FACOLTÀ DI ECONOMIA
UNIVERSITÀ DI BOLOGNA
SEDE DI FORLÌ

Corso di Laurea in Economia
delle Imprese Cooperative
e delle Organizzazioni Nonprofit

SOCIAL PREFERENCES:

from the experimental lab to the economic theory

Gianandrea Staffiero

Working Paper n. 18
Giugno 2005

in collaborazione con

AICCON
CULTURA COOPERAZIONE | NONPROFIT

Gianandrea Staffiero
University of Navarra.

Informazioni :
Facoltà di Economia di Forlì - Corso di Laurea in Economia delle Imprese Cooperative e delle ONP
Tel. 0543-374620 – Fax 0543-374618 e-mail: nonprofit@spfo.unibo.it   website: www.ecofo.unibo.it

Abstract

We present a wide collection of experiments which show how human behavior deviates substantially with respect to the predictions derived from standard homo economicus assumptions.

Then we review the theoretical literature that this evidence has stimulated. In particular some models are found to be consistent with evidence from a large set of games. As fundamental differences exist among these proposals, new experiments were devised to contrast their effectiveness in predicting behavior. We argue that inequality aversion models are to be preferred to intention based models because the additional predictive power the latter may have comes at a very high cost of complexity. We also find that equality considerations are more relevant than efficiency motives in most economically relevant settings. Results are not conclusive and this gives scope to further research over these issues.

1. Introduction

In this paper we are going to explore experimental evidence and related theories over an increasingly lively debate in economics: whether it is reasonable to leave the standard assumption of selfishness as a reasonable approximation of agents' motivation and, once this is agreed upon, which new models are better fit to explain players' behavior.

This is only a part of what experimental economists are working on. In particular, the other pillar of most economic "mainstream" analysis, instrumental rationality and common knowledge of it, is another hot area of experimental investigation. Indeed, we are going to refer to this aspect insofar as it is related to the focus of our analysis. In fact, one line of attack towards other-regarding preferences as an aspect emerging from experiments consists in questioning whether experimental subjects really understand the kind of game they are playing. For instance, when players cooperate in ”prisoner's dilemma" situations it may be the case that they do not understand which choices really maximize their payoffs, either because the artificiality of the experimental lab situation does not provide the clues typically given in real life, or because they do not exert a high comprehension effort, due to the limited amount of money involved in most (though not all) experiments.

The fact that people are not completely selfish can hardly come as a surprise, and indeed is reckoned, for instance, in overlapping generation models in which the welfare of the next generation is taken into account (see Blanchard and Fischer, 1989). However, it is often assumed, often tacitly, that non-selfish motivations are the exception, so that for most situations of economic interest we could safely assume selfishness by all agents to derive game-theoretic predictions. Technically, this is equivalent to have own payoffs as the only argument of any players' utility functions.

To get back to the same, classical example, we find among the implications of this approach that if the payoff structure defines a ”prisoner's dilemma” situation, the same happens when we consider utility levels. As a consequence, defection is the best response not only to itself but also to a cooperative move by the opponent, and this holds independently of whether the game is played simultaneously or sequentially, one-shot or with repetitions, as long as they are finite (backward induction arguments apply). Possibly quite a lot of people would deny that defection is their response to cooperation in all interactions which are not going to be repeated for an indefinite amount of times. Experiments constitute a test about whether this would be a void claim or, rather, conditional cooperators constitute an important part of the population. We are in fact going to see that this is indeed the case and that suitable assumptions on preferences makes this behavior compatible with the application of familiar game theoretic techniques. Moreover, these same assumptions explain also a good bunch of other evidence where behavior does not appear to follow what arises from self interest by all players and the common knowledge of this aspect. The evidence we are going to present follows for the most part a logical consistency, which makes it hard to believe that it comes out of mistakes.

Another point already worth stressing is the economic relevance of those situations represented in experiments. While in some settings people's behavior is compatible both with selfishness and with alternative assumptions, so that for instance competitive behavior is not necessarily indication of the former, in others we find that the departure from the assumption of pure selfishness provides intuitive explanations for important phenomena such as, for instance, a higher frequency of contract incompleteness with respect to what would be caused only by excessive costs or impossibility to write complete contracts.

In the next section we are going to focus on the experiments designed especially to test whether selfishness can explain behavior in some simple games. In section 3 we explore data from prominent games such as the prisoner's dilemma and the public good games. These sections illustrate how experimental evidence prompts the search for alternative models of preferences. Section 4 shows some of these theoretical proposals, while section 5 is focused on experiments designed in order to contrast these theories, in order to look for the most successful one in explaining human behavior. Section 6 concludes.

2. Sacrifice to hurt or to help: ultimatum, gift exchange, trust and hot response games

Consider the following game. One player proposes a division of a pie worth, say, $10. The other is given the opportunity to accept the proposed division, in which case it is simply implemented, or to reject it, which would make both players get nothing. Which choices should players take? If we assume that each player is selfish and knows that the other is selfish as well, it is quite easy to derive the subgame perfect Nash equilibrium (Nash, 1950, Selten, 1975). The second player should accept any division which entails a positive amount for her, no matter how small; only a (10,0) should give her doubts about what to do. The proposer, expecting this, should leave the smallest possible amount, 1 cent, as responder's payoff and $9.99 for himself, unless for some reasons he is 100% sure that even (100,0) would be accepted, in which case he would not leave the cent either. Other Nash equilibria exist, with much more favorable outcomes for responders, but they involve non-credible threats. For instance, the best response to a responder's strategy "accept only 50 or more" is the strategy with an offer of precisely 50. This would be a Nash equilibrium, but not a subgame perfect one: if an offer is lower than 50, rejecting it is not the payoff maximizing response in the subgame consisting in the single decision node of the responders.

There are now loads of experimental evidence (see in particular Güth et al., 1982, Forsythe et al., 1994 and the review in Roth, 1995) that players behave quite differently, so that payoff distributions are much more equitable, and 50%- 50% is usually the modal division and 40% a typical average value. It may be argued that first movers simply do not understand the game well enough, a it requires a bit of inspection, though simple it can be. However, as pointed out in Roth, 1995, it happens that rejection rates make those type of offer quite sensible, as substantially uneven divisions are often rejected. On the other hand, it appears tougher to presume that second movers do not understand the situation presented at their decision node and possibly become surprised when they realize that rejection results in (0,0) distribution. It is quite clear that responders, when they reject, deliberately choose to hurt proposer even if this implies a cost for themselves. What remains a bit unclear is whether proposers give substantial amounts to responders for fear of rejection or for ”pure” fairness considerations.

In this respect it is useful to compare results in the ultimatum game with the ones in the dictator game, as in Forsythe et al. (1994) (see also Kahneman et al., 1986). The dictator game has one player deciding over how to divide a pie; it can be seen as an ultimatum game stripped of rejection possibilities. Here we find that some players decide to take the whole ”pie”, but most do not. Average amounts left to the other players are lower, but still substantial (around 25%). The comparison indicates that most likely fairness and fear of rejection are present in proposers’ behavior in the ultimatum game. However, dictator game results have been questioned, in particular by experiments involving ”double-blind” procedures. Hoffman et al. (1994) find that when such procedures imply that not only opponents but even experimenters cannot attribute choices to subjects (an assistant collects choices in numbered envelopes and give them to another one, so that nobody can link choices to faces) then the percentage of dictators keeping the whole endowment grows to 64%. It still remains significant, however, that not only with double blind procedures some subjects still fail to keep all for themselves, but especially the fact that in absence of such procedures results change in direction of increasing equality. Of course, in principle, standard predictions should not be modified just because of observability of choices by experimenters. In other words, effects of observability by the experimenters are per se a significant deviations from standard predictions, while anonymity across subjects can instead be justified by the possibility among them to make agreements before the experiment takes place or that revenges or rewards could take place out of the lab, factors that deeply change the very nature of the game. Moreover, procedures in Hoffman et al., e.g. substantial differences in instructions across treatments, were criticized and stimulated new experiments. Among them, Bolton and Zwick (1992) found no evidence that being observed by the experimenter causes any effects in ultimatum and dictator games. These contradictions are a general aspect especially strong in a non-strategic interaction like a dictator game, whose results in general are to be taken cautiously. However, their overall indication that fairness considerations do enter players' mind and behavior should not be neglected.

Also the relevance of ultimatum game results has been questioned by several economists, and in particular by Binmore, Shaked and Sutton (1985), whose claim, in a nutshell, is that proposers offer equal or close to equal split as behaving as theory commands would make it very cheap for responders to behave irrationally and reject. While the dictator game appears to show that some role is also played by fairness consideration in proposers' behavior, the relevance of how cheap it is to reject is not to be overlooked.

As a matter of fact, however, results from experiments when high stakes were involved confirm the tendency found in earlier experiments. In particular, Hoffman et al. (1995) find that whether the whole pie is worth \$10 or \$100 does not affect results significantly; Cameron (1999) and Slonim and Roth (1998) provide high stakes in countries where they are equivalent to one month wages or more, and still find approximately the same results.

Another set of games widely studied goes under the denomination of ”Gift Exchange Game”, following Akerlof's (1982) definition of labor contracts as ”partial gift exchange”. In fact the game, first conducted by Fehr et al. (1993), is especially devoted to provide intuitions about the relationship between employers and employees, for the cases where the former cannot completely control the latter's behavior. In this game, the first mover (the ”employer”) offers an amount of money, *w* (”wage”) to the second mover (the “employee”). The second mover can reject, in which case both players get zero payoff, or accept. If she accepts, she has to take a choice over a costly ”effort”, *e*, so that her payoff will be $w - c(e)$, with $c(e)$ strictly increasing in e, while the first mover gets $ve - w$. Values are set in such a way that the benefit for an additional effort unit for the employer, v, is always greater than the marginal cost for the employee.

It is easy to see that under standard assumptions the employees exert the minimum effort no matter the level of w (which they accept in any case) and, anticipating this, employers offer the minimum wage. During experiments, however, a number of employees exhibit reciprocal behavior, so that they exert high effort when the wage is high. As a consequence, employers do better, on average, when they offer high wages. This result resembles real life situations when employees have, for the nature of their jobs, some degrees of discretion over their work. It is found in this case that reciprocal behavior increase the total pie to share, while in the ultimatum game it may destroy it.

While in the gift exchange game second movers choices are responsible for efficiency levels, in the ”Trust Game” introduced by Berg et al. (1995) these are determined by the first mover. In fact the first mover decides how much to transfer to the second, but the transfer is tripled in this passage. Then the second mover decides how much to transfer back. Here standard assumption would lead to no returns by the second movers in any case, and, consequently, no money transferred by the first. Perhaps not surprisingly, second movers return a relevant proportion of what they receive so that it is a good policy for first mover to “invest” a substantial amount in the first transfer.

All these experiments share a common feature: an important fraction of players deviate from own-payoff maximization in circumstances where, as second movers, their choice leads directly to a certain payoff distribution. Also in common across these games is the fact that such deviations go in the direction one would expect when assuming the natural tendency to reciprocate nice or bad behavior. An ultimatum responder rejecting a low offer or a “worker” exerting a high effort in response to a high wage take actions which are compatible with very intuitive reciprocity criteria. Also noteworthy is the difference, though: in the ultimatum game responders sacrifice their payoff to hurt ”mean” proposers, thereby reducing to zero the total pie initially available to players, while workers are willing to spend their effort in order to help generous employers, and their choice actually increase the total payoffs the experimenter is going to distribute.

However, the trust game indicates that a second mover can also transfer money when this act does not increase the payoff sum, but only affects its distribution.

Both positive and negative reciprocity seem to play an important role. An interesting experiment in Offerman (2002) presents the “hot response game”. First movers have to decide among a helpful and a hurtful choice. The former results in getting 8 units for themselves (to sum up to a previously achieved endowment) and 4 for the opponent; the latter in getting 11 and making her lose 4. The second mover, then, has three options: a “cool” reply which makes her achieve 10 additional units, nothing to the first mover; a “reward” choice, which gives her 9 and 4 to the first mover, and a ”punish” choice which also gives her 9, and ”-4” to the first mover. It is found that after a helpful choice 75% of second movers choose to reward, 25% a cool reply, nobody chooses to punish. The latter choice, instead, is taken by 83% of second movers after a hurtful choice, the rest picking a cool reply. This evidence reinforces the tendency not to take the own-payoff maximizing choice, in second mover behavior, as the cool reply is seldom selected. More interesting, however, is the comparison with another treatment where the first mover’s choice is selected by a random mechanism. In this case only 17% punish after a hurtful choice, while 50% reward after a helpful one. Offerman interprets this as evidence of a greater effect of negative, with respect to positive reciprocity: the difference it makes that the choice depended on first mover’s will or not is much more pronounced when we observe frequencies of hurtful choices than when we look at helpful ones. In other words, as the title says, ”hurting hurts more than helping helps” in the sense that the perception of bad intentions changes second movers’ behavior much more than the perception of good intentions. This is related by the author with the existence of a ”self-serving bias” according to which a player expects that an opponent must behave nicely to her, when given the opportunity.

The evidence presented in this section suggests the need for economists to go beyond self interest to find the rationale behind human behavior. Most data we have seen so far come from games explicitly designed to test predictions arising from the hypothesis of self interest. Let us see in the next section how people behave in settings which more closely resemble ”traditional” playground in economics.

3. Prisoner’s dilemma and public good games: the cooperation puzzle

The prisoner’s dilemma is possibly the most enlightening example of what game theory is about, and in fact it is usually the first which is presented to students in standard microeconomics course to spur interest in game theory. The interest comes for the simplicity in representing a number of real life situations, besides the need for the police to make accomplices confess their crimes. For instance, nations in potential conflict spend incredibly high amounts in arming even when their population have serious nutrition problems. The reason their governments typically offer is that refraining from doing so would mean being attacked and defeated by the enemy. On the other hand, a less frequently mentioned motivation for failing to agree on not buying or constructing weapons any further is the temptation each one would have on breaking the agreement, get weapons and attack. In game theoretical terms, this and other less dramatic situations feature equilibria characterized by defection by the players involved, which create allocations which are inferior, in Pareto sense, with respect to the ones available if players chose cooperation.

Andreoni and Miller (1993) ran a prisoner’s dilemma experiment lasting 10 rounds, a duration players knew ex-ante. The well known and previously mentioned backward induction reasoning leads to predict defection in all rounds by both players in every couple. However, their results are characterized by fairly high cooperation rates in round 1 (60% approximately) and a steady decay till very low levels in the last round.

It could be argued that players are learning the ”proper” way to play, i.e. to defect, during the game. We should notice, though, that behavior in all couples appears to show a frequent type of logic by a certain number: cooperate until your opponent defects. However, it also happens that a first defection in the couple is relatively more frequent the closer we get to the last round. These results, and indeed the design of the experiment itself, are closely related to the theoretical model exposed in Kreps et al. (1982). In this model it is shown that cooperation by a self-interested agent can be rational if his expectation to meet an altruistic player, in the sense of being willing to cooperate as long as the other does, is strictly positive. Clearly, as rounds pass by and the end gets closer, the expected returns from cooperation get lower and this explains the increasing tendency to defect. So, experimental data are consistent with the model. By the same token, we should expect that if the two players are not going to be matched together again cooperation rates should not be significant. However, in the so called ”stranger” treatment also proposed in Andreoni and Miller (1993), where players are rematched at every round with a mechanism ensuring no repetition among any two players, cooperation is still substantial, although lower than in the ”partner” treatment previously described. The authors conclude that the overall evidence shows that reputation building is taking place and therefore some players behave as if they were altruistic, as shown by higher cooperation in the partner treatment, but at the same time some agents are really altruistic, otherwise no cooperation at all should take place in the stranger treatment.

Public good games can be seen as an extension of prisoner’s dilemmas. They involve two or more players and typically several alternatives, in term of contribution level. Public goods are characterized, in their typical definition, by non-rivalry and non-excludability. These features make them be produced in an inefficiently low quantity, when production decisions are taken separately. In fact, the cost of a single agent’s production is born by himself, while the benefits are reaped by the whole society (or whatever group of people in consideration). This is a case of positive externalities which leads to under-production. In experiments on the production public goods, often called “voluntary contribution mechanisms”, a typical representation of an agent *i*’s payoff is:

$$\pi_i = e_i - g_i \ + m\textstyle\sum_j g_j$$

where e_i is his endowment (often assumed to be the same among all agents) and g_i his contribution to the public good. m represents the public good technology, and as a rule $m < 1 < mN$ holds, where N is the group size. The first part, $m < 1$, implies that the marginal return of contribution for an individual, (−1+m), is always negative, so that his self interest commands zero cooperation in the basic one-shot game, and in the finitely repeated game. The second inequality, $mN > 1$ implies that the social returns from contribution are positive. So, just as in the prisoner’s dilemma, we have a Nash equilibrium solution, zero contribution by all players, which is Pareto inefficient with respect to the social optimum, reached if everybody contributes the whole endowment.

Just like the main features, also the pattern of the deviations from predictions is quite similar to the prisoner's dilemma: contribution levels are typically significant at the beginning, for instance at 40% in Isaac and Walker (1988) and the decay as rounds go by. Also here, we have evidence of what could be thought as learning the ”right” strategy. However, studies like Andreoni (1988) and Croson (1996) show examples of ”restart effect”. What happens in their data is that if the same players who played a public good experiments in which indeed a decay was observed are rematched and start the game anew, then their contribution levels resemble the ones of the first round in the previous play. Of course, this is not compatible with the idea that they were previously contributing low amounts before just because they had learned that contributing more is a bad idea.

Other interesting findings include the positive effects on contribution levels of increasing group size when the m factor is kept constant, which suggests that ”efficiency gains” matter (see, e.g., Brandts and Schram, 2001), and of allowing pre-play communication which in principle should be irrelevant “cheap talk” (Isaac and Walker, 1988).

Of particular relevance, in terms of tracking down motivations, are recent studies by Keser and VanWinden (2000) and Fischbacher et al. (2001). The former replicate evidence of steady decay in contribution in a ”stranger” treatment and observe closely behavior in a ”partner” treatment to find that agents tend to adjust to the group average, which gives evidence of ”conditional cooperation behavior. Fischbacher et al. test this possibility more directly, by allowing one member per group to revise his contribution decision; they find that 50% of players adjust to group averages as the main indicator of non-selfish behavior. Finally, the most dramatic change with respect to free-riding outcomes is found by Fehr and Gaechter (2000). They introduce a punishment mechanism which allow players to reduce any opponent's payoff, but at a cost of reducing their own. Predictions stemming from standard assumptions do not take into account this possibility in a finitely repeated game, as a selfish player would never punish. However the evidence shows that free riders are heavily punished by high contributors and, especially, that the former (instead of the latter as in standard public good games experiments) quickly adapt their behavior, so that average contribution levels go up and stay very close to full contribution till the last round of the experiment.

The overall message of this section is that intuitive deviations from strategies deduced from standard homo economicus reasoning affect substantially the results we get in games which, due to their socio-economic relevance, have traditionally been in the spotlight of economic analysis. This fact, in conjunction what we have already seen in the previous section, motivates the quest for new theories of human behavior better fit to accommodate this puzzling evidence.

4. Explaining experimental evidence: learning and social preference models

4.1 Learning models

Some interesting models based on assumptions of bounded rationality were proposed to explain ultimatum game results. In particular, Roth and Erev (1995) argue that a decisive aspect behind results consists in the fact that rejecting low offers has a mild cost for responder, while proposing an excessively – from the responder's point of view - unfair allocation which causes rejection results in a great loss, with respect to what would be achieved offering the minimum amount that a responder would take. Then a learning model based on adapting choices to payoffs in previous rounds show that self-interest is indeed compatible with ultimatum game results once we relax rationality assumptions. In the same direction goes the logit equilibrium proposed by McKelvey and Palfrey (1995), where the following logit rule is incorporated in the analysis:

$$p_i = exp(\pi^e{}_i/\mu)/\sum_j exp(\pi^e{}_j/\mu) \qquad (1)$$

where p_i is the probability that choice i is taken among the available alternatives and the error parameter, μ, determines the sensitivity of choice probabilities to payoff differences. McKelvey and Palfrey apply the analysis to the centipede game, where indeed the potential length of the strategic interaction - which is reduced to an immediate

”take” choice in a not so immediately recognizable subgame perfect Nash equilibrium - gives a solid ground to the applications of bounded rationality models. However, we would argue again that limitations of rationality should not be taken as the main explanation for simple decision nodes as the responder’s in the ultimatum game. In other words, we think that learning aspects may rather be a complement, in some complex games, to an analysis incorporating non-selfish preferences[1].

4.2 Altruism

One of the most intuitive ways of incorporating other-regarding preferences in the economic analysis is to assume that other players’ payoff enter positively some player’s utility function, in other word that these players are altruistic (e.g. Becker, 1974 and various references in Zamagni, 1995). This could explain why cooperation is observed in prisoner dilemma and public good games, for instance, but especially the fact that a fraction of players give something to their opponents in dictator games. Formally, denoting payoffs with π, a player *i* is altruistic with respect to the members of his *N* size group if:

$$u_i = u_i(\pi_1, ..., \pi_i, ..., \pi_N), \qquad (2)$$

with $\partial u_i/\partial \pi_j > 0$ for all $j = 1, ..., N$

One objection to this approach comes from Andreoni’s (1989) “warm glow” theory, in which it is argue that players value the sheer act of giving, so that, for instance, public funding does not ”crowd out” private donations.

More radical objections can be issued, however. In particular, we have observed that most non-selfish players behave as conditional cooperators, while according to (2)

[1] An illuminating example is found in Goeree and Holt (2000) as an explanation of the effects of fixed assignments by the experimenter in a two-stage alternating offer bargaining game.

opponent’s defective behavior should not, in principle, prevent totally future cooperative acts. That is to say that those players’ behavior does not appear to be consistent with (2) throughout the rounds of the experiments, while of course an alternative utility function should be stable in its ability to explain choices. Moreover, hurtful choices cannot be justified by the functional form in (2), but indeed they do occur in a variety of settings, such as the ultimatum game and the public good game with punishment.

4.3 Intention-based theories

Levine (1998) propose a model in which other players’ intentions enters the utility function as follows:

$$u_i = \pi_i + \sum_{j \neq i} \pi_j (a_i + \lambda a_j)/(1 + \lambda) \qquad (3)$$

with $0 \leq \lambda \leq 1$ and $-1 < ai < 1$ for all players.

The a_i parameter denotes a ”general” tendency of player i towards other players: if positive, i tends to be altruistic, if negative he is “spiteful”. The λ parameter is what accounts for modifications in behavior depending on opponents’: previous experience is used to estimate the value a of any given opponent. If $\lambda > 0$, then a player tends to be nicer towards the altruistic players, and meaner towards the spiteful ones.

Rabin (1993) seminal contribution, aimed at entering psychology into game theoretical analysis, is based on a “kindness function” by which a player evaluates how kind his opponent is. The model incorporates believes entertained by a player not only on how his opponent behaves, but also on what a player believes his opponent believes about his own choice; a ”fairness equilibrium” is defined as a set of strategies which are reciprocal best response and a set of rational expectations consistent with the actions involved in the equilibrium. Based on Rabin’s model, Dufwenberg and Kirchsteiger (2004) introduce the ”Sequential Reciprocity Equilibrium” in which players keep track of beliefs about intentions after observing actions in each round of a repeated game, while Rabin’s approach is focused on normal form games.

The common feature of these models is that they involve a large number of parameters and typically involve a multiplicity of equilibria. That is to say, their possibility of explaining a larger part of human behavior, such as conditional cooperation, comes at a cost of increasing substantially the complexity of the analysis.

4.4 Distributional preferences

Bolton (1991) proposes relative income as an additional factor into players' utility. In particular, in two-player games his formulation is:

$$u_i = u_i(\pi_i, \pi_i/\pi_j), \qquad (4)$$

with $\partial u_i/\partial(\pi i/\pi j) > 0$ if $\pi_i < \pi_j$, and $\partial u_i/\partial(\pi i/\pi j) = 0$ otherwise.

The intuition is quite simple: you dislike to be worse off than another player and so the more you reduce unfavorable inequality the better. If the other player is not better off, then you do not care about her payoff. This model is compatible with ultimatum rejections and with punishment following free riding in a public good game. However, it cannot explain ”nice” behavior towards opponents such as giving in dictator game, returning favors in the gift exchange and in the trust games and conditional cooperation in dilemma games.

This criticism appears to be shared by the author, as in Bolton and Ockenfels (2000) we find related but more general formulation:

$$u_i = u_i(\pi_i, \sigma_i) \qquad (5)$$

where

$$\sigma_i = \pi i/\textstyle\sum_j \pi_j \qquad if \qquad \textstyle\sum_j \pi_j \neq 0$$

and

$\sigma_i = 1/N \qquad if \qquad \sum_j \pi_j = 0$

For a given level of own payoff π_i, the maximum utility is reached when $\sigma_i = 1/N$. That is, a player is happier if his proportion of wealth is equitable. On the other hand, as we would expect, for a given proportion σi his utility is increasing in his own wealth. This model incorporates aversion also with respect to favorable inequality and is compatible with the evidence that Bolton (1991) failed to explain.

A similar logic is at the basis of the model proposed by Fehr and Schmidt (1999), in which we have:

$$u_i(\pi_i, \pi_{-i}) = \pi_i - [1/(N-1)] \sum_{j \neq i} \alpha \max\{\pi_j - \pi_i, 0\} - [1/(N-1)] \sum_{j \neq i} \beta \max\{\pi_i - \pi_j, 0\} \quad (6)$$

α measures aversion to unfavorable inequality, or ”envy”, β the aversion to favorable inequality, or ”guilt”. Both this and Bolton and Ockenfels models are compatible with behavior in a variety of games. In particular, ultimatum rejections after an unfair offer can be caused by aversion to unfavorable inequality, while second mover’s behavior in the gift exchange game and in the trust game reduces the favorable inequality arising from a generous first move. Giving in dictator game is also compatible with these models according to the same logic; the linear formulation in (6) would actually imply that players should choose between sharing equally the ”pie” or not giving anything, while a few choices are somewhere in the middle. Linearity is chosen to keep the model as simple as possible while at the same time more determinate than in (5). Conditional cooperation in prisoner’s dilemma and in public good game are in line with these models, too, as after a round where a player defects, the cooperator is worse off and therefore motivated to switch to defection. Punishment behavior is also compatible, as the damage for the punished is bigger than the expense for the punisher. However, the fact that only free riders are punished is more in line with (6): if instead a player’s concern were only on getting closer to the average, then it would be indifferent with respect to whom to punish. Evidence in Falk et al. (2000) remarks this aspect: a cooperator never hurts another cooperator even if doing so would improve his relative payoff, making it closer to the average (as some agents defected).

Moreover, both couples of authors underline - also in their titles - that not only equity and reciprocity (Bolton and Ockenfels) and fairness and cooperation (Fehr and Schmidt) can be explained, but also competitive behavior leading to unequal outcomes. In particular, variations of the ultimatum games include competition among proposers or among responders. In those games it actually happens that the way to play predicted by these models converges to the one envisaged by "standard models" predictions: the player alone on one side of the market enjoys outcomes in which he gets the maximum payoff while the others are left with almost nothing. In those games the advantaged player knows that all but one player on the long side of the market will be left with nothing in any case. Therefore, picking the highest offer does not affect inequality levels substantially, while improving his own payoff. On the other hand, on the long side of the market the possible presence of a single selfish player drives behavior towards bidding up offers or minimum acceptance levels. As a matter of fact, preferences are not assumed to be homogeneous in neither of the models; in particular, Fehr and Schmidt's calibrations find that the presence of a percentage (around 30%) of purely selfish players is a constant across most games. Their behavior, though, is affected by the presence of non-selfish players, so that for instance a first mover in a gift exchange game may behave generously even if he is only concerned about his own payoff.

4.5 Social welfare orientation

Charness and Rabin recently proposed a new model where players are assumed to care about social welfare as defined by the following function:

$$W_i(\pi_1, ..., \pi_i, ...\pi_N) = \delta \, min\{\pi_1, ..., \pi_i, ...\pi_N\} + (1 - \delta)\textstyle\sum_j \pi_j \tag{7}$$

As we can see this function combines efficiency in term of payoff sum with a Rawlsian maximin criterion. Then an agent i who assigns weight λ_i on this social component maximizes:

$$u_i = (1 - \lambda_i)\pi_i + \lambda_i W_i(\pi_1, \ldots, \pi_i, \ldots\pi_N) \tag{8}$$

This model captures evidence found by the authors themselves and others (see below) of efficiency oriented behavior. For instance, they find that most players when choosing as dictators prefer (own payoff, opponent's) the allocation (400,700) to (400,400). However, this model fails to capture punishment and ultimatum rejections. For this reason, they propose an extension in which, more or less in the same logic as in Rabin (1993) or in Levine (1988) players evaluate each other's social welfare orientation, i.e. the parameter λ. A player who is found to put a "too low" weight on social welfare stimulates negative reciprocity and therefore may be punished. They acknowledge the complexity involved in this variation and argue that among simple models, theirs is to be preferred to the ones involving inequality aversion in terms of ability to explain a wider array of economically relevant interaction.

With this argument we are introduced to an exciting area of investigation which we deal with in the next section.

5. Which social preferences?

So far we have seen alternative models which are able to explain an important part of that reality that models based on assuming pure selfishness fail to capture. Pure altruism is probably the criterion that proved to be the weakest among the ones proposed above. In the next paragraphs we present the evidence about the two main comparisons raised in this respect.

5.1 Equality versus reciprocity

Models based on reciprocity have a very intuitive appeal: most people would agree that it is a ”good thing” to ”help nice people and hurt the mean”. However, incorporating reciprocity comes at a cost of high complexity. This is not the case with inequality aversion: the formulation in Fehr and Schmidt, in particular, only involve two parameters. The question is whether simplicity comes at the cost of failing to explain many relevant results in social and economic interaction.

A couple of variations of the ultimatum game point in this direction. Blount (1995) replaces first mover's choice with a random draw. Knowing this, second movers are more willing to accept unfair outcomes. This is evidence that responders take proposers' intentions into account, when they reject unfair offers. In Falk et al. (2000) and in Brandts and Solà (2001) a (8,2) (proposer's payoff first) proposal is more frequently rejected when the alternative available to proposers is (5,5) than if it is (2,8) or (10,0); this shows that how equitable alternatives were also matters. With a rationale similar to Blount's, Falk et al. (2000b) show evidence that in a sequential games results change depending on whether first mover's choices were intentional or randomly determined, but distributional concerns are also relevant, in deciding whether and how much to punish or to reward first movers.

Other studies, however, found results showing little or no importance of intentions. For instance, Cox (forthcoming) find that the replacement of an intentional move with a random mechanism does not weaken the tendency found in trust games to give back a substantial part of the investment made by the first mover. Bolton et al. (2000) find that in a sequential game where second movers have to pick among five alternatives payoff allocations, the distribution of their choices does not vary depending on whether this decisional node was reached after a good, a bad or a neutral choice by the first mover, in terms of an easy comparison with the alternative path he could have taken. We argue that in all those settings intentions were given a very good chance to affect outcomes and it is not always the case that they do, unlike what could be expected taking into account that we are talking about choices by players which do not comply with predictions based on self interest. There is clearly scope for further research, in particular about which factors drive apparently contradictory evidence. By now, we cautiously argue that the present evidence does not indicate compelling reasons to incur in the cost of complexity required by incorporating intentions into economic models, as simpler inequality aversion models can explain almost as much with lower complexity and much easier ways to find unique predictions[2] .

5.2 Equality versus efficiency

The previously mentioned model by Charness and Rabin predicts that subjects are motivated by a function which includes the total payoff to be distributed. In the same paper they present evidence that in dictator games many players are willing to take choices in this direction, even if they imply increasing inequality. Several authors performed tests giving different results. Bolle and Kritikos, for instance, confirm that the payoff sum matters for several players, while Güth et al. (2003) and Morgenstern (2003) find that efficiency plays no role.

Some tentative conclusions may be derived: it appears that efficiency concerns vanish when strategic interactions are happening. In the latter case, agents resist with particular strength to the idea of allowing opponents to go with higher payoff, and also show willingness to transfer money even when this does not increase the total pie (as in the trust game). This led Fehr and Schmidt (2002) to underline that ”the Dictator Game is different from many economically important games and real life situations” and ”where both players have some power to affect the outcome, the surplus maximization motive is less important”. Moreover, as previously stressed, dictator game results are in general the least stable with respect to small variations in procedures, as the evidence reported here also confirms, while results in strategic interactive settings are often confirmed even varying money stakes or cultural contexts.

6. Conclusions

We explored the evidence contradicting predictions arising from standard “homo economicus” assumptions of rationality, selfishness and common knowledge of these two

[2] Charness and Rabin themselves use (8) but not the reciprocity-based modification to evaluate the data they collected, which shows how difficult it can be to derive clear-cut predictions from models incorporating intentions.

features. These data are gathered not only in games created in order to test the validity of those assumptions, but also in others which, due to their relevance, have always been a main focus of economic analysis, such as the prisoner's dilemma and public good games.

These findings stimulated the research towards ways to incorporate aspects of "social preferences" which appear consistent with data into the economic analysis. Therefore models were created which include altruism, reciprocity (the so-called "intention-based" models), inequality aversion and "social welfare". The authors of these models stress their predictive power on a wide range of phenomena observed in the experimental lab, although it is typically the case that each model fails at some relevant game.

New theories, in turn, stimulated new experimental research to further test their predictive power. The findings are by no means definitive, although there is possibly wide agreement on the weakness of ”pure altruism” with respect to the other alternatives to self interest. However, we claim that by now most evidence is in concordance with inequality aversion models; incorporating reciprocity can enhance predictive power but not as much - in our view - as to justify the great cost in terms of complexity and, often, equilibrium multiplicity. Efficiency criteria, on the other hand, seem to have a scope limited to particular settings where strategic interaction is absent.

A final note is about a further direction of research, which consists in applying those models to a wide range of games. An obvious, already explored but still worth investigating setting is the relationship principal-agent, in particular in terms of whether contracts should be made as complete as possible or not. Indeed, fairness minded agents may exert high efforts even if - or perhaps especially if - contracts are left incomplete and control mechanisms are relaxed. Other results (Huck et al., 2001) show that inequality aversion drives towards puzzling evidence in Stackelberg competition, namely upward sloping response functions by followers. This indicates that areas of applications of these models really abound. One merit of this research, indeed, is to have shown that the scope of economic analysis is not limited by these aspects of human behavior, but actually enriched.

References

Akerlof, G. (1982): ”Labor Contracts as Partial Gift Exchange”, *Quarterly Journal of Economics*, 97, 543-569

Andreoni, J. (1988): ”Why free ride? Strategies and Learning in Public Good Experiments”, *Journal of Public Economics*, 37, 291-304

Andreoni, J. (1989): ”Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence”, *Journal of Political Economy* 97, 1447-1458

Andreoni, J. and Miller, J.H. (1993): ”Rational Cooperation in the Finitely repeated Prisoner’s Dilemma: Experimental Evidence”, *Economic Journal*, 103, 570-585

Becker, G. (1974): ”A Theory of Social Interactions”, *Journal of Political Economy*, 82, 1063-1093

Berg, J., Dickhaut, J., and McCabe, K. (1995): ”Trust, Reciprocity and Social History”, *Games and Economic Behavior*, 10, 122-142

Binmore, K., Shaked, A. and Sutton, J. (1985): ”Testing Non-Cooperative Bargaining Theory”, *American Economic Review*, 78, 837-839

Blanchard, O. and Fischer, S. (1989): *Lectures on Macroeconomics*, Cambridge MA, MIT Press

Bolton, G.E. and Zwick, R. (1995): ”Anonymity versus punishment in ultimatum bargaining”, *Games and Economic Behavior*, 48, 287-292

Bolle, F., and Kritikos, A. (2001): ”Distributional Concerns: Equity or Efficiency Oriented?” *Economics Letters* 73, 333-338

Bolton, G.E., Brandts, J. and Ockenfels, A. (2000): ”Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game”, *Experimental Economics* 3, 207-221

Bolton, G.E., and Ockenfels, A. (2000): ”A Theory of Equity, Reciprocity and Competition”, *American Economic Review* 100, 166-193

Brandts, J. and Solà, C. (2001): ”Reference Points and Negative Reciprocity in Simple Sequential Games”, *Games and Economic Behavior*, 36, 138-157

Brandts, J. and Schram, A. (2001): ”Cooperation or Noise in Public Goods Experiments: Applying the Contribution Function Approach”, *Journal of Public Economics*, 79, 2, 399-427

Cameron, L. (1999): ”Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia”, *Economic Inquiry*, 37:1, 47-59

Charness, G. and Rabin, M. (2002): ”Understanding Social Preferences with Simple Tests”, *Quarterly Journal of Economics*, 117, 817-869

Croson, R. (1996): ”Partners and Strangers Revisited”, *Economics Letters* 53, 25-32

Dufwenberg, M. and Kirchsteiger, G. (2004): ”A Theory of Sequential Reciprocity”, *Games and Economic Behavior*, 47, 2, 268-298

Falk, A., Fehr, E. and Fischbacher, U. (2000): ”Informal Sanctions”, Working paper N.59, University of Zurich

Falk, A., Fehr, E. and Fischbacher, U. (2000b): ”Testing Theories of Fairness - Intentions Matter”, Working Paper No. 63, University of Zurich

Fehr, E. and Gaechter, S. (2000): ”Cooperation and Punishment in Public Goods Experiments”, *American Economic Review*, 90, 4, 980-994

Fehr, E. and Schmidt, K.M. (1999): ”A Theory of Fairness, Competition and Cooperation”, *Quarterly Journal of Economics* 114, 817-868

Fehr, E. and Schmidt, K.M. (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

Güth, W., Kliemt, H. and Ockenfels, A. (2003): ”Fairness versus Efficiency: An Experimental Study of (Mutual) Gift Giving”, *Journal of Economic Behavior and Organization*, 50, 4, 465-475

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993): ”Does Fairness Prevent Market Clearing? An Experimental Investigation”, *Quarterly Journal of Economics*, 108, 437-460

Forsythe, R., Horowitz, J.L., Savin, N.E., and Sefton, M. (1994): ”Fairness in Simple Bargaining Experiments”, *Games and Economic Behavior*, 6, 347-369

Goeree, J. and Holt, C. (2000): ”Asymmetric Inequality Aversion and Noisy Behavior in Alternating-Offer Games”, *European Economic Review*, 44, 1079- 1089

Güth, W., Schmittberger, R., and Schwarze, B. (1982): ”An Experimental Analysis of Ultimatum Bargaining”, *Journal of Economic Behavior and Organization*, 3, 367-388

Hoffman, E., McCabe, K. and Smith, V. (1996): ”On Expectations and Monetary Stakes in Ultimatum Games”, *International Journal of Game Theory*, 25, 289-301

Huck, S., Müller,W. and Normann, H.T. (2001): ”Stackelberg Beats Cournot: On Collusion and Efficiency in Experimental Markets”, *Economic Journal*, 111, 749-766

Kahneman, D., Knetsch, J.L., and Thaler, R. (1986): ”Fairness and the Assumptions of Economics”, *Journal of Business* 59, S285-S300

Keser, C. and van Winden, F. (2000): ”Conditional Cooperation and Voluntary Contributions to Public Goods”, *Scandinavian Journal of Economics* 102(1), 23-39

Kreps, D.M., Milgrom, P, Roberts, J. and Wilson, R. (1982): ”Rational Cooperation in the Finitely repeated Prosoner’s Dilemma: Experimental Evidence”, *Journal of Economic Theory*, 27, 245-252

Levine, D. (1998): ”Modeling Altruism and Spitefulness in Experiments”, *Review of Economic Dynamics* 1, 593-622

Morgenstern, A. (2003): ”Responsibility, Delegation, and Incentives”, mimeo, University of Bonn

Nash, J.F. (1950): ”Equilibrium Points in N-person games”, *Proceedings of the National Academy of Sciences* 36, 48-49

Offerman, T., (2002): ’Hurting Hurts More Than Helping Helps’, *European Economic Review* 46, 1423-1437

Rabin, M. (1993): ”Incorporating Fairness into Game Theory and Economics”, *American Economic Review*, 83(5), 1281-1302

Roth, A.E. (1995): ”Bargaining Experiments”, in Kagel, J, and Roth, A.E. (eds.): *Handbook of Experimental Economics*, Princeton, N.J., Princeton University Press

Selten, R. (1975): ”Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games”, *International Journal of Game Theory*, 4, 25-55

Slonim, R. and Roth, A. (1998): ”Financial Incentives and Learning in Ultimatum and Market Games: an Experiment in the Slovak Republic”, *Econometrica*, 66, 569-596

Zamagni, S. (ed.) (1995): The Economics of Altruism, Edward Elgar Pub., Brookfield, VT