

# On Preferences for Fairness in Non-Cooperative Game Theory

Loránd Ambrus-Lakatos

23 June 2002

Much work has recently been devoted in non-cooperative game theory to accounting for actions motivated by fairness considerations (see the overviews in Fehr-Schmidt (2001) and in Sobel (2001)). One may think of several reasons why this topic has received so much attention in the last decade. First, the neglect of moral motivations in game theory can issue in intellectual uneasiness on the part of its practitioners. Indeed, if fairness considerations are excluded from a framework for studying rational deliberation, then agents cannot be conceived of as following dictates of both morality and rationality. Second, many game theorists hold that the challenges presented by certain simple games, such as the Prisoner's Dilemma and the Ultimatum Game, ought not be left unanswered. Others claim, next, that certain focal economic phenomena, like patterns of contracting on putatively competitive markets, can be explained only by means of models that explicitly involve attitudes towards fair behavior (Falk-Fehr, 2002). All the three of these reasons point to the need of completing game theory, so that it could exhibit agents who act upon moral reasons.

Now, in my view, non-cooperative game theory cannot be revised so that it could account for actions that are undertaken because of fairness considerations. There are many game situations in which fairness considerations cannot be represented at all, and yet there are many others in which fairness considerations can indeed be represented but they call for actions that no version of game theory can account for. In this paper I will argue that one well-known attempt for revision, that of Rabin (1993), certainly fails to achieve its objective. I would like to emphasize that my criticism is not issued by the facile and misguided view according to which game theory unduly focuses on selfish motives as against altruistic ones. Also, while it may be thought that my claim cannot be properly defended without reference to a thesis, or to a full-fledged theory, on how fairness considerations and their motivating force could be conceived, for my current purposes, I will not need to propose any such view or theory.

Below, I will first present, in section 1, the standard or orthodox view of game theory on the issue of fairness in games and will also assess to what extent it is correct. Then in section 2, I will discuss and critically examine Rabin's concept of fairness equilibrium. Section 3 offers concluding remarks.

## 1. THE STANDARD VIEW AND ITS LESSONS

The case for including fairness or moral considerations into non-cooperative game theory is many times expressed in the following, rather familiar, way. Suppose two agents share a situation in which both can take one of two actions, and these are labeled as  $C$  and  $D$ , respectively. Any pair of actually chosen actions determine an outcome, and agents form preferences over each such outcome. Suppose also that the potential outcomes are in fact evaluated by the numbers indicated in the (bi)matrix below (*Table 1.1*). Each cell of the matrix represents an outcome, and in each cell the left hand side number shows how much (subjective) value the agent-player who chooses a row

attaches to the corresponding outcome. Similarly, right hand side numbers show the evaluations of the agent-player who is to choose a column.

Table 1: A standard Prisoner's Dilemma game

	<b>Cooperate</b>	<b>Defect</b>
<b>Cooperate</b>	4, 4	0, 6
<b>Defect</b>	6, 0	1, 1

The situation is modeled then as game, taken in the sense of game theory. According to the standard typology, the two agents face a Prisoner's Dilemma.

Now if the agents follow the precepts of game theory, they both choose  $D$ . This precept can be justified in various ways, but it suffices here to invoke an argument that invokes the so-called the Domination Principle. According to this Principle, if there is an action for an agent that yields more payoffs than any other no matter what the other does, this action is to be undertaken. Now if  $D$  is chosen, one gets (strictly) more than if  $C$  is chosen, no matter what the other does. That is,  $D$  (strictly) dominates  $C$ , for both. However, it may be also argued that by opting for  $D$  the players manifestly disregard moral considerations, as the outcome that would result from both choosing  $C$  would be better for both. Therefore, if an agent carries out the advice of the theory, thereby she shows herself as not being motivated by demands of morality. This argument may also emphasize that by adhering to the Domination Principle each cares about what would be best for her, instead of what would be best for both. Now it is possible to hold that in generic practical situations agents should try to attain the outcome that they consider as the best together with the claim that what is best here is the outcome generated by each choosing  $C$ . Hence the view that game theory, as a framework for studying rational deliberation, fail to recognize what rational agents, in some situations including that of the Prisoner's Dilemma, should regard as the best outcome and therefore what should motivate them the best. Indeed, one may add that game theory also fails as a predictive account, stressing that many agents would indeed take the  $C$  option because that is what they think leads to the best outcome. Nevertheless, game theory cannot make that prediction.

According to many game theorists those who forward this argument are mistaken about what the theory is all about. These theorists favor what I will call the Standard View (in honor of Binmore (1994)). This view recognizes a canonical model for non-cooperative games, holds that what is to be done in games is a matter of employing a relevant solution concept, and concludes that in the Prisoner's Dilemma rational agents are to choose the option  $D$  as this is what the relevant solution concepts recommend one to do.

Below, I will only consider finite non-cooperative games in normal form. The canonical model of these games consists of a collection of mathematical objects (Osborne-Rubinstein, 1994: 11). The first of these is the finite set  $N$  of agents who participate in the situation. The second is the set of actions  $A_i$ , or courses of actions, available to each of the participants,  $i = 1, \dots, N$ . Next, each

possible combination of actions determines an outcome the situation may reach. Finally, according to this model, each agent  $i$  ( $i = 1, \dots, N$ ) has formed a preference ranking  $R_i$  over the potential outcomes. If their preferences meet, as it is standardly supposed they do, some further structural criteria, they can be represented by numbers, called payoffs: that is, for each agent, and for each two outcomes, the outcome that has a higher payoff attached to it is also preferred to the other one. The agents are then taken as wishing to achieve as high a payoff number as possible. As a matter of postulation, this is what motivates them to act in this or that way, that is to choose this or that action from among those available to them. Of course, they also have to take into account what their counterpart will most likely to do.

The precepts of game theory are formulated in terms of solution concepts, that is statements about what each participant, if rational, will do in a given game or scenario. Solution concepts need to meet some general criteria. They have to be, first of all, reasonable. Second, they need to be justifiable to each agent. One such solution concept is the "equilibrium in strictly dominating actions", claiming that when each agent has an action that strictly dominated all others, each is to take that action. Thus in a Prisoner's Dilemma situation, the Domination Principle provides a justification for choosing action  $D$ , as the suggestion that one is to choose an action that yields the most no matter what one's partner will do is to be regarded as reasonable. And as the situation is symmetric, this justification is addressed to both players. Solution concepts are assessed by further criteria, including the one according to which they need to be applicable in as many situations as possible. They also need to be simple and relevant, and, if possible, are to give a unique answer to the question "what shall I do in this situation?". While the solution concept based on the Domination Principle does not meet some of these latter requirements, game theorists agree that it should be used when it is applicable. Now both playing  $D$  is also a Nash-equilibrium here. For a pair of action to be a Nash-equilibrium, it has to be true that for each his action is the best response to the equilibrium action(s) of the other(s). So on the Standard View, whoever argues that action  $C$  should be chosen in a Prisoner's Dilemma ought to put forward a candidate solution concept that while meeting the above criteria also recommends taking  $C$ . The critical argument presented at the beginning of this section did, however, fail to identify such a solution concept.

The argument according to which, in supplying precepts for how to play in the Prisoner's Dilemma, game theory unduly neglects moral considerations points to the fact that the outcome generated by each player choosing  $C$  is worse for both than the outcome generated by each choosing  $D$ . It continues with claiming that this fact should induce them to opt for  $C$ . Now while this claim may well be right, it is a claim that game theory cannot issue as it seeks an answer to the question: "which action should I choose so that by that I could attain as high a payoff as possible?"; and aiming at an outcome that is relatively better for all agents is outside of the scope of this question. Besides, as outcomes are determined by the actions of each, there is of course no guarantee that by choosing  $C$  one achieves the putatively moral outcome. Still, the argument also implies that game theory asks the wrong question, as the outcome generated by the dictates of its favored solution concept can be described as an outcome that is morally unwarranted.

However, on the Standard View, game theoretical analysis has to take for granted what the agents actually value or prefer the most. Surely, the significance of a Prisoner's Dilemma game is

enhanced if it models a, let us say, real situation in which agents could have concerns other than achieving as high a payoff as possible. In the game model, however, players are taken as already having formed their preferences over the various potential outcomes, and these preferences may or may not reflect moral considerations on their part. Preferences being thus given, the only residual question there is for the players to answer is how to achieve as good an outcome as possible; that is the question of which solution concept is to be employed. On the Standard View, the outcomes and the payoffs rendered to them are not to be further interpreted, as this is not the task of game theory. Suppose that Peter and James face a Prisoner's Dilemma where the label  $C$  stands for the option of leaving a blind child in a forest on a very cold December night, and  $D$  stands for the option of sheltering and otherwise caring for him in a warm and cozy home, on the same night. Suppose also that the payoffs of these wicked men are as in *Table 1.1*. It would be clearly absurd to claim that the morally warranted outcome in this case is that spanned by both choosing  $C$ .

On the Standard View, a particular game model is prior both to its interpretation and to its analysis. This could be expressed by means of the following pair of claims. A model of a non-cooperative game lays out all the ex ante reasons there are for actions in terms of structure of the available actions themselves and of the possible payoffs they may determine. And, second, beliefs about what the other will most likely do would also count as ex ante reasons, but there is no grounding in the theory of non-cooperative games for such beliefs. Therefore solution concepts can be based only on available ex ante reasons for actions. Let me explain.

In case an agent opts for the action  $D$  in the Prisoner's Dilemma in *Table 1.1* because this is what the Domination Principle dictates, her decision can only be based on the structure of her own possible actions and payoffs, she does not need to take into account beliefs about what her partner will do. In case she chooses  $D$  because that is part of a Nash-equilibrium, she is to consider that the other also will opt for  $D$  and that  $D$  is the best response to the other's doing  $D$ . But the first of these considerations is not available to her straightforwardly, it is a matter of a hypothesis provided by the theory itself. In so far as game theory is to account for the deliberation of the agents, in this situation, in terms of Nash-equilibrium, it views the agents as choosing the best response given an imputed belief. So the account of deliberation is completed by a consideration of what an equilibrium could be, namely that each agent happens to believe that the other will choose his equilibrium action. But  $(D, D)$  is an equilibrium partly because both agents deliberate properly in this equilibrium, given their imputed beliefs, that is they both take the action that is the best response to what the other is (supposedly) believed to choose. Next, one may be able to explain or predict what has been or will be done in a non-cooperative game only if rationality, as conceived by the theory itself, is assumed to direct the choices of the agents. One is to conclude that solution concepts combine considerations of deliberation, criteria for equilibria, and the ambition to explain or predict actions in game situations.

It is, at the same time, indisputable that in Prisoner's Dilemmas, each agent achieves a relatively low payoff, lower than what they could possibly achieve if they disregarded the precepts inscribed into the solution concepts of game theory. Given this, the theory can be, and has been, regarded as self-defeating. But being self-defeating does not, of course, imply that the theory is defeated because it fails to incorporate obvious moral considerations. It may only mean that it fails to provide good

precepts to agents for attaining what they value most. Again, this, in principle, could be something that is valued on moral grounds. As agents are taken as justified to have the preferences they have, they cannot fail because they do not prefer the most the outcome recommended by some moral consideration. They form their preferences over a list of potential outcomes, and the canonical game model provides no information about the basis of these preferences. So we conclude that game theory is indeed cold to fairness or moral considerations.

Nevertheless, in my view, it is reasonable to consider taking action  $C$  in a Prisoner's Dilemma situation. Let us grant to the theory that the payoffs in this situation represent a legitimate ranking of the possible outcomes in terms of how much one would gain in each of them. At the same time, one may recognize that the other player also has a legitimate claim to attain the largest possible level of payoffs. Then one may also be capable of ranking the outcomes in terms of how fair they are. However, no agent can determine alone a particular outcome, as the action of the other will also affect that. So two considerations compete: that of gaining as large a level of payoffs as possible and that of attaining as fair an outcome as possible. At the same time these considerations cannot straightforwardly traded against each other, as they call for two different sorts of deliberative reasoning. This is because while determining how to reach the maximal payoff level one needs to attend to the structure of one's payoffs and that of the available actions for each, together with an assessment of what the other will likely do; and for determining what is fair to do it is enough to find out what the most fair outcome is, and there is no need to consider what the other will likely do. So *in lieu* of a principle that could hedge between these two considerations, it may be reasonable to decide that the second trumps the first and choose  $C$  indeed. Especially if one happens to judge, on the spot, that the other also will choose  $C$ . The recommendation of game theory to always choose  $D$  appears therefore as unreasonable. When an agent deliberates about what to do in a Prisoner's Dilemma, she may not be ready to adopt the priority thesis on which the Standard View is based.

To summarize, if the Standard View is accepted there is no room for fairness considerations in game theory. It remains to be seen whether game theory can indeed abandon the Standard View.

## 2. RABIN'S MODEL OF FAIRNESS

The priority thesis, that is the stance according to which a particular game model is prior to its interpretation and to its analysis, is especially difficult to defend in the case of the so-called Ultimatum Game. In this game, two players are to distribute among themselves a certain sum, say 100 units. One of them, the Proposer, can first offer a split of this sum to the other, and we stipulate that offers have to be phrased in integer multiples of units. Next, his partner, the Responder, can decide whether she accepts or rejects his proposal. In case she accepted it, both are paid what was offered; if she rejected it, neither gets anything. Now, this being a game in extensive form, the focal solution concept to be employed is that of subgame perfect Nash-equilibrium, and accordingly the Proposer is to offer 99 units to himself which the Responder is to accept. This prescription is in sharp contrast with the corresponding empirical evidence: when this game is played only once, Proposers rarely offer more than two-thirds to themselves and Responders accept abusive offers even more rarely. Indeed, the fifty-fifty split is quite a frequent (albeit not a focal) outcome. How

can game theory deal with this apparent anomaly?

First of all, it is to be acknowledged that the Ultimatum Game fits the canonical model of non-cooperative games in extensive form: its players, the available moves, the possible outcomes are all identified. Nevertheless, the most common reaction to the empirical evidence cited above is to claim that the description of this game is incomplete. The shares to be gained from a split comprise only the pecuniary or "material" payoffs the players may attain, whereas their reluctance to deviate considerably from the fifty-fifty split suggests that they also care about the fairness of the eventual outcome. Hence it is natural to propose that they have preferences concerning the fairness of the outcome and that players care both about pecuniary and fairness payoffs in this game. So this is the proposal that mounts a challenge to the priority thesis. Nevertheless, the proposal needs to be completed by the articulation of a specific form the putative fairness preferences are to take, as no solution concept can operate without a full description of the objectives of the players in a game.

Of course, the problem posed by the Ultimatum Game may be approached in other ways as well. One could argue, for instance, that it is mistaken to take for granted the "refined" Nash-equilibrium solution concept that yields an counter-intuitive prescription for this game.

In his paper, "Incorporating Fairness into Game Theory and Economics" (1993), Matthew Rabin offers an analytical tool that both provides a suggestion for how fairness preferences are to be formulated and promotes a new solution concept for games where fairness considerations are supposed to play a role. It is accepted that the players in the Ultimatum Game try to earn as much money as possible, that is they seek their own material self-interest. But they also act so as to manage certain emotions that reflect their judgments concerning the motives of their counterparts. They are taken as seeking to be helpful to those who seem to be helpful to them, and seeking to hurt those who seem to hurt them. These latter objectives generate their preferences for fairness.

However, as I will argue, Rabin's approach fails to give an adequate account of how ordinary people play the Ultimatum Game. There are two main reasons for this. First, his new solution concept does not meet the demands issued by the Standard View. Second, his fairness payoffs are formulated in a rather *ad hoc* way.

**2.1. Rabin's Model.** Rabin's analysis accepts the canonical model for finite normal form games presented in the previous section, and he concentrates on two-payer games. Let us denote by  $a_i$  the actual strategy of player  $i$ , by  $b_j$  his belief about what player  $j$  will play, and by  $c_i$  his belief about what player  $j$  thinks that he will play. Then a pair of actions  $(a_1, a_2)$  forms a *fairness equilibrium* if two conditions hold:

- (i)  $a_i$  maximizes the overall utility function  $U_i(a_i, b_j, c_i)$ , for both players;
- (ii) and  $a_i = b_i = c_i$ , for both players.

The overall utility function is made up of two parts:

$$U_i(a_i, b_j, c_i) = \pi_i(a_i, b_j) + f_j(b_j, c_i)(1 + f_i(a_i, b_j)),$$

that is, it incorporates the material payoffs  $\pi_i(a_i, b_j)$ , and also fairness payoffs, taking a specific form. The function  $f_i$  stands for the assessment of player  $i$  about how kind he is to his partner, that

is about how much resentment he may generate given the choices  $a_i$  and  $b_j$ ; then the  $f_j$  function stands for how kind player  $i$  perceives his partner to be to himself, again given  $b_j$  and  $c_i$ . These functions also have an internal structure. In order to explicate what this structure is, one needs to introduce further notations. So:

- $\pi_j^h(b_j)$  is the highest payoff player  $j$  can get if he chose  $b_j$ ;
- $\pi_j^{min}(b_j)$  is the smallest payoff player  $j$  can get if he chose  $b_j$ ;
- $\pi_j^l(b_j)$  is the lowest Pareto-efficient payoff player  $j$  can get if he chose  $b_j$ ;
- $\pi_j^e(b_j)$  is the equitable payoff player  $j$  can get if he chose  $b_j$ , defined as the average of  $\pi_j^h(b_j)$  and  $\pi_j^l(b_j)$ .

Now:

$$f_i(a_i, b_j) = \frac{\pi_j(b_j, a_i) - \pi_j^e(b_j)}{\pi_j^h(b_j) - \pi_j^{min}(b_j)}, \text{ with } f_i(a_i, b_j) = 0 \text{ if } \pi_j^h(b_j) = \pi_j^{min}(b_j),$$

and

$$f_j(b_j, c_i) = \frac{\pi_i(c_i, b_j) - \pi_i^e(c_i)}{\pi_i^h(c_i) - \pi_i^{min}(c_i)}, \text{ with } f_j(b_j, c_i) = 0 \text{ if } \pi_i^h(c_i) = \pi_i^{min}(c_i).$$

Note that both  $f_i(a_i, b_j)$  and  $f_j(b_j, c_i)$  can take values only in the closed interval  $[-1, 0.5]$ ; and that given that the other is perceived to deserve resentment the fairness payoffs increase if one also gives rise to resentment, and given that the other is perceived as helpful one's fairness payoffs increase if one is more helpful to the other.

The overall utility function  $U_i$  describes then an amended game in which beliefs concerning what the other will do and beliefs concerning what the other believes about what one will do generate additional payoffs. Rabin therefore proposes to construct from ordinary normal form games psychological games, in the sense of Geanakoplos, Peirce, and Stachetti (1989). These authors define psychological games as games in which first- or higher order beliefs about the other yield payoffs for the players (I will return to this concept later in this section). In turn, his fairness equilibrium is a Nash-equilibrium for the amended game with the restriction that in equilibrium beliefs about the other have to be correct. So he indeed both makes a proposal for how to incorporate fairness preferences into game theory and offer a particular solution concept by means of which games so described are to be analyzed.

**2.2. Fairness equilibrium and the Ultimatum Game.** By accepting that no analysis of the Ultimatum Game can proceed without introducing preferences for fairness and by acknowledging also that Rabin does exactly that, one also expects that his approach can account for the evidence of how the Ultimatum Game is played in experiments.

However, his analysis does not yield respectable results. As his fairness equilibrium is suited to games in normal form only, we are to consider the following, somewhat modified Ultimatum Game. Players are to distribute  $X$  units among themselves. The Proposer is to decide about a split of this

sum, in the form of  $(x, X - x)$ , and the Responder is to make a decision about the level of  $x$  above which he will reject any offer (let us denote this level by  $r$ ); these choices are made simultaneously. There are very many Nash-equilibria of this game, their number depends on the unit in which offers have to be expressed, but they include  $(0.5X, 0.5X)$  and  $(0.6X, 0.6X)$  and every feasible split that falls between these two together with the readiness to accept that. Experiments in which this modified game are played do not show much difference from what happens in experiments when the standard Ultimatum Game is played. However, its fairness equilibrium is unique, and it is formed by the pair of actions  $x = X - (X/(2X + 1))$ , and  $r = x = X - (X/(2X + 1))$ , respectively. So if  $X$  was 100, the Proposer gets 99 plus  $100/201$ , and the Responder gets  $101/201$ , that is, the ensuing distribution is absolutely unfair. There are many reasons why the fairness equilibrium concept issues this result, and let me start with explicating the first and most important of these.

This reason can be best appreciated if we first attend to Rabin's analysis of a much simpler game, that is usually called the Battle of the Sexes. This game can be described by the payoff matrix below:

Table 2: A Battle of the Sexes game ( $X > 0$ )

	<b>Opera</b>	<b>Boxing</b>
<b>Opera</b>	$2X, X$	$0, 0$
<b>Boxing</b>	$0, 0$	$X, 2X$

Let us abbreviate *Opera* by  $O$  and *Boxing* by  $B$ . First we are to note that playing  $(Opera, Boxing)$  may be a fairness equilibrium. This is because, first,  $f_1(O, B) = -1$ ,  $f_2(O, B) = -1$ , and second,  $f_1(B, B) = 0$ . Accordingly,  $U_1(O, B) = 0$  and  $U_2(B, B) = X - 1$ , which means that Player 1 has no incentive to deviate to *Boxing* if  $X < 1$ . We can also check that as  $f_2(O, O) = 0$ , we get  $U_2(B, O) = 0$  and  $U_2(O, O) = X - 1$ ; that is player 2 also lacks incentive to deviate from the proposed equilibrium action if  $X < 1$ . Therefore if  $X < 1$ ,  $(Opera, Boxing)$  is indeed a fairness equilibrium, whereas it is certainly not a Nash-equilibrium (as  $2X > 0$  and  $X > 0$ ).

Why would player 1 consider playing *Opera* when he believes that player 2 plays *Boxing*, and that player 2 believes that he will play *Opera*? In case of playing *Opera* he gets material payoffs at the level of 0, but he also gets fairness payoffs, albeit also at the level of 0. This is because he resents that his partner to choose *Boxing* when she believes that he is to opt for *Opera*, and he also recognizes that the other resents him for his *Opera* choice that keeps her, player 2, from getting the material payoffs of  $2X$ . Could he improve of his situation by choosing *Boxing* instead? Now in that case he would get material payoffs of  $X$  and his partner would stop resenting him as she accesses her  $2X$  in this case. However, he, player 1, still resents her for keeping him from getting  $2X$  ( $f_2(O, B) = -1$  still), and indeed as much as if he intended to respond with *Opera* to *Boxing*! So in case of deviating to *Boxing*, player 1 gets fairness payoffs of  $-1$ , as he responds with kindness to something that is to be resented. But what sustains resentment, and at the same level, when now player 1 is to choose the best response to the other playing *Boxing*? In his discussion,

Rabin argues that the sustained resentment is due to the fact that by choosing *Boxing*, player 2 keeps player 1 from getting  $2X$ , his highest potential material payoff. That is, he considers that she chooses *Boxing* despite of he wishing to do *Opera*, so he is to punish her because of her motives. However, player 1 never gets to know the motives of player 2: he only considers that she will play *Boxing* and tests the respective gains from the two possible responses to that.

In order to see better the absurdity of the  $(Opera, Boxing)$  fairness equilibrium, let us examine why  $(Boxing, Opera)$  may be a fairness equilibrium as well. The relevant resentment values in this case are  $f_1(B, O) = -1$ ,  $f_2(B, O) = -1$ , and  $f_1(O, O) = 0$ , respectively. It is easy to check that if  $X < 1/2$ , then this pairs forms a fairness equilibrium. Let us trace the reasoning of player 1 that leads to him playing *Boxing* now. If he believes his partner will play *Opera* and also believes that he will play *Boxing*, then by indeed doing *Boxing* he gets material payoffs of 0, and fairness payoffs of 0, as well. Now shall he deviate to play *Opera*? In that case he gathers material payoffs  $2X$  and the resentment to be felt against him will change to 0 (as  $f_1(O, O) = 0$ ). However, his resentment towards player 2 is still the same, namely of the value  $-1$ . This is because, as Rabin would explain, player 1 resents that player 2 wants to play *Opera* when he wanted to play *Boxing* and therewith he is deprived of material payoffs of  $X$ . But, even though by responding with *Opera* to *Opera* he would get the highest possible payoffs in this game ( $2X$ ), this will not be enough to induce him to do so, because the overall resentment he feels is even larger than in the case when he wanted to play *Boxing* ( $-1$  instead of 0).

Now one can attend to the task of showing why  $(0.5X, 0.5X)$  is not a fairness equilibrium in the modified Ultimatum Game. In general, this can also be traced back to counterfactual assessments on the part of the Responder of the fairness of his counterpart.

**2.3. Note on psychological games.** The above analysis can rightly be judged as somewhat muddled. We can sharpen it if we reflect on the relationship between the concepts of fairness equilibrium employed in it and of the Nash-equilibrium. In order to test whether an action may be part of a Nash-equilibrium, a player is to suppose first that his partner plays her part of the Nash-equilibrium. Then he is to consider whether he could get more by deviating from his part of the Nash-equilibrium or not, and while employing this test he is to sustain his hypothesis about what the other will do. So if indeed his part in the Nash-equilibrium so tested brings him the most, he will do the best in case his hypothesis will turn out to be true. Now, in the context of fairness equilibrium, part of what player 1's counterpart does is forming a belief about what player 1 will do. So in a proposed equilibrium she is supposed to make a move (*Boxing* above), and is supposed to believe that player 1 will do his part of the equilibrium (that is *Opera*), further she is supposed to be right in thinking that player 1 thinks that she will do *Boxing*. So when player 1 considers a deviation from his proposed move, then he has to sustain his hypothesis about the beliefs of player 2. This is why he considers, according to Rabin, that by choosing *Boxing* she keeps him from getting material payoffs of  $2X$ . But player 1 has no basis for the belief that she thinks that he will choose *Opera*, indeed this belief is tested by considering the switch to *Boxing* on his part. But then why resent her with the same intensity in case of a deviation to *Boxing*? In an equilibrium, beliefs happen to coincide, but this eventual coincidence ought not to be taken for granted while

deliberating for one's choice. Indeed, it is not taken for granted when a player is about to decide whether a certain action is part of a Nash-equilibrium or not.

[Psychological games, psychological equilibrium, counterfactual assessments, deliberation]

### 3. CONCLUDING REMARKS

There are other works that use some version of Rabin's doctrine of fairness in order to account for phenomena like the usual behavior in the Ultimatum Game (see Falk-Ursbacher (2001) and the references therein). However, as these all employ the concept of psychological equilibrium, they are open to the same criticism as Rabin's model itself. One can list two more kinds of strategies by means of which fairness motivated actions could be incorporated into game theory and economics. The first of which is to derive actual preferences of agents in distribution problems using an axiomatic approach. In this vein, Karni and Safra (2002) study how to combine self-interested and fairness preferences into one actual preference pattern. I note that they also cannot exclude that agents in an Ultimatum Game would arrange a radically unequal sharing of the surplus. But it is more urgent to state that their work is vulnerable to the criticism that their characterization of fairness is no more than *ad hoc*. On the other hand, calibration theories are not to succeed either, as general revealed preferences arguments could show (Osband-Green, 1991).

But, at the end of the day, we may also decide that game theory was really not conceived for the study of moral phenomena. It was meant to be a tool for the analysis of strategic situations.

#### REFERENCES

- [1] BINMORE, KEN (1994): *Playing Fair*, Cambridge: MIT Press;
- [2] FALK, ARMIN - ERNST FEHR (2002): "Psychological Foundations of Incentives", mimeo., Center for Economic Studies Munich (forthcoming in *European Economic Review*);
- [3] FALK, ARMIN - URS FISCHBACHER (2001), "A Theory of Reciprocity", mimeo., Center for Economic Studies Munich;
- [4] GEANAKOPOLOS, JOHN - DAVID PEARCE-ENNIO STACCHETTI (1989): "Psychological Games and Sequential Rationality", *Games and Economic Behavior* 1.1: 60-79;
- [5] GILBOA, ITZAK - DAVID SCHMEIDLER (1988): "Information Dependent Games", *Economics Letters* 27: 215-221;
- [6] GREEN, EDWARD - KENT OSBAND (1991): "A Revealed Preference Theory for Expected Utility", *Review of Economic Studies* 58.4: 677-696;
- [7] KARNI, EDI - ZVI SAFRA (2002): "Individual Sense of Justice: A Utility Representation", *Econometrica* 70.1: 263-284;
- [8] OSBORNE, MARTIN - ARIEL RUBINSTEIN (1994): *A Course in Game Theory*, Cambridge: MIT Press;
- [9] RABIN, MATTHEW (1993): "Incorporating Fairness into Game Theory and Economics", *American Economic Review* 83.5: 1281-1302;