LUISS Guido Carli
LIBERA UNIVERSITÀ INTERNAZIONALE DEGLI STUDI SOCIALI

Department of Economics and Finance

Chair of Games and Strategies

COOPERATIVE BEHAVIOR IN PRISONER'S DILEMMA GAME

Different models to capture the role of fairness

Supervisor:

Marco Dall'Aglio

Candidate:

Costanza La Serra

ID: 166891

A.Y. 2013/2014

Index:

1. INTRODUCTION

Even the most primitive human societies seem like complex machines designed for growth and survival. There are different opinions about the behavior of individuals in a social environment. Functionalism, an ancient but still influential school in anthropology and sociology bases its beliefs on the assumption that behaviors and institutions are built only in order to promote the healthy functioning of social groups as every single's mental state depends on the role he plays in the system of which he represents a part. On the other hand, the conviction that people are selfish brings different sociologists to believe that human actions have to be interpretedsolely in terms of personal benefits: any group benefits are casual, an accidental side effect of selfish individual decisions. Economists in particular through game theory have explained this last school of thought for many times.

From what we can see in everyday life, seems that people are often willing to cooperate with each other for different reasons and therefore a lot of further research has been made on this topic. The principle of "fairness" has been introduced in game theory and economics for the first time by Rabin in 1993 and after that many experiments and other models for capturing the role of fairness in human choices have been developed. From both these new theories and empirical evidences seems that there are situations in which the standard mathematical self-interest model previously created to understand people's decisions when dealing with some specific gaming environment, is not any more appropriate.

Individuals are sometimes moved by "goodness" or influenced by others' actions. In many situations they are not only interested in achieving their material payoffbut rather cares about "social" goals or may be willing to help people that have been kind with them, while punishing subjects that have behaved in a too selfish way. All these emotions we will see that have economic implications.

Different mathematical frameworks will be described for this purpose and applied to some simple games in order to understand and eventually predict human choices.

It has to be said that the natural structure of some kind of games is more suitable for inducing cooperative attitudes by its players, while in others there are factors such as high competition that makes very difficult to reach common fairness.

The gaming environment in which we are particularly going to concentrate is the one of social dilemmas, in the specific we will focus our attention on the simple prisoner dilemma game and some applications will be done.

The prisoner's dilemma is the most famous example of game with a unique Pareto-inefficient Nash equilibrium. The main characteristic of this game is that even if the substantial gains achievable through cooperation, non-cooperation (defection) represent the dominant strategy for all the players. The theoretical result is that each agent will avoid cooperation even if through joint defection the final payoffs are lower than the ones that could have been achieved through mutual cooperation.

In general we define social dilemmas as those specific games in which it arises a tension between personal interest and collective interest. These are very interesting and challenging situations because acting in one's immediate self-interest is tempting to everyone involved, even though everybody benefits from acting in the longer-term collective interest and in real life we continuously have to deal with such kind of situations. For this reason, generally speaking, more knowledge about social dilemmas should be very useful in order to understand not only the theoretical puzzles of why people cooperate (or not) but also the ways in which cooperative behavior in organizations or broadly in the society could be sustained or promoted.

1.1 Social Value Orientation

Since social dilemmas are so important, they have been studied and analyzed in different disciplines such as psychology, sociology, political scientists and economists. To better comprehend the environment in which the models we are interested in are developed, we have to say that people's personal interests can be very different in the society and when they are put in front of a choice they react based on the *type*of person they are.

Social Value Orientation (SVO) proposes that when people share something valuable between themselves and others, their personal SVO lets them weight payoffs differently to self and other people and accordingly redistribute the resources.

Three kind of SVO have been identified in social games, dividing the population into three different kind of subjects and this are the cooperative, competitive or individualistic one.

Cooperators are those people who aim at maximizing the joint outcome to self and the other players, competitors are those who act such as to maximize the difference in outcome between self and the others, in order to gain a relative advantage and finally individuals show their only interest in maximizing their individual payoff with little or no regard for the outcome of the rest of the players.

SVO was introduced to explain different behaviors in cooperation in social dilemmas, which are situations in which people may either decide to act in their own interest, called *defection*, or to act in the collective interest, called *cooperation*. Thus, previous studies referred that people with a cooperative SVO more frequently were willing to collaborate in social dilemmas with respect to individualists and competitive individuals.

Several methods has been developed in order to assess SVOs, but the most common is the "decomposed game". This method uses the so called triple-dominance measure of social values (TDMSV) to distinguish among the three categories of individuals. In the TDMSV players have to choose nine choices among three alternative distributions of values to themselves and another unknown participant. The three alternatives corresponded to an individualistic, a cooperative and a competitive one. The proportion of the different categories of people can vary of course from study to study but overall it seems to be rather stable.

It is clear that this method clearly asses competitors and individualistic motives, but it fails in determining the actual cooperative motive that could be either considered the intention of achieving equal outcomes or the one of maximizing joint outcome.

Van Lange (1999) hypothesized that prosocials' incentive was bot joint payoff and equality, but D. Eek and T. Garling in their found out that equal outcome is probably the only motive. They argue that the prevalence of equal payoffs among cooperative people reflects an equality motive based on fairness considerations. This means that prosocials act aiming at achieve equal outcomes, not minimizing the differences between their own and the others' reward. Thus, if equality can't be put in place, cooperatives would rather choose another alternative instead of minimizing differences in outcome.

All that said it seems essential to have mathematical framework aimed at capturing the relevance of altruism in social dilemmas.

2. THE MODELS

In this chapter we are going to focus on the description of the most relevant models for fairness that started to be developed in the early 90s after reasoning on the evidence that strict self-interest mathematical framework were not representative of people behavior and the results they provided were therefore unrealistic. The new model proposals are sometimes very different from each in the structure, as they capture different "emotions" and fairness in interpreted in different ways. It represents inequity aversion as well as altruism or again could be intended as the action of responding with kindness to goodness while badly to evilness.

2.1 RABIN (1993)

Matthew Rabin, in 1993 was the first that developed a game-theoretic structure to incorporate fairness into different economic models.

The basic assumption in this model is that people act in such a way that is "good" with those persons that have been gentle to them, while hurting those who have not been kind. Players are basically involved in what Geanakopolos, Pearce and Stacchetti (GPS,1989) named a "psychological game", where payoffs depends on both their own actions and on expectations about others' actions.

The outcomes reflecting such behaviors are called *fairness equilibria*and they can be of two types: the mutual-max outcome is the results of situations in which both the players act to maximize the other's material payoff and the mutual-min outcome where each agent aims at minimizing the opponent's payoff. Fairness equilibria do not constitute neither a subset nor a superset of Nash equilibria, they can just provide other solutions and eliminate others. Rabin shows that each mutual min and max Nash equilibrium is a fairness equilibrium and if payoffs are relatively small the fairness outcomes are approximately the set of mutual-max and mutual-min payoffs; if instead players have large payoffs the fairness equilibria seem to be the set of Nash equilibria.

The model starts adopting the GPS framework of "psychological game" and therefore explicitly incorporates beliefs that make the analysis more complicated but it is just necessary to better capture the aspects of fairness.

For two player normal form game the sets of pure strategy are A_1 , A_2 from which the mixed strategy sets S_1 and $S_2$are derived. The material payoffs are $\pi_i : S_1 \times S_2 \rightarrow R$.

The following notation are used: $a_1 \in S_1$ and $a_2 \in S_2$ are the strategies chosen by the two participants; $b_1 \in S_1$ and $b_2 \in S_2$represent respectively player 2 beliefs about which strategy is chosen by player 1 and player 1 beliefs about his opponent's action; $c_1 \in S_1$ and $c_2 \in S_2$finally are the expectations of each single player about what they think the other is thinking about their action.

If player i believes that the opponent j is going to choose b_j as his strategy, and consequently decides to play $a_i$determining a couple of payoff from the set :

$$\prod(b_j) \equiv \{(\pi_i(a, b_j), \pi_j(b_j, a)) | a \in S_i\}$$

We can consider an equilibrate payoff for player j that is an average of his highest and lowest payoff among points that are Pareto-efficient in $\prod(b_j)$, i.e. among the points that produce an outcome that makes every player at least as well off and at least one player strictly better off. In other words, a Pareto Optimal payoffs for $j$on $\prod(b_j)$ cannot be improved upon without hurting at least the other player. So given $\pi_j^h(b_j)$ and $\pi_j^l(b_j)$ as the extreme payoffs, we derive the equilibrate outcome that exactly corresponds to the payoff that player j would receive if his opponent decides to equally split the difference with him among all the Pareto-efficient outcomes and the Pareto frontier is limit.

$$\pi^e(b_j) = \frac{1}{2}\left(\pi_j^h(b_j) + \pi_j^l(b_j)\right)$$

Considering then the minimum possible payoff that player j can achieve in all the set $\prod(b_j)$ as $\pi_j^{min}(b_j)$ we derive Player i goodness to his opponent j in choosing the action a_i as the following equation:

$$f_i(a,b_j) = \frac{\pi_j(b_j, a) - \pi_j^e(b_j)}{\pi_j^h(b_j) - \pi_j^{min}(b_j)}$$

If we have $\pi_j^h(a, b_j) = \pi_j^{min}(a, b_j)$then $f_i = 0$and it happens only when player *i* is willing to give his opponent *j*his equitable outcome. When we have $\pi_j^h = \pi_j^{min}$ all the responses by player *i* to action $b_j$will give him the same payoff and therefore there is no incentive for *i* to play kindly.

In case $\pi_j^h > \pi_j^{min}$ we can have both positive or negative values of f_i. If $f_i < 0$we have a situtation in which player *i* is providing his opponent with a less than equitable outcome and this could be due to two reasons: either player *i* is taking more than the amount given to him among the Pareto frontier, or he is taking a value that is inefficient in this set.

On the oppposite case if $f_i > 0$ the player is giving his opponent more than his fair payoff, but of course we have to consider that the Pareto frontier is not represented by a single point and therefore $\pi_j^h \neq \pi_j^e$.

Since we already stated that in this game expectations are considered, we can define$\tilde{f}_j(c_i,b_i)$ as player *i*'s belief about the goodness of player *j* in his regards:

$$\tilde{f}_j(b_j, c_i) \equiv \frac{\pi_i(c_i, b_j) - \pi_i^{\mathrm{e}}(c_j)}{\pi_i^{\mathrm{h}}(c_i) - \pi_i^{\min}(c_i)}.$$

Same considerations have to be done in this case as regard the values of π_i^h, π_i^{min} and π_i^e that can again determine $\tilde{f}_{\mathrm{j}} = 0$ when $\pi_i^h = \pi_i^{min}$, meaning that every kind of action played by *j* with respect to the movement made by *i* doesn't influence *i*'s final payoff and therefore there's no expectation about the level of kindness to be received.

$\tilde{f}_{\mathrm{j}} < 0$ expresses beliefswhen*i* is expecting the opponent to act unfairly with him$\tilde{f}_{\mathrm{j}} > 0$ otherwise.

Supposing that *i* thinks that *j* will play b_j, then player *i* will choose action a_i to maximize his utility:

$$U_i(a_i,b_i,c_i) = \pi_i(a_i,b_i) + \tilde{f}_i(b_i,c_i)\,[1 + f_i(a_i,b_i)]$$

The utility *i* will obtain will be given by the material payoff derived from the actual action he plays, that is represented by the first term of the equation, plus his expected gain from the kindness the opponent will have for him but also with respect to the level of goodness that himself is going to play.

If player *i* expects that his opponent is playing unkindly, meaning$\tilde{f}_j(\cdot) < 0$, then he will decide to play badly in turn, choosing an action such that $f_i(\cdot)$ will be negative or very low. On the contrary if player *j*is acting kindly to player *i*, $\tilde{f}_j(\cdot)$ will be a positive value and in turn player *i* will respond with a positive $f_i(\cdot)$. Of course these utility functions must be related to the preferences for material payoff and in some cases pecuniary gains could exclude fairness considerations.

The model incorporates the stylized fact that the larger the monetary payoffs the less the players' attitude to care about fairness, the game is in this sense susceptible to the scale of pecuniary payoffs.

Rabin reaches the solution imposing the additional condition that beliefs match in all the cases the actual behavior of the agents.

A pair of strategies (a_1,a_2) is a fairness equilibrium if for $i = 1,2$ and $j \neq i$,

(I) $a_i \in \arg\max_{a \in Si} U_i(a,b_j,c_i)$

(II) with $a_i = b_i = c_i$

This two conditions are the same that need to be satisfied when finding Nash Equilibrium.

2.2 FEHR AND SCHMIDT

Ernst Fehr and Klaus Schmidt in 1999, discussed a new framework that can be labeled the *inequity-aversion model*and that represents one of the most relevant contribution to fairness studies.

Their model, as the Rabin's model, is based on the notion of an equitable outcome (the fairness outcome), but differently the two researcher understood that the focus Rabin made

on the role played by intentions was too linked to psychological game theory, and it was mathematically difficult to deal with it.

Fehr and Schmidt therefore decided not to treat intentions explicitly, allowing their model to be suitable not only for a two-person game but also in *n*-person games.

First of all, they define an inequity individual as a subject that dislikes outcomes that are not perceived as fairly balanced. The perception and measure of fairness is not an easy matter and it is commonly based on a kind of "reference outcome", that is the result of a complex process of social comparison.

This means clearly that *relative*outcomes(comparison with the others' pecuniary payoff) do affect people's satisfaction of their own outcome and consequently their behavior. This means that along with the monetary payoff, the relative payoff is a mean of motivation for individuals and for this reason they have to be included in a person's utility function. It is also evident from much of the literature that for some people payoffs gained by the others constitute an important constraint when making a choice and could drive the subject even to give up relevant amounts.

In general people are sensitive to inequities in favor of as well as against them when making decisions about how to divide outcomes among them and the others. Sometimes theycould perceive overcompensation and feeling "guilty" because they are obtaining a too high share of the amount, other times they could be simply unsatisfied for having gained an undeserved and too small reward.

Standard game theory's assumption that players only aim at achieving the highest possible reward has been revisited in what can be called a "motivation model" where intentions of people do matter a lot in order to make their final choices.

It has to be specified that the disutility derived by unfair allocation of output takes two forms: there is the loss that comes from material disadvantages and the loss deriving from material advantages.

The equation of inequity-aversion is presented as a linear function and this allows simplicity.

We can describe the utility function of individuals as it follows:

$$U_i(\{x_i, x_j\}) = x_i - \frac{\alpha_i}{n-1} \times \sum \max(x_j - x_i, 0) - \frac{\beta_i}{n-1} \times \sum \max(x_i - x_j, 0),$$

where:

$x_{i,j}$ is the payoff of player *i* (or *j*)

α_i is a parameter of "jealousy" for the others' rewards

β_i is a parameter expressing feeling of "guiltiness"

The maximizations are mathematical notations that mean they are only taken into account when in the parenthesis the value of the differences between the payoffs there exists, therefore are higher than zero.

It is generally assumed that $0 < \beta_i < 1$ and $\alpha_i > \beta_i$ since the disutility that comes from a position of disadvantage is commonly higher than the disutility that comes from a position of advantage. Even if people are altruists it is the case that if they are the ones "damaged" by the unequal distribution will be surely worst off then if they are in the position of receiving the highest amount.

Since the model is implemented in games with more than two players we have to average the sum of all the payoffs 'inequalities for the *n* players in the game minus 1.

Further studies will observe that the size of the sample is relatively important for the choices of the participants since people are naturally lead to interact and influence each other, and this is particularly clear in the example of the public good.

Several experiments and empirical studies provide good indicators of the fact that in many cases people are more cooperative than what it is assumed to be in the basic self-interest model.

2.3 BOLTON AND OCKENFELS

Bolton and Ockenfels in 2000 designed a new model, a bit more complex than the previous ones that they named ERC (equity, reciprocity and competition). Due to its structure this new framework can be either applied to normal form game and to extensive form game. A subject's payoff is derived completely by his own monetary and own relative payoff, making for a quite parsimonious model. This kind of structure allows robust results, even if

we have to consider that in real life the quantitative data for many games are influenced deeply by external factors such as cultures and environment, meaning that there is always some ground for bias.

A simple version of ERC model can provide quantitative solutions in situations formulated as a dilemma-game the one we are really interested in.

The model in general can be determined for *n*-players, $i = 1,2,\ldots n$ .The pecuniary payoffs are positive values represented as $y_i \geq 0$.

Each agent act such as to maximize the value of his *motivation function:*

$$v_i = v_i\,(y_i,\sigma_i)$$

Motivation functions can be considered as a particular kind of utility functions that highlights the objectives that stimulate players' behavior. Bolton and Ockenfels consider in their research the fact that the weights that people give to these objectives can of course change over time, depending on their particular characteristics.

From what we can see in the equation above, the motivation function of each player is dependent from two parameters. $y_i$represent each player monetary gain and $c = \sum_{j=1}^{n} y_j$ is the total monetary payout.

The second parameter $\sigma_i$expresses each player's share of payoff, that can be

$$\sigma_i = \sigma_i\,(y_i,c_i,n_i) = \begin{cases} \frac{yi}{c} & if\ c > 0 \\ \frac{1}{n} & if\ c = 0 \end{cases}$$

Fixing σ and given two alternatives where $v_i\,(y_i^1,\sigma) = v_i\,(y_i^2,\sigma)$ *and* $y_i^1 > y_i^2$ *, player i will choose* (y_i^1,σ). this implies that for a given relative outcome, the player's decision is consistent with the standard assumption made about preferences for money, more is better than less.

A typical payoff function for two players game is given by:

$$v_i\,(y_i, \sigma_i) = a_i y_i - \frac{b_i}{2}\left(\sigma_i - \frac{1}{2}\right)^2 a_i \geq 0 \quad b_i > 0$$

Each player's profile can be described by its preferences' ratio a_i/b_i, that are the weights attributed to the pecuniary and relative components of the motivation function. In the case for which the ratio $a_i/b_i = 0$ means we have a strict relativism, people care the outcome of their opponents that much that almost are not interested in their actual gain. The parameter *c* is the total monetary outcome.

The first component of the equation represents the standard preferences for the monetary payoff, while the second component determines the influence of the comparative effect, it is in fact the value *b* spread over the all population (in this case $n = 2$). What we can infer from this is that, as the difference between player *i* and his opponent's amount increases, the greater the loss he will incur.

As already seen with the other self-interest model, also applying ERC, if a game gives rise to a trade-off between monetary payoff and relative interests, the behavioral pattern observed contradicts the standard theoretical expectations. Unfortunately there is another very important factor that arise in many situations, people do not always play fairly, and for this reason competitive behavior in many cases may determine traditional Nash Equilibrium being the ERC equilibria.

2.4 RABIN AND CHARNESS (2002)

The research on cooperative behavior has gone further during the years focusing even more on social preferences, starting from the assumption that people are both self-interested and also care about the outcomes of the others. In 2002 Matthew Rabin again together with Gary Charness derived another and simpler model to explain "helpful sacrifice" by players.

In their simple linear framework, they assume that propensity of an agent to give up part of his payoff in order to let the opponent gain some more is determined by three parameters:

- The weight on the opponent's payoff when the player gains more (p)
- The weight on the opponent's payoff when the player gains less (σ)
- The change in weight when the opponent behave badly (θ)

Let's consider player 1 and 2, and assume that their respectively monetary payoffs are π_1 and $\pi_2$the formulation of player 2's preferences is

$$U_2(\pi_1, \pi_2) = (pr + \sigma s + \theta q)\, \pi_1 + (1 - pr - \sigma s - \theta q)\, \pi_2 \ ,$$

where

r = 1 if $\pi_2 > \pi_1$ and r = 0 otherwise;

s = 1 if $\pi_2 < \pi_1$ and s = 0 otherwise;

q = -1 if player 1 has behaved badly and q = 0 if he behaved good.

The function expresses the fact that utility of player 2 is determined by a weighted sum of his own monetary payoff and his opponent's gain. The weight that the second player attributes to the payoff of the other depends clearly whether player 1 is receiving a greater or a smaller payoff with respect to his, but also on the good or bad behavior of player 1.
Another way of expressing this utility function is dividing it into two opposite cases:

$$\text{when } \pi_2 \geq \pi_1 \ , \ U_2(\pi_1, \pi_2) = (1 - p - \theta q)\pi_2 + (p + \theta q)\, \pi_1$$

$$\text{when } \pi_2 \leq \pi_1 \ , \ U_2(\pi_1, \pi_2) = ((1 - \sigma - \theta q)\, \pi_2 + (\sigma + \theta q)\, \pi_1$$

The three parameters catch different aspects of social preferences. The first two parameters *p* and σ depends only on the outcomes and do not have any reliance on reciprocity, while the last parameter θ actually provides a procedure for capturing reciprocity but to keep things simple and in an environment of complete information it will not be taken into account.
When considering simple competitive preferences it is assumed that player 2 will always prefer to do the best as possible with respect to his opponent, but at the same time will also take care directly about the payoff that is given to player 1. This means "people like their outcomes to be high relative to the others' payoff" and can be represented through the

assumption $\sigma \leq p \leq 0$. Considering $\sigma \leq p$says that the preferences for outcomes relative to the opponent is at least as great as when gaining less as when gaining more.

3. APPLICATION ON PRISONER'S DILEMMA GAME

In this chapter we are going to see how it is possible to apply the several models described above in a social context, particularly in the Prisoner's Dilemma.

In this well known game, the solution normally provided by game theory is a unique Nash equilibrium in which both player, given the possibility either to *cooperate* or to *deviate*, decide to act both in the most selfish manner and therefore opting for deviation.

Payoffs for the players are symmetric and we will redesign them using the different frameworks. Sometimes they seem not to vary that much, other times more parameters are involved and we are going to look for which levels of these it is possible to sustain cooperation.

We are interested in this, because the NE in Prisoner's Dilemma can be surely considered strategically the best outcome possible, as it is the only stable equilibrium, but we know that in many cases it is not the most efficient one. Cooperate can bring to a most efficient joint outcome in many situations and therefore is important to look at the conditions that can allow for this.

3.1 FAIRNESS MODEL

Rabin's fairness model allows concluding that altruism could drive in some cases each player to sacrifice in order to help the opponent. This consideration admits the existence of one more fairness solution a part from the Nash equilibrium in the Prisoner's Dilemma game: the cooperative outcome(cooperate, cooperate) is in fact another possible equilibrium since we have said that if one player knows that the other will play kindly to him it will cooperate as well. Each of them therefore will be willing to help the other as long as the material payoffs derived from defecting are not too large to overcome fairness motivation.

Considering a Prisoner's Dilemma described by the following table we can rearrange the payoffs using the fairness model:

Table 1

	Cooperate	Defect
Cooperate	*(c-d) ; (c-d)*	*-d ; c*
Defect	*c ; -d*	*0 ; 0*

For a generic player *i:*

1. $U_i(C,C) = \pi_i(C,C) + \tilde{f}_j(C,C)\,(1 + f_i(C,C))$ with both positive values of $\tilde{f}_j$(C,C) and f_i(C,C)

2. $U_i(C,D) = \pi_i(C,D) + \tilde{f}_j(C,D)\,(1 + f_i(C,D))$ with $\tilde{f}_j$(C,D) negative and f_i(C,D) positive

3. $U_i(D,C) = \pi_i(D,C) + \tilde{f}_j(D,C)\,(1 + f_i(D,C))$ with $\tilde{f}_j$(D,C) positive and f_i(D,C) negative

4. $U_i(D,D) = \pi_i(D,D) + \tilde{f}_j(D,D)\,(1 + f_i(D,D))$ with both negative values of $\tilde{f}_j$(D,D) and f_i(D,D)

In the first utility both players play kindly and therefore their expectations are positives also about the other's action. In the second equation there is a situation in which *i* plays kindly,having the belief that the other is defecting, and he's actually doing so. In the third case player *i* acts unkindly being aware that the other is playing cooperative and finally both act unkindly and consequentially their expectations are negatives.

The prisoner's dilemma shows two issues that have also previously discussed. First issue is that the notion of "pure altruism" by the player is inconsistent, as it is true that both people can cooperate reaching the fairness equilibria, but it is also true that if each of them expects the other to defect they will both end up defecting. Moreover player *i*, knowing that player *j* will cooperate, would decide to cooperate as well if and only if he is willing to give up the extra amount *d* and pay a direct cost for cooperation and come in favor his opponent giving him the possibility to gain *c*.

The second issue highlights the role of intentionality in behaving fairly or not. Since people determine their choices also considering possible actions of the others, strategy that potentially could be played (but actually are not) are as much important as the ones chosen.
We can rewrite the utility functions with parametrical values to see more clearly what happens to changing in the value of these numbers when preferences of the players are at the extreme cases.
First thing we have to determine in order to rearrange the payoffs is the value of $f_i$in the different cases.
Using Rabin's equation to determine f_i we have:

$f_i(C,C)= \frac{(c-d)-\frac{(c-d-d)}{2}}{(c-d)+d} = ½$

In this case we have considered $\pi_j^h = (c\text{-}d)$ because it is the maximum outcome possible for *j*given that he decided to cooperate and $\pi_j^{min} = -d$ in the case player *i* decides to deviate. The expected outcome is the average between the two values and therefore will be
$\pi_j^e = \frac{c-d-d}{2}$.

$f_i(D,C)= \frac{-d-\frac{(c-d-d)}{2}}{(c-d)+d} = - ½$

Now again the highest payoff possible is represented by$\pi_j^h = (c\text{-}d)$ and also the minimum still remains $\pi_j^{min} = -d$. The expected outcome is $\pi_j^e = \frac{c-d-d}{2}$.

$f_i(C,D)= \frac{c-\frac{(c+c-d)}{2}}{c-(c-d)} = ½$

This time the maximum payoff for *j*is $\pi_j^h = c$ and the lowest possible considering that i is always cooperating is $\pi_j^{min} = (c\text{-}d)$. The expected payoff in this case is $\pi_j^e = \frac{c+c-d}{2}$.

$f_i(D,D)= \frac{0-\frac{c}{2}}{c-0} = ½$

In the last case the maximum payoff $\pi_j^h = c$ while the lowest in case the opponent defects as well is $\pi_j^{min} = 0$. We have the average payoff represented by $\pi_j^e = \frac{c}{2}$.

All that said we now can consider that when a player believes that the other will deviate, he will feel bad and his utility will decrease; in the case $f_j(\cdot)$ will be equal to – ½ .In the

opposite case if the player is sure that the other will cooperate we will attribute $f_j(\cdot)$ the opposite value $f_j(\cdot) = ½$.

The same procedure has to be applied to the expectation about the opponent's action $\tilde{f}$ that will assume as well either the value of ½ or – ½ depending on the other's behavior.

Utilities finally appear as follows with $c>d$

$U_i(C,C)= (c-d) + ½ (1 + ½) = (c-d) + 0.75$

The player decides to pay the cost of cooperation*d* and the utility obtained from this action is not only the material payoff that remains*(c-d)* but 0.75 additional utility derived by the fact that he feels better responding kindly to the opponent's fairness.

$U_i(C,D) = -d + (- ½)(1+ ½) = - (d + 0.75)$

The player receives the negative material payoff but also suffers a loss given by the feeling of dissatisfaction having behaved good with the opponent cheating on him.

$U_i(D,C) = c + ½ (1- ½) = c + 0.25$

Player gainsthe maximum possible material payoff responding with defection to a kind action of the other and moreover since he has cheated,he doesn't care about the fact that he behaved unfairly and enjoy the other's kindness.

$U_i(D,D) = 0 + - ½ (1- ½) = - 0.25$

They both defected and behaved badly receiving zero pecuniary payoff and moreover suffering a loss deriving from the fact that received unkind behavior of the opponent in turn.

The table now appears as follows:

	Cooperate	Defect
Cooperate	*(c-d)+0.75 ; (c-d)+0.75*	*-(d+ 0.75); c+ 0.25*
Defect	*c+ 0.25 ; -(d+0.75)*	*-0.25 ; -0.25*

We can state, even from what we see in the table that fairness equilibrium outcome is either strictly positive or weakly negative and there will always be a certain symmetry of behavior. It will never be the case in fact that one of the two players at the equilibrium behave kindly and the other unkindly.

Imposing cooperation as the optimal choice when the other acts cooperative we pretend $(c-d)+0.75 > c+ 0.25$. This means that with these level of preferences for all the values of$d$smaller than 0.5 we will have two Nash Equilibrium (C,C) and (D,D) with cooperation that is clearly more advantageous for all the agents.

This would be the case in which the level of goodness of the players and their trust in the fairness of the other's movement, can allow the subjects to feel better and more satisfied than if they would have defected in response to the opponent's cooperation.

If we try to impose cooperation as a dominant strategy, we would require also that $U_i(C,D)> U_i(D,D)$meaning that $-(d + 0.75) > -0.25$. This inequality holds if $d < -0.5$ but because of the construction of the game, d is suppose to be a positive number and therefore it is never the case for these level of payoffs that cooperation could represent the only NE.

3.2 INEQUITY AVERSION MODEL

The inequity aversion model is suitable for different applications in several kind of games. We will see how could be applied in the Prisoner's Dilemma, and then due to its versatility we will develop also a model for the voluntary contribution game, that involves clearly more than two players.

Each player fully understand the game that he faces and behave rationally, that is, asituation in which a social outcome is determined depending on a whole profile ofplayers' choices and they differ in their preferences over social outcomes.

Each player is characterized by the combination of his envy parameter and guilt parameter (α, β), and we have already said that three types are recognizable: the cooperative, competitive or individualistic. We could describe the cooperative's preferences as high-envy-high-guilt with $\alpha = 1$ and $\beta = 1$, meaning that he is willing to choose C if his opponent plays C,due to the fact that choosing D would make him feel guilty. However, if the opponent is playing D, his "evilness" will not allow the other to gain an excessive higher payoff with respect to his and will decide to play D as well. Cooperation and equitable outcomes at the end will be achieved in any case.

The competitive subject is a high-envy-low-guilt type with preferences $\alpha = 1$ and $\beta = 0$ he will not hesitate to choose D both if the opponent defects or plays C because he feels little

guilt in responding to C with D and his utility will be lowered a lot if while him playing C the other responds with D.

The individualistic type can be described by preferences $\alpha = 0$ and $\beta = 0$, since he only cares about material payoff and he has no interests in the relatives outcomes.

Thus,three different types in inequity aversions perfectly follow three different utility functions.

Let's consider a PD that is again described by Table 1 andrewrite the payoffs for player *i* where x_i , x_j as the respective monetary payoffs:

$$u_{(0,\,0)}(x_i,x_j) = x_i$$

$$u_{(1,\,0)}(x_i,x_j) = x_i - 1 \text{ x } max\{x_j - x_i,\, 0\}$$

$$u_{(1,\,1)}(x_i,x_j) = x_i - 1 \text{ x } max\{x_j - x_i,\, 0\} - 1 \text{ x } max\{x_i - x_j,\, 0\}$$

The individualistic feels neither envy or guilty in pursuing his pecuniary payoff x_i, the competitive will feel envy if $x_{i<}x_j$ and therefore his utility will decrease by a disadvantageous inequality $x_i - x_j$ but he doesn't feel guilty is the opposite situation occurs where $x_{i>}x_{j.}$

The cooperative type feels as guilty as envy if payoffs between him and his opponent are not equals, therefore in this case his material gains will be lower by both advantageous and disadvantageous inequality.

Appling the utility functions derived above we will face these results in the Prisoner's Dilemma:

For type (0,0)

	Cooperate	Defect
Cooperate	*(c-d)*	*-d*
Defect	*c*	*0*

For type (1,0)

	Cooperate	Defect
Cooperate	*(c-d)*	*-c*
Defect	*c*	*0*

For type (1,1)

	Cooperate	Defect
Cooperate	*(c-d)*	*-c*
Defect	*-d*	*0*

We can construct six different bimatrixes, mixing the combinations' type of the two players, but it is possible to notice that the Nash Equilibrium will stay the same in five of the cases, with both players defecting (D,D) with exception only in the case either player *i* and player *j*are cooperative. In this situation we find another Nash Equilibrium in (C, C) that also gives higher payoffs to both of the agents. Still *cooperate*does not represent a dominant strategy for the agents, but due to their strong inequity aversion (that would make impossible for them to choose uncoordinated actions) if they know the preferences of the opponent (1, 1) in this case, they will surely prefer higher payoff since they are rational and will act cooperative.

This example can clearly express the importance of the social preferences among the population when such dilemmas occur and that if all the people would adopt a cooperative approach, having perfect information on the others' type, a most efficient solution would be achieved making all the players better-off.

Considering now player 1 with generic preferences (α_i, β_i), we can derive his utilities in PD, considering the parameters in the table as the monetary payoffs.

We will refer to the PD expressed in Table 1, where in the specific the value of *d* represents the direct cost of cooperation, and *c* is the payoff that is given to the opponent.

Adjusting the utility functions with the inequity-aversion model we will derive the following:

$U_{1(C,C)} = (c\text{-}d) - \alpha_i(0) - \beta_i(0) = (c\text{-}d)$

$U_{1(D,C)} = c - \alpha_i(-d - c) - \beta_i(c + d) = c + \alpha_i(c + d) - \beta_i(c + d)$

$U_{1(C,D)} = -d - \alpha_i(c + d) - \beta_i(-d - c) = -d - \alpha_i(c + d) + \beta_i(c + d)$

$U_{1(D,D)} = 0 - \alpha_i(0) - \beta_i(0) = 0$

	Cooperate	Defect
Cooperate	$(c\text{-}d)$, $(c\text{-}d)$	$-d - \alpha_i(c + d) + \beta_i(c + d)$; $c + \alpha_i(c + d) - \beta_i(c + d)$
Defect	$c + \alpha_i(c + d) - \beta_i(c + d)$; $-d - \alpha_i(c + d) + \beta_i(c + d)$	0 , 0

Clearly in case of coordinated actions (D,D) and (C,C) the utilities will coincide with the only monetary payoff, as the players would neither feel better-off nor worst as the outcomes are equitable and this is their main interest.

We can consider now cooperation as a Nash Equilibrium imposing $(c\text{-}d) > c + \alpha_i(c + d) - \beta_i(c + d)$.

In this case we can see that is some particular cases, for some specific values of the parameters, people could have more incentive to sustain cooperation because gaining more in terms of utility as a mix of monetary gain and personal satisfaction.

This situation occurs when

$$(\alpha_i - \beta_i) > \frac{d}{c+d}$$

We have said that the value of α is always higher than the value of β, and both parameters are

included in a closed interval between zero and one, therefore the term on the left of the inequality will be a very small number as the difference between the parameters decreases, but could still satisfy the equation if the difference between *c* and *d* is really high.

This fact can intuitively let understand that if there is very small difference between α and β meaning that people suffer advantageous inequality almost as much as disadvantageous, it is possible that they can allow the opponent gain a very high value of *c*if there is a relatively low direct cost *d*.

Clearly, what is more influential on people's decisions even if they are "fair" players, is the direct cost they suffer opting for cooperation, while giving extra amount to the opponent would not let them change their cooperative intentions.

3.3 EQUITY RECIPROCITY AND COMPETITION MODEL

In dilemma games we observe that if players with a very high preference for self-interest deviate from their equilibrium strategy, the pecuniary payoff of the all participants will increase and they will all be better off.

We stated that also in ERC model as well as in the inequity-aversion model, the main focus has to be done on the preferences' thresholds of the players. Cooperation can be driven by the interactions among the different type of agents and consequently their trade-off between pecuniary and relative gains.

We can better see how important are these factors in the Prisoner's Dilemma using the bimatrix we have seen until now in order to express the normal form of the game.

In order to show the relevance of the trade-offs between monetary and relative profits for ERC predictions, is possible to describe each subject with a motivation function for two players:

$$v_i\,(y_i, \sigma_i) = a_i y_i - \frac{b_i}{2}\left(\sigma_i - \frac{1}{2}\right)^2$$

The relation a_i/b_ias we have already said is the representation of the preferences of each agent, determining therefore his type. The best decision rule again for every individual with preferencesa_i/b_i is achievement of cooperation when it arises a situation in which it strictly dominates defection.

The motivation model of Bolton and Ockenfels is designed for more complicated situations than the ones that show up in PD. We have seen in fact that the model deals with

probability and reciprocity that are not really involved in the social game of our interest. Moreover the framework analyzes the shares of a total payoff because it is meant to describe utility functions when more than two players are involved and each gain is a weighted amount of the overall, implying also the fact that payoffs are not symmetric among agents.

Ultimatum game and market game in particular find very satisfactory results through the use of this model.

Nevertheless we will derivate motivation functions for the simple prisoner dilemma described above:

$$v_{i\,(C,C)} = a_i(c-d) - \frac{b_i}{2}\left(\frac{(c-d)}{2(c-d)} - \frac{1}{2}\right)^2 = \frac{a_i(c-d)}{2}$$

$$v_{i\,(D,C)} = a_i(c) - \frac{b_i}{2}\left(\frac{c}{(c-d)} - \frac{1}{2}\right)^2$$

$$v_{i\,(C,D)} = a_i(-d) - \frac{b_i}{2}\left(\frac{-d}{(c-d)} - \frac{1}{2}\right)^2$$

$$v_{i\,(D,D)} = -\frac{b_i}{8}$$

	Cooperate (C)	Defect (D)
C	$\frac{a_i(c-d)}{2}, \frac{a_i(c-d)}{2}$	$a_i(-d) - \frac{b_i}{2}\left(\frac{-d}{(c-d)} - \frac{1}{2}\right)^2, a_i(c) - \frac{b_i}{2}\left(\frac{c}{(c-d)} - \frac{1}{2}\right)^2$
D	$a_i(c) - \frac{b_i}{2}\left(\frac{c}{(c-d)} - \frac{1}{2}\right), a_i(-d) - \frac{b_i}{2}\left(\frac{-d}{(c-d)} - \frac{1}{2}\right)^2$	$-\frac{b_i}{8}, -\frac{b_i}{8}$

The total monetary outcome is therefore the sum of the pecuniary gains that appear in the table, while the relative payoff is the ratio between each player's pecuniary gain and the overall payout.

With these new values of utilities for the Prisoner Dilemma we can see which relation links the parameters in order to allow cooperation as a Nash Equilibrium.

Again we will need $v_{i\,(C,C)} > v_{i\,(D,C)}$.Meaning $\frac{a_i(c-d)}{2} > a_i(c) - \frac{b_i}{2}\left(\frac{c}{(c-d)} - \frac{1}{2}\right)^2$

Computing the inequality we obtain the following relation:

$$\frac{a_i}{b_i} < \frac{(c+d)}{4(c-d)^2}$$

For this reason, the smaller is *b*the greater the ratio $\frac{a_i}{b_i}$will become, as they are both positive values smaller than one. Consequentially as players have more selfish preferences in order to still reach cooperation we need the amount *c* and *d* to be not too divergent, so that their sum will relevantly exceed the denominator $4(c-d)^2$.

It is straightforward that as people have almost same level of preference as regard their own and the opponent's payoff, cooperation is very likely to be sustained for different values of *c* and *d*.

Looking for cooperation as the unique solution of the game, we should impose that also $U_i(C,D) > U_i(D,D)$. Computing the inequality we find$\frac{a_i}{b_i} < -\frac{1}{2d}\left(\frac{c+d}{2(c-d)}\right)^2 + \frac{1}{8d}$that represents the condition for which *cooperate*is a strictly dominant strategy for all the players. Clearly this inequality makes sense if the value on the right is a positive one, and in order to allow for this we require that $\left(\frac{c+d}{2(c-d)}\right)^2 < \frac{1}{4}$that at the end would pretend $(c-d) > (c+d)$. This is not possible since both *c* and *d* are positive numbers, therefore we conclude that there are no values of the ratio$\frac{a_i}{b_i}$that allow for cooperation as a unique Nash Equilibrium of the game.

3.4 SOCIAL PREFERENCES MODEL

As already mentioned before we will exclude the parameter θ and we will limit to consider reciprocity in the function of players' utility assigning values 0 and -1 depending on the action of the agents.

In the PD game, cooperate will mean the value of q = 0 and defect will determine q = -1.

Let's represent the game and describe the payoffs according to the model:

We will define $\pi_1 = \pi_2 = (c-d)$ when both cooperate, $\pi_1 = \pi_2 = 0$when both defect, $\pi_1 = \pi_2 =$ *c*when the player deviates and the other cooperates and $\pi_1 = \pi_2 = -d$if the player cooperates while the other does not.

Recalling the function of the model:

$U_2(\pi_1, \pi_2) = (pr + \sigma s + \theta q)\, \pi_1 + (1 - pr - \sigma s - \theta q)\, \pi_2$

Where p and σ represent the weight that players attach to the fact that one material payoff is higher than the other, in particular p is about an advantageous inequality and σ about a disadvantageous one.

$U_1(C, C) = (0 + 0 + 0)(c\text{-}d) + (1 - 0 - 0 - 0)(c\text{-}d) = (c\text{-}d)$

$U_1(D, C) = (p + 0)(-d) + (1 - p - 0)c = (1\text{-}p)c$

$U_1(C, D) = (0 - \sigma - 1)c + (1 - \sigma + 1)(-d) = (\sigma - 2)d - (\sigma+1)c$

$U_1(D, D) = (0 + 0 - 1)(0) + (1 - 0 - 0 + 1)(0) = 0$

	Cooperate	Defect
Cooperate	$(c\text{-}d), (c\text{-}d)$	$(\sigma - 2)d - (\sigma+1)c, (1\text{-}p)c$
Defect	$(1\text{-}p)c\ ,\ (\sigma - 2)d - (\sigma+1)c$	$0\ ,\ 0$

If we want to induce cooperative behavior in this prisoner dilemma, we have to impose, as we have seen for previous models, that the gain from cooperating when the other is doing so is greater than the payoff deriving from defection. This means:

$(c\text{-}d) > (1\text{-}p)c$ leading to:

$$p_i > \frac{d}{c}$$

Therefore what really matters is to look at the relationship that lies between the ratio of the two values *c* and *d.*

The intuition is that since for assumption *d* is always smaller than *c*, otherwise we would get negative payoffs for cooperation, the difference between the parameters must be relevant enough so that the payoff $(c - d)$ is more attractive than just the payoff *c* weighted for the importance the player attaches to it $(1 - p)$.

We see that also for this game is not possible to impose cooperation as a strictly dominant strategy, since the payoff given by C when the other is defecting, that is $(\sigma - 2)d - (\sigma+1)c$, is always a negative number. It is not convenient for the agent to play it in this situation but rather defect and accept zero payoff.

4. CONCLUSION

Economic models and studies have assumed for many years that people when making decisions only aimed at pursuing their own monetary self-interest, and did not take care about social objectives.

This vision was clearly to limitative, in everyday situations there is often evidence of the contrary because people are driven by numerous emotions and they clearly have an economic implication.

Different frameworks therefore have tried to incorporate these feelings into mathematical models in order not only to understand human choices but eventually to predict them, knowing the basic parameters that distinguish each individual.

The role of fairness in particular was analyzed under many aspects: in some models, fairness is interpreted as the willingness to respond with good actions to people who behaved good, while responding with unkindness to those who behaved badly. Evidence suggested that subjects are willing also to sacrifice relevant amount of their gain in order to "punish" unkind attitudes. Evidence says that players sacrifice relevant quantities of their material payoffs to reward or punish different kind of others' attitudes.

Rabin suggested that for this reason welfare economics should not only focus on the efficient allocation of material outcomes, but also should provide organizations so that subjects are satisfied about the way they interact with each other.

In the model of Fehr and Schmidt in particular, fairness was intended as the desire of sharing equitable outcomes between the players, and it was represented by the interest in advantageous and disadvantageous inequality, represented by the two parameters α and β. Their level of values describe three main *types*of players and under some conditions subjects achieve almost complete cooperation without any external enforcement.

In ERC model was taken in consideration the reciprocity factor, expressed by the share received to each single player with respect to the whole amount available, weighted for parameter of preferences. It is quite intuitive here how much people's satisfaction for their personal gains and consequentially their utility, is affected by the relation with the opponent's share. It seems unfair getting a too low share with respect to the other.

Rabin again adjusts and evolves his first model with another framework in which the focal point is represented by social preferences.

We could infer that the relation between the preferences parameters in all the models together with particular conditions about the cost of cooperation and the amount given to the other, can always provide cooperative solutions as the most efficient ones. It's harder to impose though that cooperation is the dominant strategy because with defection of the opponent, still playing cooperation seems senseless in all of our examples. When the other instead decides to play cooperative, deviating doesn't seem to be an attractive decision if people are characterized by parameters that stimulate them to behave fairly with the others, obtaining fair behavior in response and let them care about equitable distribution of the outcomes.Nevertheless, the fact that people care about relative results can be identified by a still self-centered attitude, although in a way different with respect to what the generally accepted theory says.

Players could have a propensity to cooperation because a joint success necessarily implies an individual success, as we have seen in fact that the joint outcome provided by cooperation is higher than the one obtained by selfish actions.

To conclude we should state that in general when incorporating emotions into mathematical frameworks we observe that people's satisfaction about their interaction with others plays an important role and it is often as much important as the material payoff they receive.

BIBLIOGRAPHY

Bolton, Gary E. and Ockenfels, Axel (2000), ERC: A Theory of Equity, Reciprocity, and Competition, *American Economic Review* **90** (1): pp. 166-193

Camerer C. and Thaler R.H. (Summer 2003) In Honor of Matthew Rabin, *The Journal of Economic Perspectives* **17** (3): pp. 159-176

Charness G. and Rabin M. (August 2002) Understanding Social Preferences with Simple Tests, *The Quarterly Journal of Economics***117** (3): pp. 817-869

Charness G., Frechétte G. R. and Qin C. (February 2005) Endogenous Transfer in the Prisoner's Dilemma game, *Games and Economic Behavior* **60** (2): pp. 287-306

Doebeli M. and Hauert C. (2005) Models of Cooperation Based on the Prisoner's Dilemma and the Snowdrift Game, *Ecology Letters* **8**: pp. 748-766

Eek D. and Garling T., (2008). A new look to the Theory of Social Value Orientation. *In*: Biel, A., Eek, D., Gärling, T., Gustafson, M., eds. *New Issues and Paradigms in Research on Social Dilemmas.* Springer, pp. 10-26.

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

Geanakoplos J., Pearce D. and Stacchetti E. (1989) Psychological Games and Sequetial Rationality, *Games and Economic Behavior***1**: pp. 60-79

Ottone S. and Ponzano F. (November 2005) An Extension to the Model of Inequity Aversion by Fher and Schmidt, working paper n.58 *Amedeo Avogardo University of Alessandria, Department of Public Policy and Public Choice*

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