Forthcoming in Bartolini, S., Bruni, L., Porta, P.L. (Eds.), *Policies for Happiness*, Oxford University Press.

Promoting Trust through Institutional Design

Vittorio Pelligra

University of Cagliari

&

CRENoS

Abstract:

If, as economists, we are really interested in promoting people's happiness, we should devote much attention and effort in understanding the process that leads us from individual and interactive choices to happy or unhappy outcomes. In interactive situations, this process is almost always constrained by the rules of the game and quite often such rules are designed according to the prevalent economic theory, in particular, according to principal-agent theory. In recent years, however, such prevalent view has been challenged, on a descriptive ground, by behavioral economists. Intrinsic motivations, gift-giving, trust, trustworthiness and reciprocity are just some of the factors that affect real people behavior and that are inconsistent with the standard economic model where agents are supposed to be selfish, consequentialist and opportunistic. The paper discusses the role of two different sets of behavioral principles, namely intra-personal and inter-personal mechanisms. These mechanisms describe how real people react to individual internally-generated and social relationally-generated incentive. Since laws and institutions can be though of as (dis)incentive providing systems designed to favor a desired conduct and hinder dysfunctional behaviors, the understanding of the dynamic behind intra-personal and inter-personal mechanisms is essential to the design of efficient institutions and normative schemes.

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

Acknowledgement: while writing this paper I have benefited from discussions with many people: Stefano Bartolini, Luigino Bruni, Bruno S. Frey, Benedetto Gui, Shaun Hargreaves-Heap, Margit Osterloh, Alessandra Smerilli, Robert Sugden, Tullio Usai, Stefano Zamagni ans Luca Zarri. I wish to thank them all, while retaining the whole responsibility for the final result.

1. Introduction.

"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" (Rabin, 1993, p. 1283). If, as economists, really we are interested in people's happiness, we should devote much attention and effort in understanding the process that leads us from individual and interactive choices to happy or unhappy outcomes. In interactive situations, this process is almost always constrained by rules of the game and quite often such rules are designed according to the prevalent economic theory, in particular, according to principal-agent theory. In recent years, however, such prevalent view has been challenged, on a descriptive ground, by behavioral economists. Intrinsic motivations, gift-giving, trust, trustworthiness and reciprocity are just some of the psychological factors that affect real people behavior and that, at the same time, are difficult to account for within the standard economic model where agents are supposed to be selfish, consequentialist and opportunistic.
The descriptive inadequacy of the standard model of *homo economicus* is not neutral from a positive viewpoint. As noted by Robert Gibbons, in fact: "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" (1998, p. 130). In much the same vein studies on social dilemmas, happiness and economics and social interactions (see Bruni & Porta, 2005; Gui & Sugden, 2005), all point out that the standard concept of economic rationality is inadequate to account for these phenomena and it has to be amended in order to incorporate relational factors such as trust, reciprocity, intentions and social emotions.

This shift from an ultra-simplified view of the economic agent to a more realistic one is similar to what happened when the profession realized that the negative implications of an ultra-simplified view of the functioning of the market. Asymmetric information, externalities, increasing returns and other real life phenomena have been incorporated into more sophisticated models that not only lead to a better understanding of the economic realities but also provide useful insights on how to manage the emerging complexity. Much in the spirit of Gibbons' quotation, the development of more realistic models helps us in delivering policies and designing new institutions capable to foster markets' efficiency and enhance people's welfare.

Behavioral economics, with its emphasis on cognitive limitations but also on pro-social behaviors represents a second shift that deepens our understanding on the functioning of incentives. Since, laws and institutions can be though of as (dis)incentive providing systems designed to favor a desired conduct and hinder dysfunctional behaviors, the understanding of the dynamic behind intra-personal and inter-personal mechanisms appears to be essential for policy-makers, normative and institutional architects. In this paper I discuss some of the research

lines that point in this direction by questioning, in particular, the assumptions of self-interest, consequentialism and opportunism.

Economic models are based on the assumption that agents are individual utility maximizers, that is, actuated by the desire to achieve the preferred among the outcomes their actions could lead to. A corollary of this assumption is that material rewards play a dominant role in shaping agent's preference orderings However, in the last decade a growing body of theoretical paradoxes and empirical evidence have begun to be accumulated that cast doubts about the descriptive accuracy of this model. These limitations call for an enlarged version of rationality, where the maximization of material utility is no longer the only motive to action, and agents are described as non-pure-consequentialists. Here I shall explore ways to complete that picture of human agency focusing, as Gibbons suggests in the opening quotation, on the role of intrinsic motivations and social relations. Having discussed, the basic tenets of the classical agency theory (2), the paper analyses the working of two different kinds of incentive mechanisms, namely intra-personal (3) and inter-personal and discusses experimental results that emphasize the empirical relevance of the latter (4). A theoretical framework based on social approval, reciprocity and trust is provided, that accounts for the empirical evidence discussed earlier (5) and produces important normative and institutional implications (6). Conclusions close the paper (7).

2. The basic assumptions of agency theory.

Agency theory assumes two kinds of subjects, the principal and the agent. Principals have some interest that cannot be pursued without the participation of the agent(s). The participation in the interest of the principal is a source of disutility for the agents. Thus, the agency problem reduces to find a way to make the agent "incapsulate" principal's interests. Two main facts characterize the principal-agent relationship so described: first, their interests are conflicting and second, agent's actions or characteristics are only imperfectly observable by the principal. Consider for simplicity a relation between employer and employee. The employer aims at maximizing profits, which positively depend on the employee's effort. The employee, in turn, is effort–adverse and the principal cannot directly observe the level of effort actually performed. An incentive (the wage), which constitutes a cost for the employer and a source of utility for the employee, has to be provided to the employee in order to persuade her to carry out some level of effort. In this sense, an incentive provision system (as implied by a contract, for instance) is a device designed to align the conflicting objectives of employer and employee. The wage provided by the employer must satisfy two 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's effort will maximize her net utility.
This classical theory is grounded on three main assumptions:
1) Monotonicity: the higher the wage, the higher the effort exerted;
2) Consequentialism: people are interested only in the outcomes their actions lead to;

3) Opportunism: given the asymmetry in the informational structure, whenever it will be possible, the agent will behave opportunistically.

Most of the recent developments in agency theory[1] aim at finding optimal compensation schemes capable to reduce the risk of opportunism while contextually making the contract enough attractive to be accepted by the agent. Notwithstanding the lively debate ongoing in this area, the three assumptions above mentioned have remained mostly unchallenged. In what follows I shall discuss some phenomena that are at odds with those assumptions. In particular, in the next section I shall describe patterns of behavior that disconfirm the positive relationship between material incentives and performance, asserted by the assumption 1), that refute the consequentialist orientation of the classical agency theory, as described by assumption 2) and that will shed light upon the far too simplistic nature of assumption 3).

3. Extrinsic and intrinsic incentives: Intra-personal mechanisms.

Economic wisdom maintains that assumption 1) is a general law of human behavior. So general and well established that it has gained almost the status of an axiom. While, on the one hand, it is true that the relationship between incentive and effort implied by the assumption has received some degree of empirical validation, 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: the more you pay a subject, the more she will perform, as the assumption maintains, or 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 an "incentive matter-argument" and a "selection-argument". Below I shall present examples that directly address these two arguments. Let consider first the latter.

The "selection-argument" entails that high monetary rewards are 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 blood banks to pay donors for the blood they gave. However, contrary to what they would have expected, the result turned out to be worse along all the dimensions, if compared to the donor system.

Titmuss' explanation is based on the different kinds of motivations that may underlie the same action (blood donation). In the case of the voluntary donor, in fact, the motive is altruistic and other-regarding, grounded on intrinsic reasons;

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

while in the case of the remunerated donor, the motive is materially self-interested, grounded on extrinsic motivations. According to Titmuss, the introduction of monetary incentives lead to a self-selection of potential donors, attracting those more interested in the material reward. This people is subject to a stronger temptation of opportunism, of being, for instance, less truthful about the risk of serum hepatitis. At the same time, the intrinsic motivated donors were displaced by the introduction of a monetary reward. Consistently with a sort of "Gresham's Law of human motivations" such a self-selection strongly affected the quality of the blood actually given.

In a 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 large degree of discretion, while group B is strictly monitored. The underlying idea is that as the monitoring become more stringent it will be easier to observe each agents' effort and reward it accordingly. That strict correlation between effort and reward should induce an increase in the level of effort itself. However, Barkema reports a puzzling result as, in fact, the performance of group B, the one more strictly monitored, is poorer than that of Group A, which 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 during 20 weeks how people reacted to the introduction of a fine for those parents who were late at picking their children. The fine was intended, by imposing an additional cost to dysfunctional behavior, to reduce the number of latecomer parents. At the end of the 20 weeks, however, Gneezy and Rustichini observed an increase in the number of latecomer right in the group subject to the fine. And yet, this number remained stable even after the fine was removed.

A potential explanation that may account for all these phenomena refers to the so-called "motivational crowding-out theory" (Frey, 1997; Frey & Oberholzer-Gee, 1997). In certain cases the subjects' willingness to perform a given action is decreased (instead of increased, as the theory would suggest) by the prospect of a monetary or material reward. The motivational crowding-out theory assumes that the *same* person may have both extrinsic and intrinsic reasons, and that when one tries to incentive, through extrinsic rewards, certain classes of actions ruled by intrinsic motivations, the underlying motivation is transformed, from intrinsic to extrinsic, and the overall result may turn out to be a decrease in the agent' willingness to perform those actions.

While Titmuss stresses the risk of adverse selection associated to the provision 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[2]; among them, particularly relevant are those concerning subject's self-determination and self-esteem (Frey, 1997, ch. I). Crowding-out occurs when - "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" (p.47). Also 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

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

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).

While motivational crowding-out theory accounts for the 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 related 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 I will try to provide some elements for a unifying framework with which both classes of violations could be accommodated.

4. Consequentialist or intentional rationality?

The examples I have been discussing so far show how material incentives may be ineffective (and even counterproductive) in achieving their aim because of the often neglected, complex interplay of intrinsic and extrinsic motivations. However, considering different sources of personal motivation is just the first step towards a more descriptively accurate picture of economic agency. A next step, in fact, should include those sources of motivations that originate within an interpersonal relationship, those sources of motivations that are relational or belief-dependent (Geanakoplous, Pearce & Stacchetti, 1989; Battigalli & Dufwenberg, 2005). Jack Hirshleifer claimed 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 a 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. According this view, Players order their preferences over actions according to their preferences over the consequences these actions lead to. If action *a* produces outcome *a,* action *b* produces outcome *b* and action *c* yields to outcome g, action *a* is preferred to *b* and to *c*, as long as outcome *a* is preferred to *b* and the latter to *g*. Experimental evidence, however, show that the same outcome may be variously assessed depending on the history of play that lead to it. That means that when deciding how to behave in a strategic situation, real people take into account not only the prospective outcomes of their joint actions but also other backward looking elements.

Consider the games G1 and G2 depicted in figure 1 and 2 (Falk *et al*., 1998). Player *A* makes an offer to *B* of either 2 or 5 in G1, or 2 or 8 in G2, player *B* can either accept or reject *A*'s offer. If she accepts, the division is implemented and the players are paid accordingly, if she refuses to accept, both players get nothing.

[insert figures 1 and 2 about here]

Given that players are assumed to be self-interested, rational maximizers and consequentialist, the theory predicts first, that *A*s would offer the smallest amount of money, and second, that *B* would not reject any positive offer. Given that the outcomes conditional to *A*s choosing "H" are identical in G1 and G2, we should observe in both situations the same, or a very similar rate of refusal. However, when real people play the games we observe, first, that *B*s reject *A*s' offers more often than would have been "rational"; second, and more interestingly, that the number of refusals is higher in G1 (44%) than in G2 (18%). That result is surprising because, once *A* proposes 2 to *B*, from *B*'s perspective the two games are identical, at least in term of outcomes, payoff distribution and consequences.

A similar result is reported in Pelligra (2003), where respondents' behaviour in a "gratuitous investment game" is compared to that of the proposers in a "dictator game". The data show that, despite from the respondents' viewpoint (in the investment game) and from the proposers' viewpoint (in the dictator game), the two games are identical in terms of the consequences they lead to, in the investment game the respondents sends back on average 11 Euros, while in the dictator game the average offer is of only 5 Euros.

Analogously, in their 2002 experiment, 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 player 2. If she sends a positive amount, that is tripled and given to player 2, 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 the proposer has to decide whether to implement a sanction, if what she receives back from player 2 is less that what she had expected. The game theoretical solution to both games is for player 1 to send nothing to player 2, and for player 2 to send back nothing to player 1. However, what has been observed is not only that most of the experimental subjects decided to invest and to send back substantial amounts of money but, most strikingly, that such amount increases in the treatment when the sanctions are available even if, actually, not used. This result is 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).

All these situations, although theoretically equivalent in term of outcomes, are different if we consider the history of the play, in particular with respect, not only to what each player did but also to what they could have done and did not. What emerges is that for real people bygones are relevant as well are the choices actually made. Other studies (Blount, 1995; Charness, 1998; Nelson, 2002; Charness & Levine, 2005) have highlighted similar patterns of anomalous behaviour.

These results emphasize two crucial points: first, people are more trustful and trustworthy that predicted by the theory, and second, people do not care only about consequences, but also about others' 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 figure out what the opponent could have done and did not.

5. Reciprocity and trust as relational incentives: inter-personal mechanisms.

The evidence I have been presenting so far should have made clear how retaining assumptions 1 and 2 means 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 seems to suggest. 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. What the data suggest is that a more satisfactory version of agency theory should incorporate a behavioral explanation of how incentives work and a more realistic model of the relational dynamics of the economic agent. In this section I shall discuss two of the principles that may help in developing the relational part of the theory, namely, reciprocity and trust. These two concepts and their effects are tightly intertwined and quite often confused. In most theoretical and experimental studies trust is considered as an expectation of a reciprocal behavior. However, the relation between two concepts, although very close, is richer and subtler than that. Not always trust can be considered as an expectation of reciprocity since there are instances of trusting interactions where trustworthy behavior seems not to be motivated by reciprocity (i.e. the "gratuitous" trust game in Pelligra, 2005). Therefore trust and reciprocity should be kept conceptually separated.

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 (Fehr and Gächter, 2000; Fehr and Schmidt, 2002) and theoretically (Rabin, 1993; Dufwenberg & Kirchsteiger, 2006) 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 benefits that an agent's choice attributes to another player. Another behavioral principle is consistent with that evidence and is also able to explain behaviours that are at odds with the norm of reciprocity, namely trust responsiveness (Bacharach, Guerra, Zizzo, 2002; Pelligra, 2002a, 2002b, 2005). 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 an 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".

It is important to notice that, since the logic of trust responsiveness is symmetrical, that is, an act of trust elicits trustworthiness as well as an act of distrust induces opportunism, trust responsiveness may synthetically subsume the crowding-out effect.

In the case of crowding-out effect, 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 associated, material, social and psychological consequences (Pelligra, 2005). While the motivational crowding-out, although non–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).

6. Normative and institutional implications.

As Gibbons acutely points out, there exists a possibility that: "management practices based on economic models may dampen (or even destroy) non-economic realities such as intrinsic motivations and social relations" (1998, p. 130). If constitutions, institutions, contracts and any other set of rules are based on the assumption that real people behave as the so-called *homo economicus* who obeys to assumptions 1, 2 and 3 of agency theory, there may emerge the risk of counterproductive effects. That is the main reason behind the normative implications that an enlarged picture of economic agents produces for institutional design and policy.

If people 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 resources. To avoid conflicts between intrinsic and extrinsic motivations, 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 willingness 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 communities capable to activate those elements also because, as James Coleman argued, when someone ask another to be trustworthy – "he does so because it brings him a needed benefit; he

does not consider that it does the other a benefit as well by adding to a drawing found of social capital available in time of need" (1998, p.S117). A too strict, mistrustful monitoring, reducing the room for socially approved intrinsic trustworthiness, goes in the opposite direction and 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 lead to Pareto-inferior outcomes. Trust is a matter of signals. The trustee must know that the trustor is relying upon her. Is this signal that motivates the trustor to perform trustworthily. If the trustor hadn't trusted the trustee and signaled that to her, perhaps she would not have been trustworthy. Designing rules, one has to leave room for this sort of signal.

Reciprocity and trust are norms enforced also by social (dis)approval. 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 community. An "atmosphere" can be then created where the desire for others' praise may be encouraged and that may ultimately favor agents' praiseworthiness or at least agents' desire for social praise.

If we want to take seriously Rabin's view about the connection between economic design and people's happiness, 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 community. Neglecting this point would produce counterproductive effects with consequent waste of resources and harm for the community's efficiency.

The activity of institutional design aims, on the one hand, at regulating people's interactions at different levels, by creating formal institutions capable to coordinate agents' interests and to direct them towards the desired outcome and, on the other hand, at sustaining and promote with sanctions and other forms of (dis)incentives certain classes of behaviors.

Institutional design is traditionally based on an anthropological model that is very similar to that of the *homo economicus* (Goodin, 1996). According to this view, since the harm associated to a violation of the norm is bigger than the benefit deriving from compliance, rules must be created assuming that people are "sensible knaves", to use Hume's expression, that is, opportunistic and self-interested. In Hume's own words: "Fixing the several check and controls of the constitution, every man ought to be supposed a knave, and to have any other end in all his actions that private interest" (1875: 117-8). This designing philosophy has been defined as *deviant-centered* (Pettit, 1996) and is ultimately based on the very same assumptions of the agency theory I have critically discussed above. The

basic idea is that of creating more deterrence than is necessary for most people, in order to make sure that it would be sufficient for all. However, as we have already noticed, this exercise is not neutral, as, in fact, it tends to erode most of people's intrinsic motivations and interpersonal trust. The *theoretical* descriptive inadequacy of the standard assumptions, thus, may lead to devise systems of rules that *practically* discourage compliance.

The drawbacks of the *deviant-centered* system call for an alternative designing approach to be developed. An approach where the focus is on how to foster compliance more than how to discourage deviance. This approach can be thus defined as *complier-centered.* Within this approach, all the motivational sources, intrinsic and extrinsic, are taken into account and carefully managed to avoid conflicts, and the relational incentives are used to foster compliance while at the same time discouraging opportunism.

At the core of the *complier-centered* systems three general principles stand:

i) first, before providing (dis)incentives is important to select the pool of individuals that will be subject to those (dis)incentives;
ii) second, the (dis)incentives aims principally to favor cooperative behaviors than to punish the opportunistic ones.
iii) third, (dis)incentives must not neglect the risk of opportunism.

The *complier-centered* systems as well as the *deviant-centered* one, have at their center what seems to be an inescapable dilemma: how to discourage knaves if any sanction can erode the willingness to comply of intrinsic motivated people? A way out of the dilemma could be represented by what Ayres and Braithwaite (1992) define "dynamic regulatory institutions". These institutions embody systems of rules primarily based on the power of dialogue and persuasion more than deterrence. Rules and sanction are organized in a hierarchy in which lower level sanctions have informal nature and are associated to minor violations; they are neutral, in the sense that do not erode intrinsic motivations. Higher-level sanctions are formal and are associated only to repeated violations; these sanctions are more concerned with minimizing harm that safeguard intrinsic motivation. This system is organized as a pyramid: at the basis (large and diffuse) there is dialogue and persuasion based on trust, in the middle, deterrence based on calculativeness and on top (small and rare), incapacitation aimed at reducing the harm caused by the incompetent and untrustworthy agents..

Here I cannot go further in analyzing these principles, I have done it diffusely elsewhere (Pelligra, 2002b), but, by now, the message should be clear enough. The different typologies of interpersonal relationships through which communities activates their resources should be accurately structured in order not to provoke crowding-out effects, and, more generally, trust and social capital erosion. Work relations, public administration, consumers and the civil society at large are all environments where this recommendation applies.

7. 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).

Our alternative position maintains that:

1) because of the interaction between intrinsic and extrinsic motives, it 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.

4) the tension between rules and trust (deviant-centered systems vs. complier-centered systems) turns out not to be inescapable, though it calls for a changing in the designing logic of relational structures.

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

References

Ayres, I., Braithwaite, J., 1992. Responsive Regulation. Oxford University Press, Oxford.

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

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

Battigalli, P., Dufwenberg, M. 2005. Dynamic Psychological Games. Mimeo. IGIER-Bocconi

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

Blount, S., 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. Organizational Behavior and Human Decision Processes 63:131–144.

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

Bruni, L., Porta, P. L., 2005. Handbook of Happiness in Economics. Edward Elgar, Cheltenham.

Charness, G., 1998. Attribution and reciprocity in a simulated labor market: An experimental investigation. Working paper, Universitat Pompeu Fabra.

Charness, G., Levine, D., 2005. The Road to Hell: An Experimental Study of Intentions, Mimeo, University of California SB.

Coleman, J., 1998. Social Capital in the Creation of Human Capital. American Journal of Sociology 94 (Supplement), S95-S120.

Dufwenberg, M., & Kirchsteiger, G., 2004. A Theory of Sequential Reciprocity. Games and Economic Behavior 47:268-98

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

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

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

Goodin, R., (Ed.), 1996. The Theory of Institutional Design. Cambridge University Press, Cambridge.

Falk, A., Fehr, E., Fischbacher, U., 1998. Intentions Matter. Mimeo, University of Zurich.

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

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.

Fehr, E., Gächter, S., 2000. Fairness and Retaliation: The Economics of Reciprocity.
Journal of Economic Perspectives 14: 159-181.
Fehr, E., List J., 2002. The Hidden costs and Returns of Incentives – Trust and Trustworthiness among CEOs. Mimeo, University of Zurich.
Fehr, E., Schmidt, K., 2002. Theories of Fairness and Reciprocity: Evidence and Economic Applications. In: M. Dewatripont, L. Hansen and St. Turnovsky (Eds.), Advances in Economics and Econometrics - 8th World Congress, Econometric Society Monographs. Cambridge, Cambridge University Press.
Frey, Bruno S., 1997. Not Just for the Money: An Economic Theory of Personal Motivation. Elgar, Cheltenham, UK.
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.
Gibbons, R., 1998. Incentives in organizations. Journal of Economic Perspectives 12, 115-132.
Gui, B., Sugden, R., (Eds.) 2005. Economics and Social Interaction: Accounting for Interpersonal Relations. Cambridge University Press, Cambridge.
Hirschleifer, J., 1978. Natural Economy Versus political Economy. Journal of Social and Biological Structures 1, 319-37
Hume, D., 1875. On the Independence of Parliament, in Philosophical Works, Vol.III, Green and Grose Ed.: London.
Nelson, R.W., 2002. Equity or Intention: it is the Thought that Counts. Journal of Economic Behavior and Organization 48, 423-430.
Pelligra, V., 2002. Fiducia R(el)azionale, in: P.L. Sacco and S. Zamagni (Eds.), Complessità Relazionale: Fondamenti del comportamento economico. Il Mulino, Bologna.
Pelligra, V., 2002b. Rispondenza Fiduciaria: Principi e Implicazioni per la Progettazione Istituzionale. Stato e Mercato 65, 330-353.
Pelligra V., 2003, "Consequences vs. Procedures : an Experimental investigation", Mimeo, Università di Cagliari.
Pelligra, V., 2005. Under trusting Eyes: The Responsive Nature of Trust", in B. Gui and R. Sugden (Eds), Economics and Social Interaction: Accounting for Interpersonal Relations. Cambridge University Press, Cambridge.
Pettit, P., 1996. Institutional Design and Rational Choice, in: Goodin, R., (Ed.) The Theory of Institutional Design. Cambridge University Press, Cambridge.
Rabin, M., 1993. Incorporating Fairness in Game Theory. American Economic Review 83, 1281-301.
Rotter, J., 1966. Generalized Expectancy for Internal versus External Control Reinforcement, Psychological Monographs 80, 609-16.
Smith, Adam, 1759/1976, The Theory of Moral Sentiments (Liberty Classics, Indianapolis).
Titmuss, Richard, 1970, The Gift Relationship (Allen and Uwin, London).
Williamson, Oliver, 1985, The Economic Institutions of Capitalism (The Free Press, New York).

Figure 1: “G1 - Reduced Ultimatum Game”

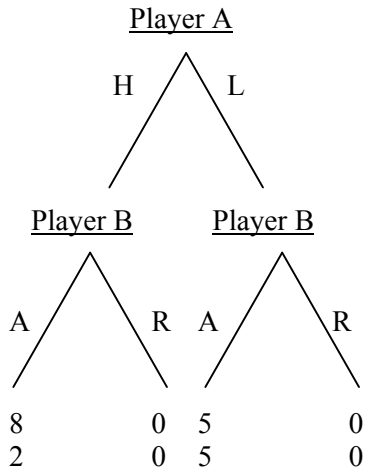

Figure 2: “G2 - Reduced Best-Shot Game”