UNIVERSITA' DEGLI STUDI DI TRENTO - DIPARTIMENTO DI ECONOMIA

NORM COMPLIANCE: THE CONTRIBUTION OF BEHAVIORAL ECONOMICS MODELS

Marco Faillo
and
Lorenzo Sacconi

Discussion Paper No. 4, 2007

The Discussion Paper series provides a means for circulating preliminary research results by staff of or visitors to the Department. Its purpose is to stimulate discussion prior to the publication of papers.

Requests for copies of Discussion Papers and address changes should be sent to:

Dott. Stefano Comino
Dipartimento di Economia
Università degli Studi
Via Inama 5
38100 TRENTO ITALIA

NORM COMPLIANCE: THE CONTRIBUTION OF BEHAVIORAL ECONOMICS MODELS[♣]

Marco Faillo
Department of Economics, University of Trento.

Lorenzo Sacconi
Department of Economics, University of Trento and Econometica.

1 Introduction

Self-interest and opportunism have recently been the target of criticism by behavioural economics theorists, who urge the introduction of a more complex - and hopefully more realistic - account of economic agents' motivations. A number of researchers have devised models of choice based on the idea that the evidence about how people play, for example, ultimatum, dictator and public goods games can be explained in terms of preferences for equity and/or reciprocity. Articles on the pros and cons of the various models have been published in the most authoritative economics journals.

The question addressed here is whether these models can also be applied to the case in which a norm of fairness is explicitly introduced into the game, for example by *letting the players choose the norm behind a 'veil of ignorance'*. We investigate under what conditions players who have contributed to the choice of the norm actually comply with it even *when such compliance is not compatible with the pursuit of their self interest* and *when the norm cannot be enforced*. Dealing with this problem requires one to abandon the standard self-interest-based model of the rational agent and shift to more sophisticated models based on the idea that agents are driven by a more complex motivational system. But which theory captures the motivations involved in this kind of problem?

We seek to answer this question by considering the results of an experiment based on a one-shot game – called 'the Exclusion Game' – which had three main features: (i) the players could agree on a norm of fairness behind a veil of ignorance (ii) the players' compliance with the norm also affected the wellbeing of individuals who did not participate to the game but who endorsed the norm (iii) the norm was not enforced by any kind of sanction. This situation, which included all the key elements of our problem, can be used as

♣ To appear in Patrizia Sbriglia and Alessandro Innocenti (eds.) , *Games, Rationality ad Behaviour*, Palgrave, London (forthcoming).

a qualitative test to identify which assumptions about the agent's motivation best explain compliance with a shared and impartial norm of fairness.

We begin with a short presentation of the experimental evidence (Section 2). We then assess whether this evidence can be explained using social preferences, reciprocity, and two other recent behavioral models. For this purpose, we recall the main features of some of these models and try to apply them directly in explanation of the experimental results (Sections 3, 4, 5). We conclude with some critical remarks (Section 6) and suggestions for future research (Section 7).

2 Norm compliance: Experimental evidence on the Exclusion game

Sacconi and Faillo (2005) ran an experiment in which subjects had to choose how to play a simple division game (the 'Exclusion Game') before and after they had agreed on a general principle and a specific rule on how to play this type of game.[1] The subjects were divided into three-person groups and were asked to play the Exclusion Game, in which just two players - G1 and G2 (active players) - were able to move effectively in the game, while the third player - G3 (who was a dummy player) – did not have an active choice, and his/her payoff was determined by the active players' decisions. The experimenters provided a sum of money S (with S = €12), and the two active players had to choose how much of S to ask for themselves from among three options: 'Ask for half of S' (€ 6), 'Ask for one third of S' (€ 4), 'Ask for one fourth of S' (€ 3). If both the active players asked for half of S, the third player would receive zero (s/he would be excluded from the benefit), while if they both asked for one third of S, the sum would be split equally among the three players (see figure 1; the third number in each cell is the payoff of the non-active player).

The experiment was divided into three phases. In *phase one* the subjects played the exclusion game twice.[2] In *phase two*, without knowing the outcome of the previous games the subjects were assigned to new three-person groups, the members of which had to agree

[1] The experiment was conducted at the end of 2005 and the beginning of 2006 at the *Computable and Experimental Economics Laboratory (CEEL)* of the University of Trento. It consisted of ten sessions, with 15 subjects for each session, for a total of 150 participants.

[2] In this phase the subjects played the game three times, in three different rounds. At the beginning of each round the three roles were randomly assigned to the members of the group. The selection mechanism was designed so that each player was able to take each of the three roles G1, G2 and G3 in turn. The subjects were told that at the end of the experiment the software would extract one of these three rounds at random, and the player's earnings for phase 1 would be determined according to the outcome of that round.[2] The game was played anonymously and subjects were not aware of the previous rounds' outcomes. This procedure produced two observations for each player in this phase: his/her choice in the G1 role and his/her choice in the G2 role. That is, we had two choices at phase one for each player.

on a general principle and on a more specific rule on how a sum of money should be divided among two active players and a non-active player.

Figure 1. The Experimental Exclusion Game. Payoff matrix

		G2		
		Ask for 3	**Ask for 4**	**Ask for 6**
G1	**Ask for 3**	3,3, 6	3, 4, 5	3, 6, 3
	Ask for 4	4, 3, 5	4, 4, 4	4, 6, 2
	Ask for 6	6, 3, 3	6, 4, 2	6, 6, 0

The experimenters proposed two alternative principles: Principle 1 stated that '*Every player should share in the benefit:, in particular, the player who has not been able to choose should not receive less than the others*', while according to Principle 2, '*Players who are able to choose should claim a larger share of the benefits*. A more specific division rule was derived from each principle (i.e. a rule imposing a 33%-33%-33% division was derived from principle 1, and a rule imposing a 50%-50%-0% division was derived from principle 2). Subjects knew that they would again play the Exclusion Game in the third phase, but they did not know in which role they would do so. They had to reach agreement by means of a voting procedure: the three members had to reach unanimous agreement on both the principle and the rule within a limited number of trials, lest they be excluded from the game. Groups that were able to select a principle and a rule passed to *phase three*, where they were assigned the active or non-active roles and were invited to play the Exclusion Game again. They were now allowed to choose whether or not to implement the rule selected. If they decided to implement the rule, the corresponding game strategy would be selected; if they decided not to implement the rule they could choose one of the alternative strategies. For example, if a player was part of a group which agreed on Principle 1 and on the 33%-33%-33% division rule, and in Phase 3 s/he decided to implement the rule, then strategy 'Ask for 4' was selected; otherwise s/he was allowed to choose one of the other two strategies. In addition, players were asked to predict the opponent's choice by guessing the outcome of the game.

If we assume that players are motivated only by a desire to pursue their maximum monetary payoff, then the unique dominant strategies Nash Equilibrium in of the Exclusion Game is the one in which both the active players choose to ask for half of S (Ask for € 6),

leaving nothing for the third player. The same outcome should be expected in phase one and phase three because, in the absence of enforcement mechanisms, the second phase (selection of the norm behind a veil of ignorance) should not have any effect on their preferences.

Sacconi and Faillo found that:

- about 85% of the players chose 'Ask for 6' at least once in *phase one*.
- considering the players who in phase one chose the 'Ask for 6' strategy (the more selfish ones) at least once, and in phase two chose a rule that implied an equal splitting of S (about 60% of them), those who expected that the other active players would implement the rule (60% of the latter) also chose to implement the rule.

Thus, choosing a rule induced a change in the behaviour of a significant number of players. The most interesting pattern of behavior was that of the class of players (about one third of the sample) which started by choosing a selfish strategy, agreed on the rule imposing an equal division of S (among active and non-active players and behind a veil of ignorance), and complied with the rule expecting compliance by their opponents as well. For these individuals, agreeing on a rule of fairness seemed to be a sufficient condition for compliance with that rule, and they expected the same to hold for their partners. Moreover, this happened even if (i) compliance with rule implied a cost in terms of material self-interest, (ii) no enforcement mechanisms were in operation, (iii) the game is one-shot, i.e. no reputation effects can emerge, (iv) the decision of each active player had no direct effect on the payoff of the other active player but only affected the payoff of the third passive player.

3 Social preferences and reciprocity. Toward a more complex view of human motivations.

Can the results of Sacconi and Faillo's experiment be explained by using social preference and reciprocity models? To answer this question let us recall the origins and the key features of some of these models.

Since the late 1980s an increasing body of economic theory has sought to explain a set of systematic behavioral deviations from standard self-interested hypotheses observed in laboratory experiments, conducted by both economists and social psychologists, and based on very simple games like Ultimatum, Dictator, Public Goods and Trust Games.[3]

[3] For an extensive survey of the early experimental evidence see Kagel and Roth (1995); see also Fehr and Schmidt (2000) and Camerer (2003).

The evidence from these experiments suggests that there exist motivations which cannot be reduced to the consistent pursuit of self-interest.[4] In particular, it seems that people are prone to:

1. punish other subjects who choose a course of action that produces negative effects on their own welfare, even if punishment is costly (*Negative reciprocity*; e.g. behavior of the responders in Ultimatum Games).
2. reward other subjects who choose a course of action that produces positive effects on their own welfare (*Positive Reciprocity*; e.g. behavior of the trustee in the Trust Game).
3. cooperate (do not cooperate) when they expect cooperation (non-cooperation) by other subjects (*conditional cooperation*; e.g. repeated Public Goods Game).
4. punish subjects who violate a shared norm, even if punishment is costly (*punishment of unfair behavior*; e.g. the behavior of the punisher in the Public Goods Game with punishment; and the behavior of the third player in the Third Party Punishment Game).
5. give away part of their wealth in order to improve the wealth of others (*pure altruism*; e.g. the behavior of the dictator in the Dictator Game).

The impossibility of explaining these findings within the standard rational self-interest framework has induced some authors to work out theories based on the idea that human agents are also motivated by consistent other-regarding motivations. These model do not explain deviations in terms of mistakes or even cognitive limitations in the ability to make consistent self-interested choices. On the contrary, according to this approach, individuals are still utility-maximizers - and their choices can consequently be modeled with nearly standard Game Theory tools such as best replies, Nash (psychological) equilibria, etc. - but their utility functions now represent more complex systems of preferences.

Researchers usually divide these theories into two classes. The first comprises theories based on 'Social Preferences' models in which players are assumed to care about the distributive consequences of their choices. The second class consists of 'Reciprocity' theories where players' choices depend on their beliefs about the other players' reciprocal kindness. In what follows, we outline some of the models most representative of these two classes. The literature generally focuses on the relative explanatory power of each model. Researchers working in this area generally start by identifying the set of standard experimental results explained by the relevant model, and they often design new experiments intended both to test some specific feature of the theory and to compare the

[4] See Fehr and Schmidt (2000) and Fehr and Camerer (2002).

performances of alternative models. We will restrict our presentation to the main motivational features of each model, without going into the details of particular experimental results.[5]

3.1 Social preferences.

Social preferences theories are based on the assumption that the utility of generic player *i* is a function of both his/her own monetary payoff (x_i) and the monetary payoffs of the other players involved in the game. Assuming a set N of players, the utility function of player *i* takes the form:

$$U_i = F_i(x_1, ..., x_N)$$

where F_i may be either a linear function either of the players' payoffs or of some social welfare functions.

Alternative theories differ in their assumptions about the way the other players' payoffs enter player *i*'s utility function.

One of the first attempts to deal with other-regarding preference has been the research conducted by the psychologists Loewenstein, Bazerman and Thompson (1989), who sought to estimate individual utility functions by collecting data on subjects' satisfaction with the alternative outcomes of a hypothetical interaction in which they had to divide losses or costs between themselves and an imaginary partner. The main findings of Loewenstein and colleagues were that subjects expressed concern for the comparison between their own payoffs and the payoffs of their opponents, and in particular that most of them were averse to inequality in the distribution of payoffs, with a stronger aversion for disadvantageous inequality than for advantageous inequality. Loewenstein and colleagues also observed that situational factors, like the nature of the relationship between the characters in the story and the type of transaction in which they were involved, had a significant effect on individual utility: aversion for advantageous inequality disappeared when the interaction was described as a business transaction and when the relationship between the two characters was described as negative. In settings like these, people are averse to disadvantageous inequality, but they always prefer to have a payoff higher than the payoff of their opponent.

In this same line of research, Andreoni and Miller (2002), using experimental data based on a modified version of the Dictator Game, observed that the choices of most of their subjects could be explained by referring to three different types of preference: selfish, where players maximize their own material payoffs; utilitarian, where players maximize

[5] A detailed description of the application of these theories to the experimental evidence can be found in in Fehr and Schmidt (2000) and Camerer (2003).

the total surplus; and Rawlsian, where players maximize the payoff of the worst-off subjects.

These three types of preferences are at the basis of the so called 'quasi-maximin' preferences theory introduced by Charness and Rabin (2002), who assume that individuals can be conceived as characterized by a utility function which is a linear combination of three components representing different types of preferences. Heterogeneity among subjects, as stressed by Andreoni and Miller, is captured by allowing individuals to differ in how they weight the different components of the function. Given a set of N players, the utility function of player *i* is defined as:

$$V_i = (1-\gamma)x_i + \gamma W(x_1,...,x_N)$$

where $\gamma \in [0, 1]$ and $W(x_i... x_N)$ is what they call the 'Social Welfare Function', in which the Rawlsian and the utilitarian component are combined:

$$W(x_1,...,x_N) = \partial Min\{x_1,...,x_N\} + (1-\partial)(x_1+...,x_N)$$

with $\delta \in [0, 1]$.

Players are heterogeneous with respect both to their sensitivity to the social component of the utility function - with the extreme case of pure self-interested players for whom $\gamma=0$ – and to the two components of the social welfare function - with pure utilitarian players for whom $\delta=0$, and pure Rawlsian players for whom $\delta=1$.

Fehr and Schmidt (1999) elaborate on Loewenstein et al.'s (1989) hypothesis to provide a formal model of inequity aversion. In their model Player *i* evaluates the consequences of his/her choice in terms of absolute differences between his/her own material payoff and the material payoff of every other player. Mathematically, this is captured by a utility function of the form:

$$U(x_1,...,x_N) = x_i - \frac{\alpha}{n-1}\sum_{j\neq i} Max\{x_j - x_i, 0\} - \frac{\beta}{n-1}\sum_{j\neq i} Max\{x_i - x_j, 0\}$$

with $\beta_i \leq \alpha_I$ and $\beta_i \in [0, 1]$.

Thus *i* suffers both disadvantageous and advantageous inequality, but the utility loss is greater in the former case. Individuals are considered to be heterogeneous with respect to their degree of inequity aversion (α and β).

An alternative model of inequity aversion has been introduced by Bolton and Ockenfels (2000). In this case a player compares his/her payoff with the average payoff of the

population of players, and shows aversion for both favorable and unfavorable inequity. In the case of N players the utility function of player *i* has the following form:

$$U_i = U_i(x_i, \sigma_i)$$

where σ_i is *i*'s relative payoff:

$$\sigma_i = \begin{cases} \dfrac{x_i}{\sum_{j=1}^{N} x_j} & if \sum_{j=1}^{N} x_j \neq 0 \\ \dfrac{1}{N} & if \sum_{j=1}^{N} x_j = 0 \end{cases}$$

The utility function is strictly concave and reach is maximum in $x_i = 1/N$. Unlike in Fehr and Schmidt's model, the player does not care about comparison between his/her own payoff and the payoff of each other player; s/he is only concerned about his/her relative position with respect to the average payoff.

3.2 Reciprocity and Intentionality

Models belonging to the class of 'reciprocity theories' have in common the assumption that players not only assess the interaction consequences on the basis of a welfare criterion but also account for the process by which these consequences are produced. Put differently, they evaluate actions in terms of their consistency with a given purpose. On to this view, in fact, what matters for players of this kind are their own intentions and those of their opponents, and in particular (in a two person case) one player's beliefs about the (intentional) way in which s/he is being treated by the second player, given what the second player believes about the treatment by the first.

These concepts were first formalized by Rabin (1993).[6] He started from the observation that individuals are willing to punish people whose actions have negative effects on their own personal well-being and to reward people whose actions have positive effects on their own personal well-being, even if punishment and reward are costly. The strength of this attitude is assumed to increase as the cost for punishment and reward decreases.

Rabin converts these ideas into a utility function which is a linear combination of a material component reflecting standard self-interest preferences, and a psychological component, which depends upon the player's beliefs about how s/he is being treated by

[6] See Levine (1998) for an alternative model of reciprocity.

his/her opponent. In the case of two players, the mathematical representation of the function is:

$$U_i(a_i, b_j, c_i) \equiv x_i(a_i, b_j) + f'(b_j, c_i)\left[1 + f(a_i, b_i)\right]$$

where $x_i(a_i, b_j)$ is the material component, i.e. the monetary payoff, whereas $f(b_j, c_i)$ and $f'(a_i, b_j)$ are two 'kindness' functions which Rabin defines as measures of how kind player *i* is toward *j*, given his/her beliefs (b_j) about *j*'s strategy, and how kind player *j* is toward *i*, given *i*'s (second order) beliefs about *j*'s (first order) beliefs about *i*'s strategy. In particular, the kindness of a player is measured as the distance between the actual payoff that s/he can induce by choosing a particular strategy and an 'equitable' payoff (x^e) defined as the average of the maximum (x^h) and minimum (x^l) payoffs that s/he can induce:

$$f(a_i, b_j) \equiv \frac{x_j(b_j, a_i) - x^e(b_j)}{x^h(b_j) - x^l(b_j)} \qquad f'(b_j, c_i) \equiv \frac{x_i(b_j, c_i) - x^e(c_i)}{x^h(c_i) - x^l(c_i)}$$

Given the form of the two kindness functions, player *i* has the incentive to reciprocate unkindly a player j's unkind action if s/he believes that *j* is not kind to her. On the other hand, *i* has the incentive to behave kindly if s/he believes that *j* is behaving kindly towards him/her. Thus, players are assumed to condition their action to the perceived motives of the opponents, these being deduced from the strategy that they are believed to choose.

To deal with a utility function which depends on a player's first and second order beliefs, Rabin departs from standard Game Theory and defines the concept of 'Fairness Equilibrium', which is an extension of Psychological Nash Equilibrium based on the theory of Psychological Games (Geanakoplos et al. 1989).[7] By applying his models to the Prisoner's Dilemma[8], Rabin shows, for example, that the strategy combination {cooperate, cooperate} is a fairness equilibrium if both the players believe that the other is playing 'cooperate'. But also {defect, defect} is a fairness equilibrium if both the players believe that the other is playing 'defect'. Player *i's* decision to cooperate depends on his/her feelings about the intentions of player *j*: if *i* expects *j* to choose to cooperate, although s/he

[7] A strategy vector is a psychological equilibrium (i) if each strategy assigns a payoff no smaller than the ones attained by any other feasible strategy, given the opponent's strategy and beliefs (standard Nash Equilibrium definition) and (ii) if the players' beliefs, of any order, are coherent with the equilibrium strategies.

[8] This model has only been applied in the context of two-player normal form games. Dufwenberg and Kirchsteiger (1998) provide an extension of the model for N-player sequential games. See Battigalli and Dufwenberg (2005) for an extension of Psychological Game Theory to dynamic games.

has the possibility to defect, then s/he will consider this choice as kind, and will reciprocate this kindness. In order to stress the importance of perceived intentions, Rabin introduces the example of a Prisoner's Non-Dilemma (p.1289) in which player *j* is forced to cooperate. In this case, according to Rabin's definition of kindness, player *i* will not be able to perceive *j*'s choice as kind and hence will choose to defect.

Social preferences models are very successful in explaining the stylized facts about deviations from purely selfish behavior, although, as shown by some experiments, their pure consequentialist nature prevents them from explaining outcomes that are unfair from a distributive point of view. In particular, if a model based on standard Game Theory is used, it is not possible to deal with situations in which the players' choices, and their evaluations of the outcome of the game, are influenced by their perceptions of the opponents' intentions. At the same time, Rabin's model of reciprocity, in which intentions have a key role, has numerous shortcomings in terms of its application to experimental evidence, and some of the predictions derived from its application to simple games are implausible (e.g. giving more than 50% in the Ultimatum Game; See Fehr and Schmidt, 2000).

Some authors have proposed a solution consisting in the combination of the two approaches and the introduction of complex models. Charness and Rabin (2002), for example, incorporate reciprocity into their model of quasi-maximin preferences. Falk and Fischbacher (2006) develop a theory of reciprocity for extensive form games with a finite number of stages, and with complete and perfect information, that can be considered an extension of Rabin's theory to N-players sequential games, with the additional introduction of a measure of kindness based on equality.

3.3 Social Preference, Reciprocity and the experimental evidence on the Exclusion Game.

We now return to our initial question: to what extent is it possible to explain what Sacconi and Faillo observed in their experiment with the theory that we have just described?

What social preferences models predict about the outcome of the Exclusion Game depends on the definition of the 'group of reference': that is, it depends on the set of subjects with whom the active player compares his/her payoff. If this set consists only of the other active player, then the unique Nash Equilibrium of the game is the strategy vector {Ask for 6, Ask for 6}. If the third player is included in the reference group, and if the weight of the social component of the utility function is sufficiently high (α and β in Fehr and Schmidt model and δ in Charness and Rabin), the Nash Equilibrium becomes {Ask for 4, Ask for 4}. Note, however, that in both cases the active players' choices should not change on moving to phase three of the experiment, i.e. phase two is irrelevant for these models. In order to explain the results of Sacconi and Faillo's experiments by using these models, we should assume that the weight of the social component of their utility function

should change as a consequence of the agreement on the division rule. But this looks like an *ad hoc* assumption, which, moreover, is not contemplated in these theories because they take player's preferences to be stable.

The limitations of these theories in explaining this kind of behavior relates to their conception of 'fairness'. The motivational assumptions at the basis of the theories of 'Social Preferences' can be represented as a combination of self-interest and other-regarding preferences assuming the form of equity concerns in Fehr and Schmidt and Bolton and Ockenfels, or quasi-maximin preferences in Andreoni and Miller and in Charness and Rabin. Outcomes are assessed by means of a combination of two points of view from which the decision maker expresses his/her preference as to consequences: self-regarding and other-regarding preferences, both understood as being the individual's preferential evaluation of the same outcome. Preferences are the individual's primitive affections for consequences. The novelty of these models is that they combine two perspectives: one is the traditional self-regarding perspective in which the individual only considers how him/herself is affected; the other is an impartial perspective in which s/he takes account of how other individuals are also affected, for example the worst off, or the degree of equality amongst all of them. Social norms are not explicitly modeled: they are simply implicitly understood as a perspective from which the individual may be affected by consequences, a perspective which is built into his/her primitive affectivity. No reference is made to the individual's decision to comply with a norm or principle; nor does his/her preference depend on the degrees of compliance by players with respect to a given norm. The decision to behave fairly is simply conceived as resulting from a fixed personal disposition.

With regard to models based on reciprocity, it is obvious that, on assuming Rabin's model of motivation, we would predict {Ask for 6, Ask for 6} as the unique fairness equilibrium of the Exclusion Game, in both phase one and three of Sacconi and Faillo's experiment. This is evident if we consider the fact that, in the Exclusion Game, the active players are just like the 'dictators' in a standard Dictator Game. Each of them is endowed with a sum of S/2 and must decide how much of that sum to give to the third player; his/her own payoff does not depend upon the action of the other active player. How the opponent treats the non-active player does not influence his/her choice. Phase two of the experiment is irrelevant in this case as well, because the agreement has no influence on the perception of players' kindness; i.e. the prediction about the active players' choice does not change on passing from phase one to phase three.

The same prediction – both active players will choose strategy 'Ask for 6' - would derive from application of Falk and Fischbacher's (2006) model, where 'intentionality' is defined as an *attitude* directed toward another player, who then perceives the opponent's

action as intended to generate a favorable or unfavorable consequence specifically for him/herself, such as a manifestation of kindness, friendship or hostility.

Rabin's and Falk and Fischbacher's theories make no reference to compliance with an abstract and impersonal norm of fairness, but only to direct reciprocity. Preference depends on a player's reciprocal and direct kindness, not on the existence of a shared norm of fairness. Thus the presence of shared norm and the presence of a third player, whose payoff depends on the active players' compliance with that norm, are both irrelevant to the outcome of the game.

Neither social preferences nor reciprocity theories model reciprocal conformity with an norm conceived as an impartial treatment. Deontology, conceived as reciprocal compliance with a norm of fairness, is not at the basis of these approaches. According to social preferences theories, fairness is a player's taste, while according to reciprocity theories, what matters is the kindness that each player expresses directly toward his/her opponents, not compliance with impartial norms which could involve any other subject.

Let us now consider two alternative theories based on a different conception of agents' motivational complexity.

4 A model of conformist preferences.

Grimalda and Sacconi (2002, 2005, see also Sacconi and Grimalda, 2007) suggest that two classes of motives for choice can be grasped by two types of preferences of the Self, and by their relative mathematical representation in the corresponding utility function, the hypothesis being that players are motivated by both consequentialist (and mainly self-interested) and 'conformist' preferences.

The model is based on the idea that the same states of affairs σ generated by strategic interaction can be described in different ways according to their relevant characteristics. A first description of states views them as *consequences*, as attributed only to the acting Self, in the case of self-interest or personal consequentialist preferences, or to any person involved in the interaction, in the case of social preferences.

The second type of preferences is what Grimalda and Sacconi call *conformist personal preferences*. Description of states is no less important here, but states are now described as sets of interdependent actions characterized in terms of whether or not they conform with a given abstract principle or ideal of fairness. A pattern of behaviors (a vector of strategies) which fully complies with the abstract principle of fairness – e.g. it maximizes a social welfare function - is defined as the *ideal*. Each player's degree of conformity with the ideal appended to each strategy choice can be determined by seeing whether the *ideal* comes

about through the choice of each player, given what s/he believes about the other parties' choices. For each strategy combination, each player's conformist preferences will depend on a measure of expected reciprocal conformity.

This characterization of the second-type preferences accords more with deontology - that is, the 'compliance' of an act with an ideal - than with consequentialism. The more a state of affairs is expected to comply with the ideal, the more it is preferred by a player (owing to a measure of expected reciprocal conformity at the basis of each player's conformist preference over states). Note, however, that deontology is not simply understood as individual compliance with a norm considered in isolation. The basis of this model is an assessment of the extent to which states resulting from players' collective interactions are consistent with an impartial principle of justice - understood as a social welfare function taking its values at each 'state' of the game generated by the players' joint actions. At the same time, social welfare function values instantiated at each state of the world do not directly enter the utility function of players, as would instead happen in a model of social preferences. On the contrary, what matters in determining the conformist part of a player's utility function is a measure of how much players are expected reciprocally to contribute to implementing the given principle of justice - that is, a measure of the extent to which the agent him/herself is responsible for generating a state consistent with the ideal, conditional on his/her expectation concerning the counterparty's action, and a measure of the extent to which the counterparty is expected to be reciprocally responsible for making the state consistent with the same norm, given his/her expectation concerning the first player's choice. Hence, a player's conformist preferences are his/her preferences for conditional and expected reciprocal consistency (conformity) in both players' actions - in a two-player case - with a shared impersonal norm which stands as both players' ideal of behavior. Such preferences are seen as the basis for a component of the player's individual utility that contrasts with his/her material utility, what we call 'ideal utility'.

Hence, the two types of preferences for a state σ are represented by a comprehensive utility function which is a linear combination of a material component (representing consequentialist personal preferences) and an ideal component (representing conformist preferences).

$$V_i(\sigma) = U_i(\sigma) + \lambda_i F\left[T(\sigma)\right]$$

where $\lambda_I \geq 0$ is an exogenous psychological parameter expressing player *i*'s sensitivity to the ideal component and F is a function representing conditional and reciprocal conformity

with the principle *T,* which in turn is a social welfare function taking its values at each state σ.

The first step in definition of the ideal component is to introduce function *T*, which formally represents a shared ideal of fairness. This is a mapping from the set of states (and first-order utilities attached to them) to a fairness ordering which ranges over states.

Grimalda and Sacconi conceive deontology as the conditional and reciprocal compliance with the fair distribution principle, represented by *T*, that players could rationally have agreed upon in an *ex ante* hypothetical bargaining situation. This contractarian characterization of the ideal principle *T* is crucial to the view that conformist preferences depend on deontology. The idea is that *if* there exists an ideal norm of justice that players accept because they could have rationally agreed on it in an *ex-ante* hypothetical bargaining situation, *then* they may gain satisfaction from expected conditional and reciprocal conformity with that ideal, and they will decide to conform if they expect conformity from their opponents. Hence conformist preferences do not view deontology as a unilateral sense of commitment, but rather as a motivation effective in inducing players to comply with the principle *T* in so far as they expect that the counterparty will also conform with the same principle. As a matter of practical reasoning and intuition – even if not of pure logic - this expectation is based on the assumption that the players could have agreed on such a principle in an *ex ante* bargaining game on the constitutional rules for their interaction. In fact, we may consider this game as 'fictitious play' or also 'cheap talk', for it does not effectively constrain the following interaction but allows players to agree on abstract and impersonal rules of fairness by giving non-binding instructions on how to play 'real' games. Nevertheless, the theory explains how such an agreement may affect preferences and beliefs, and hence become a motivational force effective in inducing players to conform with the principle agreed.

Because the principle agreed is a constitutional contract, the obvious formal representation of *T* is given by the Nash bargaining solution, i.e. the *Nash social welfare function N*

$$T(\sigma) = N(U_1, ...U_N) = \prod_{i=1}^{N} (U_i - c_i)$$

where c_i represents the reservation utility that agents can obtain when the bargaining process breaks down.

The second step is definition of function F by means of two personal indexes of conformity. By elaborating on Rabin (1993), assuming only two agents, and taking the point of view of player *i* (player *j*'s perspective is symmetrical), Grimalda and Sacconi define the following two indexes of personal conformity:

A) *Player i's index of conditional conformity*, or (better) player *i*'s degree of deviation from pure conditional conformity with the ideal principle *T* due to player *i*'s choice, conditional on his/her expectation about player *j*'s behavior, which varies from 0 (no deviation at all) to -1 (maximal deviation)

$$f_i\left(\sigma_i, b_i^1\right) = \frac{T\left(\sigma_i, b_i^1\right) - T^{MAX}\left(b_i^1\right)}{T^{MAX}\left(b_i^1\right) - T^{MIN}\left(b_i^1\right)}$$

where b_i^1 is player *i*'s belief concerning player *j*'s action, $T^{MAX}\left(b_i^1\right)$ is the maximum attainable value of function *T* given *j's* choice according to *i*'s belief, $T^{MIN}\left(b_i^1\right)$ is the minimum attainable value of function *T* given *j's* choice according to *i*'s belief, $T\left(\sigma_i, b_i^1\right)$ is the actual level of *T* when player *i* adopts strategy σ_i, given his/her belief about the other player's behavior.

B) *Player j's index of expected reciprocal conformity*, or (better) player *j*'s degree of deviation from complete reciprocity in complying with the ideal principle *T* - which is also an index that varies from 0 (no deviation at all) to -1 (maximal deviation) - as seen through player *i* 's beliefs about *j*'s action and about what s/he believes concerning player *i*'s choice

$$\tilde{f}_j\left(b_i^1, b_i^2\right) = \frac{T\left(b_i^1, b_i^2\right) - T^{MAX}\left(b_i^2\right)}{T^{MAX}\left(b_i^2\right) - T^{MIN}\left(b_i^2\right)}$$

where b_i^1 is player *i*'s *first order* belief about player *j*'s action (i.e. formally identical to a strategy of player *j*), b_i^2 is player *i*'s *second order* belief about player *j*'s belief about the action adopted by player *i* (i.e. formally identical to a player *i* strategy predicted by player *j*).

These indexes are compounded within the ideal component of the utility function, defined as

$$\lambda_i\left[1 + \tilde{f}_j\left(b_i^2, b_i^1\right)\right]\left[1 + f_i\left(\sigma_i, b_i^1\right)\right]$$

According to this definition, if player *i* perfectly conforms with the ideal, while player *j* is also expected to conform perfectly, then the two individual indexes of deviation take zero values, so that the resulting ideal utility value is λ_i, where the weight $\lambda_i \geq 0$ is an exogenous psychological parameter expressing how important conformity with the ideal is in the

motivational system of player *i*. By contrast, if a player does not entirely conform, while not expecting the other player to entirely conform either, then the two indexes take negative values (possibly −1). Thus, the utility calculation within square brackets for the conformist motivation reduces to *(1–x)(1–y)* (possibly both equal to zero) times the weight λ_i and yields less than λ_i (possibly zero) as ideal utility value.

The overall utility function V_i is the linear combination of the two components

$$V_i\left(\sigma_i, b_i^1, b_i^2\right) = U_i\left(\sigma_i, b_i^1\right) + \lambda_i\left[1+\tilde{f}_j\left(b_i^2, b_i^1\right)\right]\left[1+f_i\left(\sigma_i, b_i^1\right)\right]$$

This suggests that if a player predicts effective reciprocal conformity, so that the ideal utility component enters the utility function, as long as the weight λ_i is high enough, it is possible that the overall utility value reverses the sign of overall preference for a strategy choice σ_i with respect to the same player *i*'s simple consequentialist preference represented by $U_i(\sigma_i, b_i)$. Thus a player will conform with the shared principle if s/he expects the same behavior from the opponent.

4.1 Conformist preferences and the experimental evidence on the Exclusion Game..

By using psychological game theory, and adopting the concept of Psychological Nash Equilibrium, the theory of conformist preferences easily explains the results of Sacconi and Faillo's experiment on the Exclusion Game (for details see Sacconi and Faillo, 2005).

According to the Grimalda and Sacconi model (hereafter G-S model), in the Exclusion Game experiment both the active players, even if they have conformist preferences, should choose 'Ask for 6' in the first phase, because they have no reason to believe that the opponent is going to implement any norm of fairness (the measures of reciprocal conformity are zero). In the third phase the scenario changes. Now the players who have agreed on a fair division rule, and who are characterized by conformist preferences, measure their reciprocal conformity and decide to implement the rule *if they believe that the opponent is doing the same*.

Once the players have agreed on the principle, the strategy vectors {Ask for 4, Ask for 4} and {Ask for 6, Ask for 6} are both psychological equilibria of the Exclusion Game, and the solution of the game depends on the players' beliefs about the opponents' strategy. If these beliefs are compatible with reciprocal conformity then the rule will be implemented. This is exactly what happened in the experiment, and it is proved by the high correlation between the decision to implement the fair division rule and the expectation concerning the opponent's decision to implement it.

The second phase in the experiment explicitly modeled in empirical terms the theoretical idea that players may have second thoughts about the game they are playing and may, by a sort of fictitious play, agree on an abstract division rule for playing games like the exclusion game itself. The experiment gave the second phase only the role of allowing the players to play the same basic game again if they had agreed during the constitutional phase on the rules for playing it. Players participated in the second phase as if under a 'veil of ignorance' about their future roles (active or not) in the actual game, and agreed on a rule that did not bindingly constrain their subsequent behavior but by the activation, if so happens, of their inner conformist motivations. In the experimental situation the theory would predict that agreement on a fair constitutional rule would effectively influence the players' overall preferences and behaviors in so far as they expected each other to comply with the principle agreed - the limits of this affection being fixed by an exogenous weight representing the motivational importance of the ideal for the players. Indeed, a significant number of subjects, by entering the constitutional 'cheap talk' phase, did agree on a fair principle of division. Hence, their compliance with the principle agreed revealed that their preferences had been affected by conformity considerations, this being also suggested by the fact that they were exactly those players who said that they believed in other participants' compliance with the principle.

5 Norms, Preference and Expectations: Bicchieri's model of norm compliance.

In developing her theory of conformity, Bicchieri (2006; 2000) begins by arguing that social norms are needed when there exists a conflict of interests which can be modeled with mixed-motive games, like Prisoners Dilemma, Investment Game, Ultimatum etc. Social norms transform these games into coordination games, given that each individual expects conformity from the other players. From a purely formal point of view, the argument is similar to those put forward by Rabin and by Sacconi and Grimalda, but as we shall see, the logic underlying the decision to conform described here differs in several respects from reciprocity and conformist preferences hypotheses.

Bicchieri develops her argument in two steps. She begins by providing a rational reconstruction, in terms of preferences and beliefs, of what a norm is, identifying a set of conditions for its existence. We quote at length:

"Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motives game. We say that R is a social norm in a population P if there exists a sufficient large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:

1. *Contingency: i* knows that a rule R exists and applies to situations of type S;

2.*Conditional preference: i* prefers to conform to R in situations of type S on the condition that:

2(a) *Empirical expectations*: *i* believes that a sufficiently large subset of P conforms to R in situations of type S;

and either

2(b) *Normative expectations*: *i* believes that a sufficiently large subset of P expect *i* to conform to R in situations of type S;

or

2(b') *Normative expectations with sanctions*: *I* believes that a sufficiently large subset of P expect *i* to conform to R in situations of type S*,* prefers *i* to conform and may sanction *i*''s behavior.

A social norm R is *followed* by population P if there exists a sufficiently large subset $P_f \subseteq P$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b') are met for *i* and, as a result, *i* prefers to conform to R in situations of type S." (Bicchieri, 2006: p. 16).'[9]

According to this definition, an individual will obey a norm even if it dictates actions against his/her self-interest, if

- s/he is aware of the existence of the norm *and*
- s/he believes that a sufficiently large number of people comply with the norm *and*
- either a sufficiently large number of people think that s/he ought to conform *or*
- a sufficiently large number of people are ready to sanction him/her for not conforming.

Bicchieri assumes that what is considered to be a 'sufficiently large number of people' (the size of P_f) varies across individuals. Individuals also differ with respect to the reasons for their conformity. For some individuals, conditions 2(a) and 2(b) are sufficient to induce preferences for conformity, for others 2(b') is needed. In particular, Bicchieri distinguishes three different reasons for conformity: an individual may conform (i) to avoid negative social sanction, (ii) to gain social rewards; (iii) because s/he finds others' expectations

[9] Bicchieri defines two other types of regularity: 'descriptive norms' and 'convention'. The former apply to coordination games and their existence is based only on conditions 1 and 2(a). Here conformity is always dictated by self-interest. Conventions are defined as descriptive norms which apply to coordination games without non-strict Nash Equilibria. (see Lewis, 1969). Conventions may became social norms through the emergence of normative expectations, when braking a coordination mechanism produces negative externalities (p. 57).

reasonable or legitimate. In the first two cases, people may conform without attributing any intrinsic value to the norm, and their decision is guided by external incentives like the fear of sanctions or the desire for social approval. As Bicchieri points out, if we assume that people adhere to norms only for these two reasons, then we should not observe conformity in large and anonymous groups, where monitoring, punishing and rewarding are impossible. It is also true that there are individuals who obey because they attribute a moral value to the norm. In this last case, conformity is not a problem, and there is nothing to explain. The case of interest is the one in which we observe a person conforming with a particular norm in a particular situation even if it is not possible to punish or reward his/her behavior, but not conforming with the same norm in a different situation. In cases like this, the third reason seems to operate, although Bicchieri does not provide a clear definition of what is meant by 'reasonable expectations'. She notes that when individuals are motivated by others' legitimate expectations, if they deviate from the norm they will have to justify their behavior by offering alternative reasons.

To understand the origin of the legitimacy of others' expectations we must follow Bicchieri in the second step of her argument. When we observe the same person conforming with a norm on one occasion and not conforming with the same norm on other occasions, we should focus on the role of contextual factors in promoting or preventing conformity. The idea is that, in these circumstances, people conform with a norm only when it is made salient by some situational cues, such as environmental stimuli, that interact with and activate cognitive mechanisms like attention or focusing. Evidence on this mechanism is provided by the well-known set of experiments on littering conducted by Cialdini et al. (1990) (see also Bicchieri, 2000), which are at the basis of the so called 'focus theory of social norms'. In these studies, experimenters manipulated the environment to focus people's attention on different norms against littering. Subjects were given a piece of litter and their actions were observed in different conditions (in dirty or clean rooms, with the presence of an experimenter's confederate, etc.). One of Cialdini and colleagues' main findings was that when the subjects' attention was focused on what people approve and disapprove of (or what they call an injunctive norm) - by the action of an experimenter's confederate who picked up a piece of litter from the ground or by handbills against littering or promoting recycling - almost all of them did not litter, even in a very dirty environment.[10]

[10] For more recent experimental studies on 'focus theory' see Rege and Telle (2001) and Krupka and Weber (2004). Bohner and Zeckhauser (2004) conduct an experiment based on the Ultimatum Game in which they show that observing others' choices has a significant effect on players' decisions which can also be interpreted in terms of focussing on a particular norm.

Emphasizing the role of situational cues lead us to conceptualize norm compliance as an automatic cognitive mechanism, not as a conscious response to situational cues. This argument apparently clashes with an explanation of conformity in terms of preferences and expectations. But, as stressed by Bicchieri, her treatment should not be interpreted as a description of how people actually decide, but rather as a description of the reasons for conforming of which an individual may be unaware. We become aware of our beliefs and expectations only exceptionally: as when, for example, we encounter completely new situations, or when our decisions have important consequences. What is important is that, even if we decide by adopting rapid and automatic cognitive processes, our decision can still be explained in terms of preferences and expectations which depend on the decision context.

Eliciting a norms, however, is not a simple process; some norms, fairness for example, are very specific, and what is salient in a particular situation is not always clear (Bicchieri 1999; 2000). This produces conflicts of interpretation that may induce individuals to focus on the most favorable norm, giving rise to the self-serving bias described by Rabin (1995).

But what are the cognitive processes involved in the elicitation of expectations about what people do and expect others to do? Drawing on the work of social and cognitive psychology, Bicchieri answers this question by assuming that when we enter a situation we interpret it by means of a process of categorization which, in turn, activates schemata and scripts defined as stylized descriptions of roles and expected actions in a given situation.[11] Her assumption is that "norms are embedded in scripts" (p. 114) and consequently are activated through interpretation of the situation as belonging to a particular category.

The process of categorization, and the definition of a situation as one in which a particular script applies, is influenced by several factors, like past experience, the individuals' goals, and the way in which the situation is presented (framing). A crucial assumption in Bicchieri's argument is that social situations are 'natural categories' (or kinds) rather than 'human artifact categories'. As shown by cognitive psychology, natural kinds are characterized by high inductive potential, and when people (especially children, but also adults) are faced with objects perceived as belonging to this type of category, they

[11] This description of the decision process resembles that of the 'logic of appropriateness' described by March (1994) and according to which individual decisions are based on three fundamental questions: (i) what kind of situation is this? (ii) what kind of person am I? (iii) what should a person like me do in a situation like this? Or what are the relevant rules in this situation?

With regard to the role awareness of choice and the instrumental interpretation of the situation described by Bicchieri, March points out that (i) individuals are supposed to have an active role in identity and rule construction, and the activation and application of rules are conceived as not a completely automatic process; (ii) the strategic use of the rules is considered to be possible; (iii) the automatic internalization of norms is not considered a necessary condition for norm compliance.

tend to apply a reasoning heuristic known as 'psychological essentialism': that is, they conceive these classes of objects as characterized by an underlying and immutable nature that cannot be observed (the essence) and which they tend to infer from surface characteristics. Personality traits, for example, are inferred from physical appearance or from the individual's social role. The point is that if we categorize social situations as natural types, then psychological essentialism may be at work, and in our attempt to interpret such situations we extrapolate essential characteristics, like people's expectations and preferences, from superficial observations.

To sum up, contextual cues induce a particular categorization of the situation enacting a particular script (norm) which describes what kinds of preferences and beliefs we may expect. If we categorize that situation as belonging to a natural kind, then also the enacted interaction script is conceived as stable and invariant, and we feel that our expectations about what should happen are legitimate. Here Bicchieri makes an additional, and fundamental, assumption by stating that these expectations (what is described by the script) tend to become quasi-moral, in the sense that we feel the duty to behave in a certain way and the right to expect particular behavior by others (p. 131). If what happens does not correspond to what is described by the script, as when a person deviates from his/her expected behavior, than we may feel uneasy or be angry with him/her. Empirical and normative expectations are embedded in the script, which tells us what people normally expect and prefer in that situation. And the legitimacy of others' expectations about our behavior – the fact that we found them reasonable - as well the fact that they are ready to punish and reward us, stem from a propensity to commit the naturalistic fallacy by which we identify as right or just what normally occurs.

Bicchieri translates her hypotheses into a formal model in which a norm N_i is defined as a function that maps others' expected behavior into what *i*'s strategies: N_j: $L_{-i} \rightarrow S_i$, where L_{-i} is a subset of the set S_j of strategy profiles of players other than *i*. A given strategy profile $s=(s_1 \ldots s_N)$ *instantiates* a norm for player *i* if N_i is defined in s_{-i} and it *violates* the norm for player *i* if $s_i \neq N_i(s_i)$.

A rational player *i* is modeled as follows: his/her preferences are represented by the following utility function, which is a linear combination of *i*'s material payoff and a component which depends on norm compliance:

$$U_i(s) = \pi_i - k_i \max_{m \neq j} \left\{ \pi_m \left(s_{-j}, N_j(s_{-j}) \right) - \pi_m(s) \right) \right\}$$

where $k_j \geq 0$ measures *i*'s sensitivity to the norm. The second component measures the maximum loss that players other than norm violators (*j*) suffer because of all norm violations.

Thus a player's utility is reduced by a quantity corresponding to the maximum norm follower's loss deriving from norm violation – i.e. the difference between what the worst-off norm follower could get in case of norm compliance and what she actually gets - weighted by $k_{i.}$.

For example, in a Prisoner's Dilemma game with two players who have to choose between 'cooperate' (C) and 'defect' (D) strategies, assume that a norm of reciprocal cooperation has been established. Hence the norm dictating 'cooperation' is defined in C and not defined in D for each player's choice. If player 2 violates the norm, choosing D, player 1's utility will be:

$$U_1(C,D) = \pi_1(C,D) - k_1\left[\pi_1(C,C) - \pi_1(C,D)\right]$$

But player 1 also suffers because of his/her deviations. Indeed, if s/he chooses D, while player 2 chooses C, his/her utility will be:

$$U_1(D,C) = \pi_1(D,C) - k_1\left[\pi_2(C,C) - \pi_2(D,C)\right]$$

Player 1's utility is reduced by a quantity corresponding to the difference between what player 2, who complies with the norm, would get if player 1 complied with the norm and what s/he actually gets.

An interaction can be modeled as a game in which each player is characterized by his/her sensitivity to the norm k_i, and a probability distribution over the others' sensitivity to the norm $k_{j,}$. In the case of a Prisoner's Dilemma, for example, if empirical and normative expectations hold, and k_i and expected k_j are sufficiently large, the player will see it as a coordination game with two Pareto ranked equilibria ((C,C) and (D,D), where (C,C) is Pareto dominant). This idea is captured by a Bayesian game in which the player face a Prisoner's Dilemma or a Coordination game depending the opponent's type, defined in terms of perceived sensitivity to the norm.

Bicchieri applies this model in explanation of the results from standard experiments on the ultimatum game, as well as the evidence from experiments on mini-Ultimatum

Games:[12] ultimatum games with information asymmetries, with framing manipulations, artificial players, etc. The idea is that variations on the standard games may give rise to a change either in what is considered to be the relevant norm (as in the case of framing manipulation) or in the player's sensitivity and/or his/her beliefs about the opponents' sensitivity.

5.1 Bicchieri's theory of norm compliance and the experimental evidence on the Exclusion Game.

If we look at Sacconi and Faillo's experiment through the lens of Bicchieri's theory of norm compliance (hereafter B-theory) what we can conclude is that (i) in its 'pure form' it does not account for the experimental results because players' utility depends only on the consequences on the active players' wellbeing. If a player deviates from a norm which dictates the equal division of the sum among all the players (active and non-active), but at the same time s/he cares only about the effect of his/her deviation on the other active player's payoff, then s/he will have no reason to comply. (ii) In any case, we can extend the model to include the effect of the deviation on the wellbeing of any player (active and non-active). (iii) In Sacconi and Faillo's experiment the second phase (the agreement) induces a cognitive focusing on the rule agreed and elicits descriptive and normative expectations which are at the basis of the change in the behavior of players who in phase one chose to behave in a selfish way. Note, however, that in this context, in which the norm dictates a behavior which contrasts with the players' self interest, what matters seem to be *normative expectations* (other's expectation about the active player's compliance with the rule), which are not detected in Sacconi and Faillo's experiment. (iv) The experiment seems to suggest that agreement is an alternative source of focus for players' expectations. In this context, compliance can be explained as resulting from the emergence of empirical and normative expectations induced by the fact that players actually chose and accepted the norm.

Hence, if we assume, as we have done in the case of Sacconi and Grimalda's model, that agreement on the norm is a sufficient condition for compliance - i.e. for the emergence of empirical and normative expectations - then what we observe in the experiment is compatible with Bicchieri's theory. Some subjects change their behavior on moving from phase one to phase three because they have sophisticated preferences which depend on their

[12] In the mini-ultimatum game Proposer must choose among a limited number of possible allocations. This kind of game has been used to prove the role of Proposer's perceived intentionality in the Responder's judgement of how fair an allocation is. Consider the case of a Proposer who is endowed with 10$ and must choose between two alternative allocations in two different conditions: in condition 1, s/he must choose between ($8, $2) – where the first number is Proposer's payoff - and ($5, $5); in condition 2 s/he can choose between ($2, $8) and ($8, $2). It has been shown (Falk and Fischbacher, 2006) that the rate of rejection of offer ($8, $2) is much lower in condition 2, where the Proposer must choose between a very advantageous and a very disadvantageous allocation.

beliefs about others' behavior and on others' beliefs about their own behavior, and these beliefs are influenced by the agreement on a specific rule of fairness.

6 Critical remarks: eliciting expectation through an *ex ante* agreement

The previous sections have discussed how some of the most popular behavioural economics theories dealing with deviations from the standard self-interest hypothesis fail to explain what happens in a typical situation in which players can explicitly agree on a norm According to both social preferences and reciprocity models, individuals deviate from maximization of their own material interest because they have primitive preferences for a particular distribution of money. But introducing a norm of fairness with which the individual can agree has no influence on these preferences. To explain why these norms may have a role in influencing individuals' choices we should move to models in which they are explicitly modelled as motivational factors. In regard to this line of theorizing we considered two models where a key role is played by expectations about reciprocal compliance. These models provide two alterative explanations of how individual preferences may depend on expectations, and both of them are compatible with what we have observed in Sacconi and Faillo's experiment – this being the empirical basis for our comparative evaluation of the existing behavioural theories. In both cases, however, the evidence is explained *only if* we assume that the agreement elicits beliefs of reciprocal conformity. Further research is needed to study the cognitive mechanisms at the basis of this elicitation process. An important contribution to this endeavour is Bicchieri's 'focus theory of social norms' (hereafter B-theory). But there are some basic difficulties in this approach. They concern how normative expectations are elicited from the description of pre-existing regularities of behaviour, which can be better understood by adopting the perspective of the alternative contractarian Grimalda and Sacconi's model (hereafter G-S model). In this section we concentrate on some critical remarks on this subject. The future research agenda that we derive from them follows in the next section..

In B-theory the basis for both empirical and normative expectations is *categorization*: a given game situation is categorized as belonging to a wider class of situations which defines the range of application for a 'script'. This cognitive fact generates not just the expectation that a regularity characterizing the category's members will also identify the agents' behaviour in the reduced case, but also the prediction that these agents will entertain the normative expectation that they *should* conform to the same regularity. The theory is not very precise about the form of these expectations; but it seems reasonable to conjecture that they do not contain solely the belief that players will conform to the regularity, but also the belief that they *should* conform to it. 'Categorization' is essentially a

matter of *description*: a given situation is recognized by analogy or resemblance as belonging to a wider class of situations, so that by default ('until proof to the contrary' is forthcoming) it can be treated according to what is normally accepted to be true for the typical members of the same class. To make sense of Bicchieri's view, however, categorization must imply two kinds of inference. The first is predictive, and works as in the following example: 'case X is an exemplar of class Y; *normally* it is true for elements of class Y that they behave according to R; hence it follows that also X will behave according to R, and anybody who follows this reasoning will expect that X will behave according to R'. The second kind of inference is not predictive but normative, and it works as in the following example: 'case X is an exemplar of class Y; *normally* it is true for elements of class Y that they behave according to R. Hence it follows that X *should* behave according to R, and anybody who follows this reasoning will expect that X *should* behave according to R'. Whereas the first scheme of inference is a completely understandable example of *approximate* reasoning - that is, a kind of default reasoning [13] - that deduces a perhaps fallible conclusion from a not completely certain set of premises (which means from premises where *default conditionals* are included) - matters are quite different with the second scheme. Why, from the mere descriptive categorization of X in a class within which a regularity of behaviour is observed, should I deduce not just that also X will follow it but that X *should* follow it? Along with the dictum 'an *is* does not imply an *ought*', this seems a much more questionable inference, unless we implicitly assume that a hidden prescriptive premise is at work in the categorisation of 'X is an Y': for example, that 'it is normally *required* that members of class Y behave according to R' or 'it is normally *accepted* that member of class Y *should* behave according to R'. Once these prescriptive presumptions have been introduced, the inference becomes much more compelling, even if it still retains the nature of a 'default', i.e. the format of an approximate deductive syllogism, simply valid until proof to the contrary is given.

However, in B-theory nothing is specified that may play the role of a prescriptive premise. On the contrary, complementing B-theory with the G-S model provides the natural prescriptive component of the cognitive mechanism generating normative expectations from categorization. This is the *agreement under a fair procedure* - i.e. under the 'veil of ignorance'- introduced through the hypothesis that, by means of an *ex ante* contract, the players choose the shared distributive principle of fairness *T*, which then enters their utility functions and becomes the basis for their conformist preference. The same idea was framed in Sacconi and Faillo's experiment, where phase two introduced a setting for developing a

[13] See Reiter (1980), Ginsberg (1987), Reiter and Cresciuolo (1987), Geffner *(*1996). For an application to game theory see Sacconi (2004) and Sacconi and Moretti (2004).

'constitutional agreement' on the rules governing how the division game should subsequently be played. In phase two, the experimenters asked the subjects to agree among themselves without knowing the role that they would take in the successive phase, where they could play a division game (Exclusion Game) only if they had able to agree on some rule in the 'constitutional' phase.

Let us explain how the idea of a fair agreement on the principle *T* may integrate the cognitive model. In a pre-play phase, players make the *ex ante* agreement on principle *T*. Hence most of them subsume under principle *T* the division game that they are actually in a position to play solely because they have agreed on principle *T* for games resembling the one that they are required to play. The framework is nonetheless one in which categorization legitimates inferences employing 'if-then' conditionals based on default modes of reasoning, like those using modalities such as 'Normally true…', or 'Not inconsistent with our current state of information …' or 'True until proof to the contrary is forthcoming'. Let us assume that the situation (game) X is categorised as a T-situation according to some resemblance or (maybe imperfect) relation of belonging. Under T-situations the following default knowledge base and conditionals are cognitively admitted: (i) '*Normally* each individual agrees to principle *T* by a fair procedure '. (ii) '*Normally,* if every individual agrees to principle *T* by a fair procedure, then principle *T* is fair'. (iii) '*Normally,* if principle *T* is fair then every individual *should* act according to *T'*. These are perfectly understandable assumptions and conditionals that a player may endorse if s/he categorizes him/herself and the other players as parties to a T-situation (a situation where players have agreed on principle *T* by a fair procedure). They legitimize the following default inference: 'in so far as X is a T-situation, every players in X *should* act according to *T*'. A simple additional step is needed to get to normative expectations. Players who categorize a game as a T-situation also understand that all players described as members of the same T-situation reason *alike.* Hence by categorizing the situation, one further default conditional is accepted: '*Normally* individuals make *normally true* inferences ('true until proof of the contrary is given')'. Because players 1 and 2 are both parties to a T-situation, player 2 infers that player 1 *should* act according to principle *T* and hence *believes* that s/he should do so. At the same time, since player 1 believes that player 2 makes the above-mentioned inference, s/he also holds the related belief about player 1's *duty* to act according to *T*. This is the normative expectation we were looking for.

Note, however, that the G-S model does not need any normative expectation in order to explain compliance with the *ex ante* agreement. Mutual expectations of conformity are necessary conditions for the explanation, even though they are not sufficient ones. But they are merely *predictive* expectations of first and second order. In fact, player 1's conformity index is defined only when his/her first order belief about player 2's behaviour is given,

while player 2's conformity index is contingent upon player 1's second order beliefs about player 2's beliefs, and player 1's first order beliefs about player 2's behaviour. These expectations do not entail any normative requirement, and consequently need an independent motivational source if they are to affect behaviours. Sufficiency is thus given by an independent and autonomous source of motivation parameterized by the weight λ representing the players' 'desire to be just', or their 'sense of justice'. It may be nullified or decreased by expectations of unconformity, but it is entirely effective if the players expect reciprocal conformity. Beliefs and the dispositional will to conform are two independent components of the ideal part of the utility function, and both of them are related to the agreed principle of fairness *T*. According to the model of conformist preferences, the explanation of principle *T* as an *ex ante* agreed principle of justice suggests that parameter λ is causally related to the principle agreed. In the first phase a constitutional decision under a veil of ignorance, understood as a cooperative bargaining game, is reached. The players then participate in the actual non-cooperative game in which their utility functions are affected by the disposition to comply with the constitutional agreement in a way that also reflects their mutual expectations. Without an *ex ante* agreement on principle *T*, their disposition to comply with shared norms cannot be focused on a definite endeavour. To be sure, in principle the disposition to comply with shared principles of justice could be construed with respect to any *shared* principle, regardless of its contractarian source. But as a matter of fact, the model makes *T* an endogenous variable to the *ex ante* agreement, and Sacconi-Faillo's experiment exactly tests the effectiveness of a 'desire' to comply with a norm which has been *ex ante* agreed by a fair procedure.

Nevertheless, B-theory suggests a way in which also predictive mutual expectations of conformity can be made endogenous to the *ex ante* agreement as understood in the G-S model. The cognitive framework is again defined by *categorization* and by a set of default conditionals admitted as the *normal* mode of reasoning for agents thinking about a given game X under its categorisation as a T-situation (i.e. a game situation such that participants have *ex ante* agreed upon a norm relevant to its solution by a fair procedure). Hence, assume that the game X is categorized by a player as a T-situation, and in T-situations the following default base of knowledge and conditionals are accepted as matters of fact. (i) '*Normally* ('until proof to the contrary') every player has agreed to principle *T* by a fair procedure*'*; (ii) '*Normally* if a player has agreed to principle *T* by a fair procedure, then s/he has the willingness/intention to act upon principle *T'*; (iii) '*Normally* if a player has the willingness/intention to act upon principle *T*, s/he has a *reason* to fulfil principle *T'* (simply, the willingness to act upon it is based on the mere fact of having agreed on it by a completely free, unforced decision). This can be also elaborated into the following conditional (iii') '*Normally* if a player has the intention/willingness to act upon principle *T*

at time *t*, then s/he has an actual reason at time *t* to fulfil principle *T'* (because that intention or willingness was effectively his/her own intention to act according to *T* as revealed by a free, unforced fair agreement on *T*). Last, add the following conditional: (iv) '*Normally* if a player is *rational* and has an *actual* reason to fulfil principle *T* at time *t*, then s/he fulfils *T* at time t'.

Note that all these are just default condtionals amounting to what a player categorizing game X as a T-situation regards to be the *normal mode of reasoning* in such a situation 'until the proof to the contrary is given'. In other words, these conditionals state only what a player categorizing the game X as a T-situation is able to infer because it *is not inconsistent* with his/her present state of information - given that this consists of what can be traced back to his/her experience of agreeing upon principle *T* and of behaving in such situations. It could be said that these conditionals constitute *mental models*[14] for T-situations which the 'typical' player may hold - models that s/he adopts when s/he categorizes the game X as a T-situation.

This does not imply that these conditionals may not be contradicted by some items of evidence that emerges at a later moment. For example, evidence may be forthcoming that another player was confused when agreeing on principle *T* or did not do so in good faith. This would contradict the conditional (ii). Or conditional (iii) would be contradicted if evidence emerged that a player has changed his/her intention in the transition from the *ex ante* agreement phase to the point in time *t*, so that s/he no longer has an actual reason for acting upon principle *T*. This may also happen because evidence shows a 'weakness of the will', so that his/her willingness to act upon *T* loses its effectiveness in transmitting the intention to act at time *t*. Moreover, an item of information could reveal that, at time *t*, there are overriding reasons, other than the one that induced the agreement on *T*, for not acting upon the principle *T* (against conditional (iv)), even though the player actually has some reason to act upon it from the *ex ante* agreement. The unexpected occurrence of all this evidence cannot be logically excluded, because default reasoning is by definition non-monotonic and therefore allows for falsifications as learning proceeds. But it is not included in the categorization of the game X as a T-situation. To be clear, the mental model that a player employs after categorizing the situation as a T-situation simply admits the above-mentioned default base of knowledge and conditionals, *period*. This is the model of *normal reasoning* for a *rational player* in T-situations *until proof to the contrary is given* (i.e. until evidence arises to contradict some of the three 'normal' conditionals) and we assume that

[14] For a general introduction to the theory of mental models see Johnson-Laird (1983); Johnson-Laird and Byrne (1991); applications of this theory to the economic analysis of institutions can be found in Denzan and North 1994); Aoki (2001); Aoki (forthcoming).

(for at least a significant number of agents) no proof to the contrary has hitherto been forthcoming.

In order to complete the deduction of the predictive expectations required by the G-S conformist preference model, we need only to introduce two additional components into the typical player's default knowledge base: (v) '*Normally,* players in T-situations are rational', and (vi) '*Normally,* rational players draw *normal* inferences from the *normally* available knowledge base and conditionals'. When conditions (i)–(vi) are attached to the typical player's categorization of game X as a T-situation, the following inference follows quite naturally: 'In so far as game X is a T-situation, every player in X will do *T*'. Once this inference has been made, expectations deriving from it are also allowed to form. Because the typical player infers that each player will do *T*, s/he believes that any other player (*other* than s/he) will do *T*. Moreover, since normally rational players make the *same* inferences, also rational players *other* than the one from whose standpoint we have argued so far are expected to make them. Hence the typical player is allowed to believe that others believe that s/he will do *T*. A player *i's* second order expectation concerning player *j's* expectation that player *i* will act according the principle *T* is then derived.

This explanation entirely fits with the data yielded by our experiment. We observed that, having agreed on a rule of division by means of a fair procedure, a number of subjects entertained the first order expectation that other participants who had agreed upon the same principle and rule would implement them; and this accords with the observation that they do in fact comply with the principle/rule. The simple fact of having agreed seemed sufficient for them to predict that other players, in that they had also agreed, would comply with the principle. Summarizing intuitively the default reasoning outlined above, subjects seemed to reason as if, given that each of them chose the principle in a fair *ex ante* agreement, there were no reason for them not to comply, and there were no reason for them to expect that other similar subjects would have any reason for not complying 'until proof to the contrary' was given. Of course, not all subjects in the experiment did in fact comply. These seemed to be those who did not follow this format for categorizing T-situation because they had a mental representation of examples of T-situations where some of the default conditionals were *false* (which of course is completely legitimate and admissible given different background experiences). However, to give empirical support for the conformist preference model it suffices that a significant portion (indeed the large majority) of those who enter a pre-play fair agreement seem in fact to categorize the situation in the way that we have described. Then both the conditional willingness to comply and reciprocal expectations of conformity can be traced back to the agreement, and the players' consequent behaviour may be explained as maximizing a utility function with an effective ideal component.

Summing up, B-theory and the G-S model complement each other. The former suggests that there is a cognitive mechanism which generates expectation from categorization, this being complemented by the idea from the latter that a fair agreement is a mechanism which introduces the prescriptive and intentional premises necessary for deducing both normative and predictive expectations. On the other hand, the latter borrows from the former the cognitive mechanism of categorization in order to complement the hypothesis of an *ex ante* agreement with the idea of mental models and default reasoning necessary for deducing not just the motivational disposition to comply with fair shared principles, but also the cognitive state in which players entertain reciprocal expectations of conformity. The result is that the *ex ante* agreement becomes the sole source of the effectiveness of conformist preferences. It is nevertheless true that the two companion lines of argument are still to some extent alternatives to each other. B-theory explains the experimental evidence only in terms of other players' normative expectations (as represented by a player's beliefs) which exert motivational pressure on any player's decision to comply with a norm. By contrast, the G-S model suggests that this pressure derives from an endogenous 'desire to comply with agreed principles of justice', granted that the principles agreed are also cognitively able to induce *empirical* expectations of mutual conformity.

7. Future research agenda

Further theoretical and experimental research is needed to verify the soundness of the conjectures put forward in the previous section, and also to disentangle the two distinct lines of explanation that we have sketched.

First, experimental research should be devoted to gaining better understanding of whether and how an *ex ante* fair agreement as such may be able to elicit not just first order but also second order and purely predictive expectations of conformity - which were not directly tested for in Sacconi and Faillo's experiment. This would yield more detailed information on the reasoning process followed by players in deriving expectations, and on whether these resemble the default inferences outlined in the previous section. *Second*, analogous experimentation should verify whether an *ex ante* agreement is also by itself able to elicit normative expectations (that is, 'without any consistent regularity in the previous behaviour of the players', as happened in the exclusion game experiment, where in the first phase of the game the large majority of subjects acted against the fair division rule before they had agreed on it).

Third, in order comparatively to evaluate how well an *ex ante* agreement performs as an effective cognitive and motivational basis for both types of expectations, distinct

experimental tests of the opposite hypothesis should be devised. These should allow the subjects just to receive descriptions about the regularity of behaviour followed by other players. This means that they have not been able to participate in their settlement either by means of personal decisions or by explicit agreement, or by mere adaptation in an *ex ante* decision situation. To make this assumption, an experimental treatment should have induced them to agree in an *ex ante* decision setting over quite different terms. Alternatively they must have adapted to quite different rules, so that they do not actively endorse the same rule adopted by others, but nevertheless learn it. Even though it is natural to conjecture that the information about such a regularity will induce a predictive belief about the continuation of this pattern of behaviour, researchers should focus on whether a mere description of a behaviour regularity among others would induce a *normative* second order expectation regarding the describer – i.e. whether the subject expects the others legitimately to believe that s/he *should* comply with the same rule. Independently of elicitation of any normative expectations, it should be ascertained whether the mere *description* of a regularity in the behaviour of others exerts a motivational effect on subjects, inducing them to imitate the prevailing rule of conduct *even* if such conformity cannot be interpreted as their material best response (or as a reputation-gathering strategy). Note that the exclusion game provides a frame in which this can be verified, for there is no material advantage in conforming with a regularity which differs from simple rational egoism.

Last, in order to disentangle the two companion but nevertheless distinct lines of explanation discussed in the previous section, experiments should be devised to verify whether significant conformity with an agreed norm characterizes players that simply entertain descriptive mutual expectations of conformity but do not have expectations construable as 'I believe that s/he believes that I *should* comply with the rule *T*'. Of course, these are expectations difficult to disentangle in practice, because an *ex ante* agreement would quite naturally elicit both kinds of second order expectation, without enabling one to conclude that only normative expectations have motivational force on a player's decision to conform. However, this line of inquiry should be pursued because it typically concerns a *crux experiment*. To exemplify, consider the case in which, after an *ex ante* agreement on a principle of division, I do not conform in a first step, whereas the other players do so – with the consequence that they are now expected marginally to decrease their level of conformity in a second step. I then realize that the other players believed and continue to believe that I *should* conform, whereas it is unlikely that they still predict conformity on my part. Can this normative expectation, absent the symmetrical predictive expectation of conformity, induce me to conform on my own? To reverse the example, now assume that, after having agreed on a norm, player *i* has in fact complied with it, but some of his/her fellow partners in the agreement have not done so. While the first player entertains the

expectation that all the other players still expect him/her to continue complying in a second step, s/he obtains information that those who did not previously comply by now believe, because of a sense of 'guilt' toward him/her, that s/he has *not* a duty to conform (while nevertheless continuing to expect him/her to do so for he is believed to be and unconditional complier). Moreover, s/he predicts that, owing to this sense of guilt, many of them will switch to full conformity in the second step. Could this situation elicit conformity on the part of the first player? If yes, this would show that the agent is induced to conform, not by the normative expectations of the other players toward him/herself, but by a different endogenous desire to comply with agreed principles in so far as one may expect mutual conformity. In this case, in fact, the motivational drive cannot derive from any normative component in the other players' expectations, because these expectations are simply predictive. The motivational force instead derives from the agent's independent desire to conform, and its normative premise and causal factor reside in the fact itself of the *ex ante* agreement, provided that it is not frustrated by mutual expectations of non-conformity.

References

Aoki, M. (2001), *Comparative Institutional Analysis*, Cambridge Mass, MIT Press.

Aoki, M. (*forthcoming*), "Endogenizing Institutions and Institutional Change", forthcoming in *Journal of Institutional Economics.*

Andreoni, J. and J. Miller (2002), "Giving according to GARP. An Experimental Test of the Consistency of Preferences for Altruism", *Econometrica*, 70, 737-754.

Battigalli, P. and M. Dufwenberg (2005), "Dynamic Psychological Games", *Discussion Papers IGIER, University Bocconi*.

Bicchieri, C. (1999), "Local Fairness", *Philosophy and Phenomenological Research*, vol LIX.

Bicchieri, C. (2000), "Words and Deeds: A Focus Theory of Norms" in Nida-Rumelin and W. Spohn (eds.), *Rationality, Rules and Structure*, Theory and Decision Library, Kluwer.

Bicchieri, C. (2006), *The Grammar of Society*: *The Nature and Dynamics of Social Norms*, Cambridge University Press.

Bohnet, I. and Zeckhauser R. (2004), "Social Comparison in Ultimatum Bargaining", *Scandinavian Journal of Economics*, 103(3), 495-510.

Camerer, C.F. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton; NJ; Princeton University Press.

Camerer, C., and R. Thaler (1995): "Anomalies: Ultimatums, Dictators and Manners," *The Journal of Economic Perspectives*, 9(2).

Charness, G. and M. Rabin (2002), "Understanding Social Preferences with Simple", *Quarterly Journal of Economics,* 117(13), 817-869.

Charness, G. and U. Gneezy (2000), "What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Game". *University of Santa Barbara, Department of Economics Working Papers Series, N. 11-01*. (Forthcoming in *Journal of Economic Behavior and Organization)*.

Charness, G., E. Havury and D. Sonsino (2001), "Social Distance and Reciprocity: The Internet vs. the Laboratory", *University of Santa Barbara, Department of Economics Working Papers Series, N. 10-01*. (Forthcoming in in *Journal of Economic Behavior and Organization*).

Cialdini. R., C. Kallgren and L. Reno (1990), "A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior", *Advances in Experimental Social Psychology*, 23: 201-234.

Denzan, A. and North, D.(1994), "Shared Mental Models: Ideologies and Institutions", *Kyklos*, 47, 1-31

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

Falk, Armin, and Urs Fischbacher (2006): *A Theory of Reciprocity*, Games and Economic Behavior 54 (2), 293-315.

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

Fehr, E. and K.M. Schmidt (2000) "Theories of Fairness and Reciprocity. Evidence and Economic Applications", *Institute for Empirical Research in Economics University of Zurich Working Paper Series.*

Fehr, E. and C. Camerer (2002), "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists", *Institute for Empirical Research in Economics University of Zurich Working Paper Series.* Forthcoming in Heinrich, J., Boyd, R., Bowles, S., Gintis, H., Fehr, E. and McElreath, E. (eds.), *Foundation of Human Sociality. Experimental and Ethnographic Evidence from 15 Small-Scale Societies*, Oxford, Oxford University Press.

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

Geffner D. *(*1996), *Default Reasoning, Causal and Conditional Theories.* Cambridge Mass, MIT Press.

Ginsberg M. (ed.), (1987) *Readings in Non-monotonic Reasoning*, Morgan Kaufmann Publ. Los Altos.

Grimalda G., L. Sacconi (2002), *The Constitution of the Non-Profit Enterprise: Ideals, Conformism and Reciprocity,* University Cattaneo - LIUC, working paper N. 115 (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=402300).

Grimalda, G., L. Sacconi (2005) "The Constitution of the Not-for-Profit Organisation: Reciprocal Conformity to Morality", *Constitutional Political Economy,* Vol 16(3), 249-276.

Johnson-Laird, P. N. (1993), *Mental Models: Toward a Cognitive Science of Language, Inference and Consciousness*, Cambridge, MA: Cambridge University Press.

Johnson-Laird, P. N. and R.M.J. Byrne (1991), *Deduction*, Hillsdale, NJ: Lawrence Erlbaum Associates

Kagel, J and A. Roth (eds.) (1995), *The Handbook of Experimental Economics.* Princeton: Princeton University Press.

Krupka, E. and R. Weber (2004), "The Influence of Social Norms in Dictator Allocation Decisions" *mimeo* (website: www.andrew.cmu.edu/user/rweber/Norms%2006-03-04.pdf)

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

Lewis, D. (1969), *Convention: A Philosophical Study*, Cambridge, Ma: Harvard University Press.

Loewenstein, G. F, M. H. Bazerman and L. Thompson (1989), "Social Utility and Decision Making in Interpersonal Context", *Journal of Personality and Social Psychology*, 57(3), 426-441.

March, J. G. (ed.) (1994): *A primer on decision making: How decisions happen*. MacMillan, New York.

Pillutla, M. M. (1999), "Social Norms and Cooperation in Social Dilemmas: The Effect of Context and Feedback", *Organizational Behavior and Human Decision Processes,* 78(2), 81-103.

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

Rabin, M. (1995), "Moral Preferences, Moral Constraint and Self-Serving Biases", *mimeo*.

Rege, M. and K. Telle (2001), "An Experimental Investigation on Social Norms", *Discussion Paper No. 310, Statistics Norway, Research Department.*

Reiter R. 1980. "A Logic for Default Reasoning." *Artificial Intelligence*. 13:81-132.

Reiter R.; and Cresciuolo G. 1987. On interacting Defaults. *Reading on Non-monotonic Reasoning*. Moragn Kaufmann Publ, Ginsberg ed.

Sacconi L. (2004). *Incomplete Contracts and Corporate Ethics: a Game-theoretical Model under Fuzzy Information*, SIMPLE 6/03, Siena Memos and Papers in Law and Economics, University of Siena, http://ssrn.com/abstract=402260, forthcoming in F.Cafaggi, A. Nicita and U. Pagano (eds.), *Legal Orderings and Economic Institutions*, Routledge, London (in print).

Sacconi, L. and Faillo, M. (2005), "Conformity and Reciprocity in the "Exclusion Game": An Experimental Investigation" *Discussion Paper. Department of Economics University of Trento.*

Sacconi L. and G. Grimalda (2007) “Ideals, conformism and reciprocity: A model of Individual Choice with Conformist Motivations, and an Application to the Not-for-Profit Case” in (L.Bruni and P.L.Porta eds.) *Handbook of Happiness in Economics*, Edward Elgar, London (in print).

Sacconi L. and S. Moretti (2004), “A Fuzzy Logic and Default Reasoning Model of Social Norm and Equilibrium Selection in Games under Unforeseen Contingencies”, *University of Trento, Department of Economics Discussion Paper, December, No. 13.*

Elenco dei papers del Dipartimento di Economia

2000.1 *A two-sector model of the effects of wage compression on unemployment and industry distribution of employment*, by Luigi Bonatti

2000.2 *From Kuwait to Kosovo: What have we learned? Reflections on globalization and peace,* by Roberto Tamborini

2000.3 *Metodo e valutazione in economia. Dall'apriorismo a Friedman ,* by Matteo Motterlini

2000.4 *Under tertiarisation and unemployment.* by Maurizio Pugno

2001.1 *Growth and Monetary Rules in a Model with Competitive Labor Markets,* by Luigi Bonatti.

2001.2 *Profit Versus Non-Profit Firms in the Service Sector: an Analysis of the Employment and Welfare Implications,* by Luigi Bonatti, Carlo Borzaga and Luigi Mittone.

2001.3 *Statistical Economic Approach to Mixed Stock-Flows Dynamic Models in Macroeconomics,* by Bernardo Maggi and Giuseppe Espa.

2001.4 *The monetary transmission mechanism in Italy: The credit channel and a missing ring,* by Riccardo Fiorentini and Roberto Tamborini.

2001.5 *Vat evasion: an experimental approach,* by Luigi Mittone

2001.6 *Decomposability and Modularity of Economic Interactions,* by Luigi Marengo, Corrado Pasquali and Marco Valente.

2001.7 *Unbalanced Growth and Women's Homework,* by Maurizio Pugno

2002.1 *The Underground Economy and the Underdevelopment Trap,* by Maria Rosaria Carillo and Maurizio Pugno.

2002.2 *Interregional Income Redistribution and Convergence in a Model with Perfect Capital Mobility and Unionized Labor Markets,* by Luigi Bonatti.

2002.3 *Firms' bankruptcy and turnover in a macroeconomy,* by Marco Bee, Giuseppe Espa and Roberto Tamborini.

2002.4 *One "monetary giant" with many "fiscal dwarfs": the efficiency of macroeconomic stabilization policies in the European Monetary Union,* by Roberto Tamborini.

2002.5 *The Boom that never was? Latin American Loans in London 1822-1825,* by Giorgio Fodor.

2002.6 *L'economia senza banditore di Axel Leijonhufvud: le 'forze oscure del tempo e dell'ignoranza' e la complessità del coordinamento,* by Elisabetta De Antoni.

2002.7 *Why is Trade between the European Union and the Transition Economies Vertical?,* by Hubert Gabrisch and Maria Luigia Segnana.

2003.1 *The service paradox and endogenous economic gorwth,* by Maurizio Pugno.

2003.2 *Mappe di probabilità di sito archeologico: un passo avanti,* di Giuseppe Espa, Roberto Benedetti, Anna De Meo e Salvatore Espa.
(*Probability maps of archaeological site location: one step beyond,* by Giuseppe Espa, Roberto Benedetti, Anna De Meo and Salvatore Espa).

2003.3 *The Long Swings in Economic Understianding,* by Axel Leijonhufvud.

2003.4 *Dinamica strutturale e occupazione nei servizi,* di Giulia Felice.

2003.5 *The Desirable Organizational Structure for Evolutionary Firms in Static Landscapes,* by Nicolás Garrido.

2003.6 *The Financial Markets and Wealth Effects on Consumption* An Experimental Analysis, by Matteo Ploner.

2003.7 *Essays on Computable Economics, Methodology and the Philosophy of Science,* by Kumaraswamy Velupillai.

2003.8 *Economics and the Complexity Vision: Chimerical Partners or Elysian Adventurers?*, by Kumaraswamy Velupillai.

2003.9 *Contratto d'area cooperativo contro il rischio sistemico di produzione in agricoltura,* di Luciano Pilati e Vasco Boatto.

2003.10 *Il contratto della docenza universitaria. Un problema multi-tasking*, di Roberto Tamborini.

2004.1 *Razionalità e motivazioni affettive: nuove idee dalla neurobiologia e psichiatria per la teoria economica?* di Maurizio Pugno.
(*Rationality and affective motivations: new ideas from neurobiology and psychiatry for economic theory?* by Maurizio Pugno.

2004.2 *The economic consequences of Mr. G. W. Bush's foreign policy. Can th US afford it?* by Roberto Tamborini

2004.3 *Fighting Poverty as a Worldwide Goal* by Rubens Ricupero

2004.4 *Commodity Prices and Debt Sustainability* by Christopher L. Gilbert and Alexandra Tabova

2004.5 *A Primer on the Tools and Concepts of Computable Economics* by K. Vela Velupillai

2004.6 *The Unreasonable Ineffectiveness of Mathematics in Economics* by Vela K. Velupillai

2004.7 *Hicksian Visions and Vignettes on (Non-Linear) Trade Cycle Theories* by Vela K. Velupillai

2004.8 *Trade, inequality and pro-poor growth: Two perspectives, one message?* By Gabriella Berloffa and Maria Luigia Segnana

2004.9 *Worker involvement in entrepreneurial nonprofit organizations. Toward a new assessment of workers? Perceived satisfaction and fairness* by Carlo Borzaga and Ermanno Tortia.

2004.10 *A Social Contract Account for CSR as Extended Model of Corporate Governance (Part I): Rational Bargaining and Justification* by Lorenzo Sacconi

2004.11 *A Social Contract Account for CSR as Extended Model of Corporate Governance (Part II): Compliance, Reputation and Reciprocity* by Lorenzo Sacconi

2004.12 *A Fuzzy Logic and Default Reasoning Model of Social Norm and Equilibrium Selection in Games under Unforeseen Contingencies* by Lorenzo Sacconi and Stefano Moretti

2004.13 *The Constitution of the Not-For-Profit Organisation: Reciprocal Conformity to Morality* by Gianluca Grimalda and Lorenzo Sacconi

2005.1 *The happiness paradox: a formal explanation from psycho-economics* by Maurizio Pugno

2005.2 *Euro Bonds: in Search of Financial Spillovers* by Stefano Schiavo

2005.3 *On Maximum Likelihood Estimation of Operational Loss Distributions* by Marco Bee

2005.4 *An enclave-led model growth: the structural problem of informality persistence in Latin America* by Mario Cimoli, Annalisa Primi and Maurizio Pugno

2005.5 *A tree-based approach to forming strata in multipurpose business surveys,* Roberto Benedetti, Giuseppe Espa and Giovanni Lafratta.

2005.6 *Price Discovery in the Aluminium Market* by Isabel Figuerola-Ferretti and Christopher L. Gilbert.

2005.7 *How is Futures Trading Affected by the Move to a Computerized Trading System? Lessons from the LIFFE FTSE 100 Contract* by Christopher L. Gilbert and Herbert A. Rijken.

2005.8 *Can We Link Concessional Debt Service to Commodity Prices?* By Christopher L. Gilbert and Alexandra Tabova

2005.9 *On the feasibility and desirability of GDP-indexed concessional lending* by Alexandra Tabova.

2005.10 *Un modello finanziario di breve periodo per il settore statale italiano: l'analisi relativa al contesto pre-unione monetaria* by Bernardo Maggi e Giuseppe Espa.

2005.11 *Why does money matter? A structural analysis of monetary policy, credit and aggregate supply effects in Italy*, Giuliana Passamani and Roberto Tamborini.

2005.12 *Conformity and Reciprocity in the "Exclusion Game": an Experimental Investigation* by Lorenzo Sacconi and Marco Faillo.

2005.13 *The Foundations of Computable General Equilibrium Theory*, by K. Vela Velupillai.

2005.14 *The Impossibility of an Effective Theory of Policy in a Complex Economy*, by K. Vela Velupillai.

2005.15 *Morishima's Nonlinear Model of the Cycle: Simplifications and Generalizations,* by K. Vela Velupillai.

2005.16 Using and Producing *Ideas* in Computable Endogenous Growth, by K. Vela Velupillai.

2005.17 *From Planning to Mature: on the Determinants of Open Source Take Off* by Stefano Comino, Fabio M. Manenti and Maria Laura Parisi.

2005.18 *Capabilities, the self, and well-being: a research in psycho-economics,* by Maurizio Pugno.

2005.19 Fiscal and monetary policy, unfortunate events, and the SGP arithmetics. Evidence from a *growth-gap* model, by Edoardo Gaffeo, Giuliana Passamani and Roberto Tamborini

2005.20 *Semiparametric Evidence on the Long-Run Effects of Inflation on Growth,* by Andrea Vaona and Stefano Schiavo.

2006.1 *On the role of public policies supporting Free/Open Source Software. An European perspective,* by Stefano Comino, Fabio M. Manenti and Alessandro Rossi.

2006.2 *Back to Wicksell? In search of the foundations of practical monetary policy*, by Roberto Tamborini

2006.3 *The uses of the past*, by Axel Leijonhufvud

2006.4 *Worker Satisfaction and Perceived Fairness: Result of a Survey in Public, and Non-profit Organizations*, by Ermanno Tortia

2006.5 *Value Chain Analysis and Market Power in Commodity Processing with Application to the Cocoa and Coffee Sectors*, by Christopher L. Gilbert

2006.6 *Macroeconomic Fluctuations and the Firms' Rate of Growth Distribution: Evidence from UK and US Quoted Companies*, by Emiliano Santoro

2006.7 *Heterogeneity and Learning in Inflation Expectation Formation: An Empirical Assessment*, by Damjan Pfajfar and Emiliano Santoro

2006.8 Good *Law & Economics* needs suitable microeconomic models: the case against the application of standard agency models: the case against the application of standard agency models to the professions, by Lorenzo Sacconi

2006.9 *Monetary policy through the "credit-cost channel". Italy and Germany*, by Giuliana Passamani and Roberto Tamborini

2007.1 *The Asymptotic Loss Distribution in a Fat-Tailed Factor Model of Portfolio Credit Risk*, by Marco Bee

2007.2 *Sraffa?s Mathematical Economics – A Constructive Interpretation*, by Kumaraswamy Velupillai

2007.3 *Variations on the Theme of Conning in Mathematical Economics*, by Kumaraswamy Velupillai

2007.4 *Norm Compliance: the Contribution of Behavioral Economics Models*, by Marco Faillo and Lorenzo Sacconi

PUBBLICAZIONE REGISTRATA PRESSO IL TRIBUNALE DI TRENTO