation, that is, the opportunity to reward or to punish the agents. Principals may choose to reward or to punish, at their own cost, agents' actions by increasing of lowering agents' final payoffs.

In both treatments the game theoretical prediction is the sub-game perfect Nash equilibrium involving the lowest level of efforts and the lowest level of wage. The equilibrium is the same in both versions because punishment and rewards are considered in game theoretical terms, "non-credible threats" and therefore should not affect the strategy choices.

Figure 2: "Relationship between offered rent and actual effort" (Fehr and Gatcher, 1997)



Despite these predictions, the experimental results are different as far as the desired level of efforts, the wages offered and the actual level of effort are concerned. The desired level of effort is on average 0.70 and 0.72, respectively in the 2stage and the 3stage game, much above the 0.1 level theoretically expected. These levels are constant over all the periods without any sign of convergence towards the equilibrium. The effort levels obtained are, much higher than would have been expected: an average of 0.44 and 0.63 in respectively, the 2-stage and the 3-stage game, and it is positively correlated with the rent offered (figure 2). The actual effort is on average higher in the 3-stage game. This shows how a "non credible threat" can significantly affect behaviors.

These findings are coherent with Herbert Simon's (1991) firm belief according to which "in most organizations, employees contribute much more to goal achievement than the minimum that could be extracted from them by supervisory enforcement" (p. 31-32).

A fifth example will stress the importance of having potential material incentives or disincentives and not using them. In their (2002), Fehr and List observe the behavior of a sample of Chief Executive Officers (CEOs) in various forms of investment games. In particular they considered two variants of the game, one where the first player is endowed with a certain amount of money and has to decide which part of the endowment, if any, to send to the second player. If she send a positive amount that is tripled and given to the second player who can, in turn, decide how much, if any, to send back. The second variant of the game is similar apart from the fact that a sanction can be implemented by player one, if, what she receives back from player two, is less that what she had expected. The game theoretical solution to both games is for player one to send nothing to player two, and for player two to send back nothing to player one. However, what has been observed is that not only, most of the experimental subjects decided to invest and to send back some positive amount of money, but most strikingly, such amount increases when the sanction are available but, actually, not used. This

result in interpreted by Fehr and List as a sign that: "the *availability* of the sanctioning threat can be quite productive (…) If principals *voluntarily* refrain from using the punishment threat when it is available, agents exhibit significantly more trustworthiness than if the punishment threat is not available. Thus if agents face no punishment threat, the mere fact the principal could have used the punishment option affects the agent's trustworthiness in a positive manner" (2002, p.2). Similar results are reported by Blair and Stout (2000) with regard to the corporate law and its enforcement in the American judicial practices.

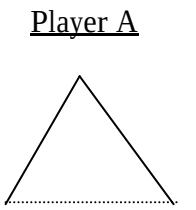
5.1. An illustrative experiment

We now present the results of an experiment conducted on 121 subjects which is explicitly designed to explore the procedural versus the consequentialist attitude of players in strategic environments. Each player was asked to play a dictator game and a modified version of an investment game as described in figure 3a and 3b.In the dictator game player 1 has to divide and endowment of 30 Euros between her and Player 2. Her decision immediately implement the final distribution. In the "gratuitous" investment game, on the contrary, Player 1 has to decide whether to keep her endowment of 10 euros or giving it to Player 2. In the former case Player 2 automatically gets a payoff of 30 Euros, while in the second case she has to decide how to divide player 1's endowment that has been tripled by the experimenter amounting now to 30 euros.

What is important in this setting is that from the second player's viewpoint both games being played are equivalent, with the only (theoretically non-significant) difference that in the investment game there is a bygone, that is, a branch of the game which is not reached. Traditional game theory produce two distinct prediction for these games:

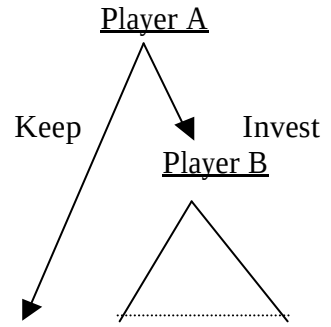
a) we should observe a minimal offer in both games;
b) the offers should not differ significantly from one game to the other.

Figure 3a: Dictator-game



30-*y* average y ≈ 5

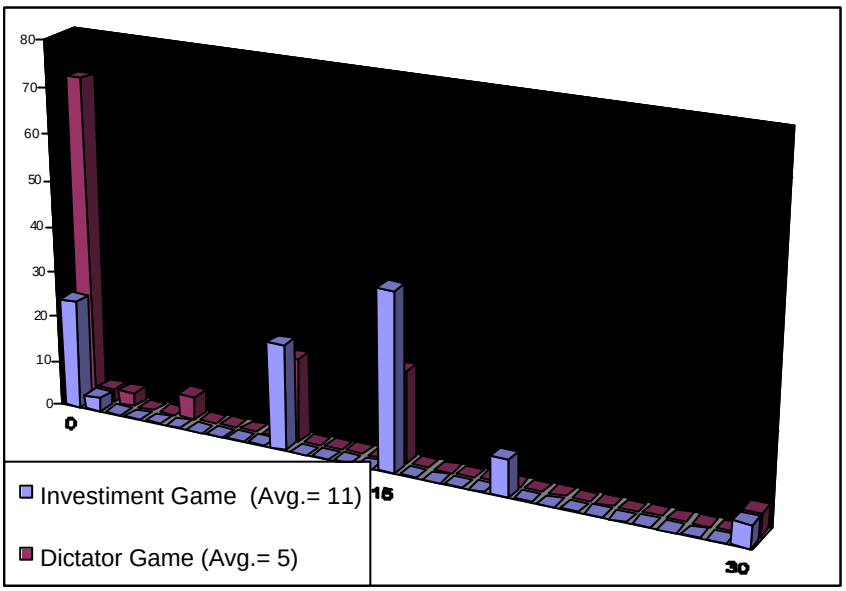
Figure 3b: "Gratuitous" investment-game



30 30-*y* average *y* ≈ 11

However, the results show, that despite the two games are identical from the point of view of the consequences they lead to, in the investment game player 2 sends on average 11 Euros, while in the dictator game the average offer is of only 5 Euros (see figure 4).

Figure 4. “Offers distribution”



Such results are consistent with those emerged from the study conducted by Falk, Fehr and Fishbacher (1998). They compare a Reduced Ultimatum Game with a Reduced Best-Shot Game. In the first game the first player is asked to divide her endowment of 10, making an offer to her opponent. If her opponent accepts, then the offer is implemented and the game ends, otherwise, if the offer is refused both players get nothing. The possible offers are only two (that is why this version of the game is called "reduced"): either 2 or 5. In the second game, the Reduced Best-Shot Game we have the same structure but the possible offers are either 2 or 8, with consequent symmetrical

outcome equal to (2,8) and (8,2). Is the offer of 2 to be considered in the same way in both games? Experimental subjects suggest not. The authors find, in fact, that the frequency of refusal for the offer of 2 in the Reduced Best-Shot Game is significantly lower than that for the same offer in the Reduced Ultimatum Game. How are we to explain such a difference? Assuming that players are altruist or inequity averse does not help. Those theories, in fact, because of their forward-looking nature, cannot explain such patterns of behavior.

The alternative strategy is that of assuming that people care not only about outcome but about outcomes *and* intentions. Intentions can be inferred by observing the chosen strategy not in isolation but within the entire strategy set; this gives to the agent the possibility to evaluate what the opponent could have done and did not. In the Reduced Ultimatum Game, in fact, the first players who offered 2, could have chosen an equal split but they did not. Because of this comparison between the actual choice and the other available options, such a choice is perceived as more unkind than the same choice in the Reduced Best-Shot Game, and punished by a higher rate of refusals.

Similar considerations about the difference between explanations based on distributional concerns and theories based on agents' responsiveness, for instance, reciprocity models, emerge from the experimental results presented in Nelson (2002).

5.2. Consequntialism, Deontologism or Relationality?

I am not endorsing here a deontological attitude. I believe that radical deontologism, implying unconditional choices would produce an even less satisfactory picture of human agency. What I am suggesting here is that players' behaviours are neither merely deontological nor narrowly consequentialist, but that their decisions are, so-to-speak, path-dependent.

The evidence I have been describing suggests that people are much more "relational" that the classical theory implies. By relational I intend that players are open to others' influences, that

they react to others' first and second orders believes and expectations and that, when playing a game, their preferences are subject to a process of endogenous (trans)formation.

In this respect, the classical defense of consequentialism refers to the fact that *in principle* the description of a consequence of an action could be thought as including the action itself that produced the state of affairs the consequence describes. As in Sen's example "A state of affairs where Brutus kills Caesar is not just one where Caesar has expired. Is one in which the killing of Caesar by Brutus figures (…) and if we decide to leave out this fact in the description of the state of affaires this is just a decision to be silent about one part of that state" (1985, pp. 181-2).

Such a defensive move is plausible only when it is possible to separate consequentialism from welfarism. It is in fact the welfarist component of the more general utilitaristic perspective that makes the actions irrelevant to the description of the end-states. But the fact that utilitarianism incorporates both consequentialism and welfarism makes it extremely difficult to keep distinct the two sides of the same coin. Even though we adopt an enlarged version of the concept of consequence, to say that action *x* is to be preferred to action *y,* is different from saying that the consequence of action *x, plus* action *x,* is better than the consequence of action *y, plus* action *y.* As we enlarge the concept of consequence such a distinction shrinks but never completely disappears. Otherwise consequentialism would become equivalent to non-consequentialism (Griffin, 1992). But consequentialism requires not only to evaluate actions entirely on the basis of their consequences but also not to consider everything else. "The result is a descriptive impoverishment from *many* perspective, *including,* among others, normative relevance" (Sen, 1980, p. 360)

Moreover, the contrast between descriptive adequacy and consequentialism has strengthened once experimental economists have started trying to operationalize the concept of consequence in real life situation attaching objective payoffs to end nodes in games.

In summarizing it can be said that the fact that bygones

matter means that we cannot focus our attention exclusively on the consequences of choices as the unique criteria that subjects use to evaluate their perspective actions; also the different paths or interactions that lead to those outcomes should become theoretically relevant.

6. Relational incentives: reciprocity and trust

The evidence I have been presenting so far should have made clear how retaining assumption 1 and 2 means, somewhat, to reduce the descriptive and explanatory power of agency theory. Moreover, from the critique of assumption 2 it follows that opportunistic behavior is not as pervasive as assumption 3 suggests. Actual behavior, in fact, depends on the particular structure of the interaction players are placed in and on the behavior of *all* the relevant players. The data seems to suggest that a more satisfactory version of agency theory should incorporate a behavioral explanation of how incentives works and a more realistic model of economic agent. In this section I shall discuss two of the principles that may help in building such a theory, namely, reciprocity and trust.

The bargaining experiments I have been discussing highlight that real people tend to behave kindly to those who have been kind to them and unkindly with who has been unkind. Those behaviors respond to the norm of (positive and negative) reciprocity. It has been shown both empirically and theoretically (Rabin, 1993) how, in certain conditions, such a norm may offset the effect of the material payoffs in the strategic decision making. The effect of reciprocity may lead the subject to act in a way that appears to be contrary to her material self-interest. The idea of reciprocity is ultimately based on the *joint* effects of material and psychological incentives. That means that the motivation that triggers (positive or negative) reciprocal behavior is ultimately based on material incentives. The perceived kindness that elicits reciprocal behavior, is a measure of material benefit that an agent's choice attributes to another player.

Following this logic, Rabin, in his model, formalizes the idea

in a way that player A feels motivated to reciprocate when, from player B's expected action, she can get a *material* payoff higher that the equitable one (which constitutes a measure of fair distribution). Such an increase refers to material payoffs. Acting in a way which repays (ignores) such an expected material gain, leads to A a psychological gain (loss).

Some of the other experiments I have reported, however, show that yet another behavioral principle may be at work in similar situations, namely, trust, in particular, in the form of trust responsiveness (Bacharach, Guerra, Zizzo, 2002; Pelligra, 2002a, 2002b, 2004). The main feature of trust in this particular meaning, refers to the fact that an explicit act of trust has the peculiarity of "inducing" or "eliciting", to some degree, a trustworthy response. In this respect is said that trust is responsive or self-fulfilling. Suppose we have two agents, A and B: according to the "responsive trust" conception, B's trustworthiness may be induced by A's choosing a trustful course of action (like, for instance, player 1 sending money to player 2 in the gratuitous investment game, or offering an above-the-minimum wage in the gift exchange game). This kind of inducement assumes the existence of a psychological mechanism according to which, A's trustful action, motivates B to reward such trustfulness, making him behave trustworthily, even though such a behavior implies some material cost. I call such a psychological mechanism "trust responsiveness".

While reciprocity, as I have said, is based on the joint action of both material and psychological incentives, trust responsiveness, is exclusively based on a psychological-moral motivation. It is the (trustful) action that motivates the trustee, and not the potential beneficial consequences for her material wealth that she may derive from it. By trusting me, you manifest to me your expectations on my behavior. If I consciously fulfill (frustrate) it, I get an increase (decrease) in my psychological payoff. Put it in another way, while reciprocity in Rabin's theory is the act of conferring benefits on people who have previously *materially* benefited you, trust responsiveness is the act of conferring benefits on people who have shown that they expect you to do so, and have willingly exposed

themselves to harm in the event that you act on material self-interest.

Such a difference becomes clear when one considers a certain class of situations where reciprocity- and trust responsiveness-based behaviors are not observationally equivalent, but leads to divergent predictions.

Consider the logic of reciprocity applied to the particular Trust Game of figure 5a below. We know that reciprocity takes the form of returning kindness for kindness and unkindness for unkindness. In this game that means that, if A expects B to play R then "A playing R" is perceived as kind by B who in turn feels motivated to play R. Consider now the instance of the Trust Game depicted in figure 5b.

What result would the logic of reciprocity produce in this particular game? Since now, if A expects B to play R, "A playing R" does not provide any benefit to B, if B expects A to play R, therefore the logic of reciprocity would suggest B to play L, straight away. What about trust responsiveness? If, on the contrary, we consider a trust responsive B player, we would observe him playing R in both situations, despite the difference in the material benefits he would obtain from A's move.

Figure 5a.
The Trust Game

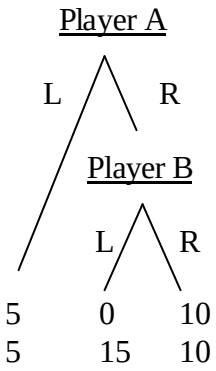
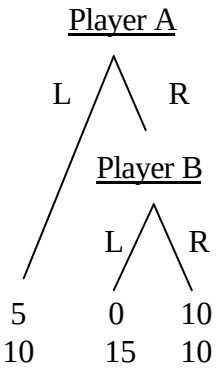


Figure 5b.
The Gratuitous Trust Game



I have labeled the latter particular form of the trust game as the "Gratuitous Trust Game" expressly to emphasize that in this game any instance of co-operative behavior, cannot be thought as an exchange, though delayed, of material benefits.

These examples shed light on the difference between trust responsiveness and reciprocity on the issue of material/non material incentives to action. In the Gratuitous Trust Game, in fact, a B player would get the same material payoff, both from A's trustful choice (followed by B's trustworthy one) and from A's distrustful choice (L). Differently from the simple trust game, the only way for B to get a payoff higher than 10, in the Gratuitous Trust Game, is to behave opportunistically. That means that if one observes a B player playing R in the gratuitous version of the trust game, that behavior cannot be explained in terms of reciprocity, while it still remains consistent with an explanation based on the trust responsiveness hypothesis. The empirical relevance of such trust-responsive pattern of behavior is documented in Bacharach, Guerra, Zizzo, (2002) and Pelligra (2002a).

It should have become clear, by now, how the logic of

trust responsiveness may subsume the crowding-out effect. A material reward, a payment to do something that the agent was willing to do on the basis of intrinsic motivations, convey a sense of distrust that elicits opportunism.

As we know from the functioning of trust responsiveness, agents have a tendency to conform to others' manifest expectations. In the case of crowding-out effects, a conflict arises between internal and external reasons for agents' action. Consider a worker that performs poorly when monitored. She will disappoint her principal but, at the same time, she will react, on the basis of her sense of worth and self-esteem, to an act of hostility by her employer (a distrustful monitoring). While crowding motivation theory explains different behaviors assuming the existence of two different types of motives for action (intrinsic and extrinsic), trust responsiveness suggests that the different effects of (dis)incentives depends on the relative weight of the social and psychological consequences. Following Smith's *Theory of Moral Sentiments*, we can isolate three main sources of motivations: material, social and internal. Material motivations are related to the material outcomes of our actions; social motivations refers to the degree of social approval and praise that derives from the others' observing our actions. The third source of motivation, the internal, refers to the extent to which we think our own actions really deserve such a praise or blame. We want, in fact, not only to want to be praise but also to be praiseworthy". In Smith's own words:

> *Man naturally desires, not only to be loved, but to be lovely; or to be that thing which is the natural and proper object of love. He naturally dreads, not only to be hated, but to be hateful; or to be that thing which is the natural and proper object of hatred. He desires, not only praise, but praiseworthiness; or to be that thing which, though it should be praised by nobody, is, however, the natural and proper object of praise. He dreads, not only blame, but blame worthiness; or to be that thing which, though it should be blamed by nobody, is, however, the natural and proper object of blame.* (Smith, 1759/1976, sec.III.2.1)

7. Implications for policy and institutional design

While the motivational crowding-out, although not standard in agency theory can easily be incorporated in a classical rational choice model (Frey, Oberholzer-Gee, 1997), reciprocity and trust, on the other hand, imply that players are responsive to other players' behavior, that is, that payoffs are endogenous. Such a characteristic makes impossible to reconcile those principles with the standard model of game theory which is essentially consequentialist (Geneakoplos, Pearce, Stacchetti, 1989; Rabin, 1993).

Apart from their theoretical implications, the behavioral regularities I have been describing, have also important normative implications, in particular as far as the design of optimal incentives schemes is concerned. If people, in fact, draw psychological utility from self-esteem, social approval, reciprocal behavior and trustworthiness, those elements should be incorporated in the incentive systems and managed as important organizational resources. To avoid crowding-out effects, material rewards have to be carefully engineered to convey a sense of support instead of a sense of control that may backfire reducing subjects' own intrinsic motivation to perform the same action the incentives were supposed to favor.

Moreover, being reciprocity and trust motivational active elements, interaction schemes should be created within the organizations capable to activate those elements. A too strict monitoring, for instance, reducing the room for socially approved intrinsic trustworthiness, risks to signal a sense of mistrust that may increase opportunism and shirking, instead of reducing it, as Barkema's experiment clearly shows. To lay down even the most specific details of a contract may subtract space for reciprocal actions and may, as we saw in the gift-exchange game, to pareto-inferior outcomes.

Reciprocity and trust are norms enforced by interpersonal social pressure and, as Fehr and Falk (2002) have

recently noticed, such norms are likely to produce strategic complementarity among agents' actions. That means that the efficacy of the social approval motive depends on others' people behavior. If others are sensitive to approval and disapproval from their peers, each agent's action will find in the desire for others' praise, a strong psychological incentive, otherwise, the material reward will always be predominant. That opens the possibility for Pareto-rankable multiple equilibria to emerge. The transition from an inefficient equilibrium to a more efficient one would depend, then, on how social incentives work within the organization. An "atmosphere" can be then created where the desire for others' praise may be encouraged and that may, ultimately favors agents' praiseworthiness or at least agents' desire for social praise.

As Matthew Rabin clearly put it: "Economics should be concerned not only with the efficient allocation of material goods, but also with designing institutions such that people are happy about the way they interact with others" (1993, p. 1283). Precisely in this sense, agents' internal, intrinsic motivations should be considered as economic realities, as well as their desire for material reward. A careful design is needed to avoid the risk of stimulating a clash between the two kinds of motivations.

Intrinsic motivations, trust and reciprocity may be thought of as important, sometime, crucial assets of each organization. Neglecting this point would produce counterproductive effects with consequent waste of resources and harm for the organization's efficiency.

8. Conclusions

In this paper I have critically discussed and challenged the three main assumptions of agency theory, respectively, that the higher the wage the higher the effort exerted, that people are interested only in the outcomes their actions lead to and that, given the asymmetry in the information structure, whenever possible, the agent will behave opportunistically. According to this standard view the agent are to be considered "self-interest-

seeker with guile", to use Williamson's expression (1985).

My alternative position maintains that:

1) because of the interaction of intrinsic and extrinsic motives, may well be possible that the use of material rewards to incentive intrinsically motivated activities, turns out to reduce the performance of such activities (motivational crowding-out);

2) people are responsive to others' behavior, therefore the same outcome may be differently evaluated depending on the strategies that lead to it;

3) for this reason people tend to behave opportunistically much less than the classical theory would suggest. Trust and reciprocity are principles that account for the observed anomalous behaviors.

Those points have important implications for the activity of institutional design. The desire for social approval, trust and reciprocity, have to be considered as organizational resources that should be carefully engineered to avoid counterproductive effects and to improve the overall performance.

References

[1] - Bacharach, M., Guerra, G., Zizzo, D., 2001. Is Trust Self-Fulfilling? An Experimental Study. mimeo, BREB, University of Oxford.

[2] - Barkema, H., 1995. Do Top Managers Work Harder When They Are Monitored?. Kyklos, 48, 19-42.

[3] - Blair, M., Stout, L., 2000. Trust, Trustworthiness, and the Behavioral Foundations of Corporate Law. Working Paper, Georgetown University Law Center.

[4] - Bohnet, I., Frey, B.S., 1999. Social Distance and Other-Regarding Behavior in Dictator Game: Comment. American Economic Review 89, 335-39.

[5] - Bolle, F., 1998. Does Trust Pay?. Europa Universitat Frankfurt (Oder), Working Paper.

[6] - Geneakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological Games and Sequential Rationality. Game and Economic Behavior 1,60-79

[7] - Gibbons, R., 1998. Incentives in Organisations. Journal of Economic Perspectives 12, 15-132

[8] - Gneezy, U., Rustichini, A., 2000. A Fine is a Price. Journal of Legal Studies 29, 1-17.

[9] - Falk, A., Fher, E., Fischbacher, U., 1998. Intentions Matter. Mimeo, University of Zurich.

[10] - Fehr, E., Falk, A., 2002. Psychological Foundations of Incentives. European Economic Review 46, 687-724.

[11] - Fehr, E. and Gachter, S., 1997. How effective are Trust- and Reciprocity-Based Incentives? In: A. Ben-Ner and L. Putternam (Eds.), Economics, Values and Organizations. Cambridge University Press, Cambridge.

[12] - Fehr, E., List J., 2002. The Hidden costs and Returns of Incentives – Trust and Trustworthiness among CEOs. Mimeo, University of Zurich.

[13] - Frey, B.S., 1997. Not Just for the Money: An Economic Theory of Personal Motivation. Elgar, Cheltenham, UK.

[14] - Frey, B.S., Oberholzer-Gee F., 1997. The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out. American Economic Review 87, 746-755.

[15] - Hirschleifer, J., 1978. Natural Economy Versus political Economy. Journal of Social and Biological Structures 1, 319-37.

[16] - Griffin, J., 1992. The human Good and the Ambitions of Consequentialism. Social Philosophy and Policy 9:118-132.

[17] - Nelson, R.W., 2002. Equity or Intention: it is the Thought that Counts. Journal of Economic Behavior and Organization 48, 423-430.

[18] - Pelligra, V., 2002. Fiducia R(el)azionale, in: P.L. Sacco and S. Zamagni (Eds.), Complessità Relazionale: Fondamenti del comportamento economico. Il Mulino, Bologna.

[19] - Pelligra, V., 2002b. Rispondenza Fiduciaria: Principi e Implicazioni per la Progettazione Istituzionale. Stato e Mercato 65, 330-353.

[20] - Pelligra V., 2003. Path-(in)dependence?: an experimental investigation. Mimeo, Università di Cagliari.

[21] - Pelligra, V., 2004. Under trusting Eyes: the responsive nature of trust", in B. Gui and R. Sugden (Eds), The Economics of Sociality. Cambridge: Cambridge University Press. (forthcoming).

[22] - Predergast, C., 1999. The provision of Incentives in Firms. Journal of Economic Literature 37, 7-63.

[23] - Rabin, M., 1993. Incorporating Fairness in Game Theory. American Economic Review 83, 1281-301.

[24] - Rotter, J., 1966. Generalized Expectancy for Internal versus External Control Reinforcement, Psychological Monographs 80, 609-16.

[25] - Sen, A., 1980. Description as Choice. Oxford Economic Papers 32, 353-369.

[26] - Sen, A., 1985. Well-Being, Agency and Freedom. The Journal of Philosophy 82, 169-221.

[27] - Simon, H.A., 1991. Organizations and Markets. Journal of Economic Perspectives 5, 25-44.

[28] - Smith, Adam, 1759/1976, The Theory of Moral Sentiments (Liberty Classics, Indianapolis).

[29] - Titmuss, Richard, 1970, The Gift Relationship (Allen and Uwin, London).

[30] - Williamson, Oliver, 1985, The Economic Institutions of Capitalism (The Free Press, New York).