Mental Equilibrium and Rational Emotions

Eyal Winter, Ignacio Garcia-Jurado, Jose Mendez-Naya Luciano Mendez-Naya, [1][2]

January 7, 2009

Abstract

We introduce emotions into an equilibrium notion. In a mental equilibrium each player "selects" an emotional state which determines the player's preferences over the outcomes of the game. These preferences typically differ from the players' material preferences. The emotional states interact to play a Nash equilibrium and in addition each player's emotional state must be a best response (with respect to material preferences) to the emotional states of the others. We first discuss the concept behind the definition of mental equilibrium and show that this behavioral equilibrium notion organizes quite well the results of some of the most popular experiments in the literature of experimental economics. We expose some attractive properties of mental equilibria which are useful for deriving the set of mental equilibria for specific games.
Keywords: Games, Equilibrium, Behavioral Economics, Emotions

1. Introduction

The tension between rational behavior as predicted by a variety of game-theoretic models and experimental results has been the focus of attention of both game theorists and experimental economists. There are two sources of rationality incompleteness that are responsible for many of the discrepancies between experimental observations and game-theoretic predictions. The first source of discrepancy arises from the fact that many strategic interactions are too complex for subjects in the lab (or outside the lab) to analyze. For example, subjects typically fail to realize that in a second-price auction it is a dominant strategy to bid the true valuation and choose an inferior strategy. Likewise, they typically fail to perform backward induction in games with more than two stages. The second source of discrepancy has little to do with complexity. While understanding the strategic considerations perfectly, players fail to maximize their own monetary rewards simply because the way they value the different outcomes of the game may be inconsistent with the maximization of material rewards. Games like ultimatum bargaining, the dictator game, and the trust game are well-known examples of this sort. Over the last decade several interesting and important models have been developed that try to reconcile the discrepancy between experimental results and

[2] Winter: Department of Economics, and Center for the Study of Rationality, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel; Garcia-Jurado: University of Santiago de Compostela; Mendez-Naya: University of Santiago de Compostela; Mendez-Naya: University of Santiago de Compostela.
[2] I wish to thank Ken Bimore, Werner Gueth and Reinhard Selten for their comments and suggestions on an earlier draft of this paper.

game-theoretic predictions, without neglecting the idea that players behave strategically. The common idea in these papers is to reevaluate the outcomes of the game for each player, taking into account inequality aversion, spitefulness, and envy, so that in the new set of utility functions the equilibrium behavior is closer to the experimental observations (see Fehr and Schmidt (1999), and Bolton and Ochenfels (2000)). The main challenge of this strand of literature is to identify the set of parameters that best explains the experimental results and use these parameters to understand players' motives in the underlying games. A somewhat different approach was proposed by Rabin (1993) with the concept of fairness equilibrium. Here the material payoffs are also altered to incorporate fairness into the utility function. The measure of fairness depends on players' actions and beliefs, which are determined in equilibrium.

In this paper we attempt to take a more general approach by recognizing the fact that different strategic environments can give rise to different types of immaterial preferences that may represent fairness or inequality aversion but also envy, spite, and a variety of other emotions. We will use the term *mental state* to represent these preferences. We will introduce an equilibrium concept called *mental equilibrium*, which is behavioral in its nature and and which seems to organize the experimental evidence for some of the most prominent examples much better than the standard concepts of Nash equilibrium or subgame perfect equilibrium. As a by-product of our analysis we will be able to derive players' behavioral preferences endogenously through the equilibrium conditions.

In simple words the concept of mental equilibrium can be described as follows. Each player, whom we assume seek to maximize only with his material/monetary payoffs, is assigned a mental state. A *mental state* is simply a utility function over the outcomes of the game (i.e., the set of strategy profiles) which is typically different from the material utility function. A strategy profile s of the game is said to be a *mental equilibrium* if two conditions hold: firstly, s has to be a Nash equilibrium in the game among the players' mental states. Secondly, no player can expect a better equilibrium outcome, with respect to his selfish material payoffs, by unilaterally changing his mental state. There are two valid interpretations of our equilibrium concept. The first interpretation involves the idea of the evolution of norms and emotions. Essential to our model is the fact that the benchmark preferences of a player are selfish and material. It is conceivable to assume that fairness, anger, envy, and revenge, which play a role in many game situations, have been developed through evolution to increase the fitness of individuals to the social environment in which they live. Our equilibrium concept can be viewed as promoting this idea. We are not proposing any specific evolutionary model to this effect, but conceptually one can view the emergence of mental states as part of such an evolutionary game. Evolutionary selection reinforces different mental states in different strategic environments, and the material payoff of the game can be viewed as measuring fitness or the degree of replication. This interpretation is in line with the indirect evolutionary approach proposed by Gueth and Yaari (1992).

The second interpretation of our equilibrium concept is that of rational emotions. In strategic environments individuals may decide to be in a certain emotional state that serves their interest. Emotional states are often induced through cognitive reasoning in full or partial awareness and are used as a commitment device. In order for the commitment to be credible, the emotional state has to be genuine and not feigned. To further explain this

point we suggest two thought experiments that demonstrate how emotions are triggered by incentives. Imagine that you are informed at the airport that your flight has been canceled and that you should report at the airline desk the next day. Consider the following two scenarios: in scenario A you observe most of the passengers leaving the terminal quietly. In scenario B you run across an acquaintance who tells you that he was rerouted to a different flight after explaining to the airline employees, in a very assertive and determined manner, that he has to arrive at his destination today. If you decide to go to the desk and request a similar treatment you are most likely to find yourself in a very different emotional state from the one you would have been in in scenario A. You are likely to exhibit signs of anger quite quickly in scenario B; in fact, these won't be only signs, you will actually be angry. You have been offered incentives to be angry and as a consequence you "choose" to be angry. The second example involves a real story. In 2006 the first co-author of this paper was interviewed for the Arab TV channel Al Jazirah. In accepting the Al Jazirah request for an interview Winter was hoping to generate sympathy for Israel and Israelis among the Arab viewership. At some stage of the interview the interviewers moved to questions about Winter's family background. Winter then told the story about his father's hardships in escaping Nazi Germany and trying to find shelter in Palestine of the 1930s. Winter remembers telling this story to friends and relatives already dozens of times in the past but it was only in front of the Al Jazirah camera that he felt so emotionally affected that he could not hold back his tears. Again Winter was offered an incentive to be emotional (in terms of his objectives) and emotional he was. The two stories told above suggest that in certain environments mental states can be thought of as outcomes of a cognitive choice. We refer the reader to an experimental testing of rational emotions by Winter et al. (2008). Taking this view one can think of mental equilibrium as an equilibrium in an amended game of credible commitments. The material payoffs here are standard payoffs in a game and not a measure of evolutionary fitness. We will be subscribing to both interpretations and will not argue in favor of one of them as we believe that different emotional reaction may fit different interpretation. In particular, for robust emotions which are irresponsive to the specification of strategic environment or to the extent of which players can see each other the evolutionary approach should favored (most "blind" experiments fall under this category). For others, which rely on mutual eye contact and strongly responsive to incentives the interpretation of rational emotions seem more appropriate. In both cases however we view emotions as a mechanism to promote self interests.

Our concept of mental equilibrium can also be viewed as a model of endogenous preferences. Players in our model select their preferences in view of their beliefs about the preferences of those with whom they interact. The remarkable feature of this concept is that while the choice of preferences is done from a self-centered point of view, the equilibrium choice of preferences may give rise to nontrivial social preferences in which the players' behavior is very far from that of a self-centered player. Indeed, in some of our examples we will restrict the set of mental states to have Fher and Schmidt (1999)-type preferences representing inequality aversion and we will be able to derive endogenously conditions on the parameters of inequality aversion that mental states must exhibit in equilibrium.

Another assumption that is rather crucial to the definition of mental equilibrium is that players' mental states are commonly known. This is a critical assumption when trying to

answer the question of how a mental equilibrium emerges. It is of lesser importance if we treat the concept of mental equilibrium as a static stability concept (as is the case with Nash equilibrium). Nevertheless, there are two tracks by which this assumption can be justified. In viewing emotions as a commitment device, it is clear that they cannot be effective unless they are observable. Indeed, body movements, facial expressions, and voice intonations are often very clear signals of emotional reactions. But even without direct eye contact players may still form consistent beliefs about the mental states of their counterparts. Through experiencing identical or similar strategic environments over and over again players can learn quite a bit about the function that maps strategic environments to players' mental states. In terms of our airport example, it is fair to assume that even a moderately socially intelligent employee who has dealt with hundreds or thousands of customers in similar situations can expect that in scenario B a passenger might be angry.

It is instructive to demonstrate the operation of the concept of mental equilibrium on the famous example of social dilemma (the n-person prisoner's dilemma game). A group of players face a game in which each player has two strategies: cooperate or defect. Each player has a dominant strategy which is to defect (i.e., this is the unique Nash equilibrium), but the maximum a player can earn is when every other player cooperates. Experimental results typically show a substantial level of cooperation in such a game. When cooperators are asked about their motives they often react by saying: "If I come with a selfish attitude that favors defection, then everybody else will defect as well and we will all lose." Note that an argument of this sort speaks in favor of social preferences as a means of promoting one's selfish goals. Of course the argument is invalidated by the fact that the game is played simultaneously. However, in spite the fact that players' action in the game are unobservable, it is often quite hard to convince a subject that his own action cannot affect that of the other players. Let us examine now how the concept of mental equilibrium relates to this game. We will concentrate on one equilibrium that results in the cooperation of all players. Suppose that each player arrives at the game with a mental state such that he prefers the outcome of total cooperation (i.e., cooperation by all players) to all other outcomes, but whenever any of the players defect he prefers to defect as well. With this set of mental states the profile in which all mental states cooperate is a Nash equilibrium. Furthermore, no single player can assign a different mental state for which a new equilibrium will arise that he prefers to total cooperation. The only relevant deviation to check here is for a mental state that always prefers to defect (i.e., the mental state whose preferences are identical to those of the selfish player). But given such a deviation and the set of other mental states, the only equilibrium outcome is for all mental states to defect, which makes the player worse off based on his selfish preferences. We will later characterize the set of all mental equilibria of this game.

The relevance and importance of our concept can be judged by two criteria: firstly, the extent to which the story behind the concept is appealing and makes sense and, secondly, the extent to which the concept is capable of explaining puzzling experimental results, particularly those at odds with standard game-theoretic concepts such as Nash equilibria or subgame perfect equilibria. Our attempt to convince the reader of the importance of our concept will address both criteria. In particular, we will introduce a battery of well-known games about which considerable experimental data has been collected and we will

compare the set of Nash equilibria to the set of mental equilibria. As we shall show, every pure Nash equilibrium is also a mental equilibrium; however, interestingly, the outcomes that emerge in experimental results of the games considered here very often correspond to mental equilibria that are not Nash equilibria.

The evolutionary foundation of preferences has been discussed in the economic literature in the past. Gueth and Yaari (1992) introduce a game of cooperation between two players and show how preferences for reciprocity (which in their model boils down to be the value of a parameter in the utility function) can emerge through evolution (see also Gueth and Kliemt (1999)). This approach, known as the indirect evolutionary approach, has also been used recently by Dekel, Ely and Yilankaya (2007), who develop a more general model than that of Gueth and Yaari (1992). They consider the class of all two-person games and interpret their payoffs as objective measures of fitness. They then endow players with subjective preferences over outcomes according to which they assume players play Nash equilibria. To select for the "optimal preferences," they impose evolutionary conditions (of selection and mutation).

Because the evolutionary story is only one interpretation of our model, we do not specify evolutionary conditions for stability. Instead, our model builds on two-level games. One level involves the mental game in which the payoffs are derived from players' mental states (emotions) and the other level involves rational players who select their state of mind. At each of these levels agents are assumed to play Nash equilibria. As a consequence of the fact that the Nash equilibrium conditions for the rational level game are less stringent[3] than Dekel et al.'s (2007) evolutionary conditions, our set of mental equilibria is typically larger than the set of stable outcomes a la Dekel et al. (2007), and our model admits a (pure) mental equilibria for any game. Furthermore, the simplicity of our structure allows us to analyze mental equilibria for games with an arbitrary number of players (and not just two-person games).

Finally, our paper departs from the existing literature also in its motivation. We are not merely interested in the theoretical properties of mental equilibria. We pay substantial attention to experimental results and use them to examine the performance of our concept. Several other papers use the indirect evolutionary approach in specific economic environments, such as Bergman and Bergman (2000) in the context of bargaining, Gueth and Ockenfels (2001) in the context of legal institutions, and Heifetz and Fershtman (2006) in the context of elections and political competition.

We continue in Section 2 with the formal definition of mental equilibrium. By this definition we assume that mental states play pure Nash equilibria. We then provide a useful characterization of mental equilibria in two-person games, which we then apply to specific examples. In Section 3 we define the "subgame perfect version" of mental equilibrium and compare it with the standard notion of subgame perfect equilibrium in the ultimatum bargaining game, partly by confining our attention to mental states whose preferences represent inequality aversion.

Section 3 deals with mental equilibria for more than two players. We show that for

[3] The sufficient conditions for an outcome to be stable in two-person games in the model of Dekel et al. is that the outcome be both a strict Nash equilibrium and efficent.

games with four or more players the notion of mental equilibria loses its predictive power, since any strategy profile in such games is a mental equilibrium. This follows from the fact that for some choices of mental states by the players the corresponding mental game may possess no pure Nash equilibrium. Following this result we show that the set of mental equilibria is nonempty for any n-person game.

To tackle the fact that the notion of mental equilibria loses its predictive power when n is at least four players, we introduce the concept of *mixed mental equilibrium* where we allow mental states to use mixed strategies (the choice of mental states by players is still assumed to be pure). We show that the awkward property of pure mental equilibria for games with at least four players ceases to exist for the new concept, and we use this concept to demonstrate its properties on some games. In particular, we demonstrate that mixed mental equilibrium need not be a Nash equilibrium. The question of a general existence result for this concept is unfortunately left unsolved[4].

2. Basic Definitions

Let $G = (N, S, U)$ be a normal form game where N is the set of players, $S = S_1 \times S_2, ..., \times S_n$ is the set of strategy profiles for the players, and $U = U_1, ..., U_n$ are the players' utility functions over strategy profiles. We refer to U_i as the benchmark (selfish/material) utility function of the players and will use u_i to represent the mental states' utility functions. A profile of mental states is denoted by $u = u_1, ..., u_n$. For a given game G we denote by $NE(G)$ the set of (pure) Nash equilibria of the game G.

Definition: A *mental equilibrium* of the game $G = (N, S, U)$ is a strategy profile s such that for some profile of mental states u the following two conditions are satisfied:

(1) $s \in NE(N, S, u)$.

(2) There exists no player i, a mental state u_i', and a strategy profile $s' \in NE(N, S, u_i', u_{-i})$ with $U_i(s') > U_i(s)$.

Observation 1: Any pure strategy Nash equilibrium s of a game is also a mental equilibrium. To see that this is the case, choose for each player a mental state whose payoff is such that s_j is a strictly dominant strategy in the game. Clearly s is an equilibrium in the mental game. Suppose that player i assigns a different mental state. Clearly in the new mental game all other players will stick to their dominant strategy. Since s_i is a best response to s_{-i} with respect to player i's material preferences (since s is a Nash equilibrium), player i cannot be any better off by assigning a different mental state.

3. Two-Person Games

[4] We have devoted much time to tackling the issue of existence. We managed to prove existence under several sets of conditions but decided that none of them is general enough to merit a result. The main stumbling block is the fact that the mapping from games to the set of delegate equilibria is not continuous in any form of definition that will allow us to use a fixed point theorem. We are grateful to Sergiu Hart, who has helped us on this issue. The example showing that a mixed delegate equilibrium may not be a Nash equilibrium is due to him. Recently Olschewski and Swiatczak (2008) proved the existence of a mental equilibrium in a 2x2 games.

In any Nash equilibrium each player attains at least his maxmin value. Proposition 1 asserts that this property characterizes the set of mental equilibria for every two-person game.

Proposition 1: Let G be a two-person game; then $s \in S$ is a mental equilibrium if and only if $U_i(s) \geq \max_{s_i} \min_{s_j} U_i(s_i, s_j)$ where $i = (1, 2)$ and $i \neq j$.

We show in the Appendix that Proposition 1 does not apply to three-person games and in fact neither of the two directions of the proposition holds true.

Proof of Proposition 1: Let v_1 and v_2 be the maxmin values of players 1 and 2 respectively. We first show that any mental equilibrium must yield each player at least v_i. Assume by way of contradiction that there is a mental equilibrium s such that at least one of the players, say player 1, earns less than v_1. Suppose that s is supported as a mental equilibrium with the mental states u_1 and u_2 respectively. If instead of u_1 player 1 deviates and chooses the mental state u_1' under which playing s_1 is a dominant strategy, then in the resulting mental game (u_1', u_2) there exists a pure Nash equilibrium and all equilibria yield a payoff of at least v_1 for player 1. This contradicts the assumption that s is a mental equilibrium, and proves one direction. We next argue that every profile yielding at least the maxmin value for the two players is a mental equilibrium. For this we construct the following mental game: Let $s = (s_1, s_2)$ be a profile that yields each of the two players at least his/her maxmin value. For the mental state of player 1 we set $u_1(s) = 1$, and $u_1(s_1', s_2) = 0$ for all $s_1' \neq s_1$. Furthermore, for every $s_2' \neq s_2$ there exists s_1' such that $U_2(s_1', s_2') \leq U_2(s)$. Otherwise the maxmin value of player 2 is greater than $U_2(s)$, which contradicts the definition of s. We now set $u_1(s_1', s_2') = 1$ and $u_1(s_1^*, s_2') = 0$ for all $s_1^* \neq s_1'$. We now define the mental state of player 2 in a similar manner: $u_2(s) = 1$, and $u_2(s_1, s_2') = 0$ for all $s_2' \neq s_2$. Furthermore, for every $s_1' \neq s_1$ there exists s_2' with $U_1(s_1', s_2') \leq U_1(s)$; otherwise the maxmin value of player 1 must be greater than $U_1(s)$, which is impossible. We now have $u_2(s_1', s_2') = 1$ and $u_2(s_1', s_2^*) = 0$ for all $s_2^* \neq s_2'$. We can now show that s is a mental equilibrium of the game supported by $u_1$and u_2. Indeed, s is clearly a Nash equilibrium under $u_1$and u_2, as the mental game never has a payoff of more than 1 for either player. To show that condition (2) in the definition of mental equilibrium applies, note that if, say, player 1 changes his mental state to u_1', then a Nash equilibrium of the new mental game (u_1', u_2) must involve a strategy profile s' such that $u_2(s') = 1$. Otherwise the mental state of player 2 will deviate. But for such s' we must have $U_1(s') \leq U_1(s)$, which implies that player 1 cannot make himself better off by changing his mental state. The same argument applies for player 2 and we conclude that s must be a mental equilibrium.

Proposition 2: For every two-person game G there exists a pure strategy profile that pays each player at least his/her maxmin value.

Proof: We prove the proposition by induction on the number of strategies of players 1 and 2 in the game. The assertion is trivial if one of the players has only one strategy. Assume now by induction that the statement is true if at least one player, say player 1, has less than m strategies, and consider now a game with m strategies for player 1. Denote this game by G and denote by G_{-m} the game obtained by eliminating the last strategy of player 1. By the induction hypothesis, G_{-m} is a game for which Proposition 2 applies. Let $s = (s_1, s_2)$ be the strategy profile in which both players obtain their maxmin value for

the game G_{-m}. If s pays each player at least his maxmin value in G then we are done. Otherwise, it must be that the maxmin value of player1 in G is greater than $U_1(s)$. This follows from the fact that player 2's maxmin value in G_{-m} is at least as large as that of G (since the game G adds a strategy for player 1 and not for player 2). Since the maxmin value of player1 in G is greater than $U_1(s)$ we must have $U_1(m, s_2') > U_1(s)$ for all $s_2' \in S_2$. Let $s_2^* = \arg\max_{s_2 \in S_2} U_2(m, s_2)$. Clearly (m, s_2^*) is a profile that pays both players at least their maxmin value in G.

Propositions 1 and 2 immediately imply an existence result for two-person games:

Corollary 1: Every two-person game possesses a mental equilibrium.

Our definition of mental equilibrium relied on the assumption that players are optimistic when contemplating deviations as it is enough that there exists at least one equilibrium in the new mental game (after player i deviates) that player i prefers to the original (putative) equilibrium in order to trigger him to deviate. A more stringent condition on deviations would require that player i deviate only if all equilibria of the new game yield player i a higher utility. Since the conditions for deviations are stronger, this equilibrium notion is weaker than the standard one. Formally:

Definition: A *weak mental equilibrium* of the game $G = (N, S, U)$ is a strategy profile s such that for some profile of mental states u the following two conditions are satisfied:

(1) $s \in NE(N, S, u)$.

(2) There exists no player i, and a mental state u_i' such that for every equilibrium, $s' \in NE(N, S, u_i', u_{-i})$ with $U_i(s') > U_i(s)$.

Clearly every mental equilibrium is a weak mental equilibrium but we shall argue that:

Proposition 3: In two-person games the set of mental equilibria and the set of weak mental equilibria coincide.

Proof: We have shown that the set of mental equilibria coincides with the set of all strategy profiles that award each player at least his/her maxmin value. It is therefore enough to show that any strategy profile that pays some player less than his/her maxmin value cannot be a weak mental equilibrium. Indeed, suppose by way of contradiction that for some profile s some player, say, player 1, gets a payoff x_1 which is less than his/her maxmin value, and that s is a weak mental equilibrium supported by the mental states $u = (u_1, u_2)$. Let s_1 be the maxmin strategy of player 1. Consider a mental state u_1' under which s_1 is a dominant strategy for player 1. Consider now the mental game $(\{1, 2\}, S, (u_1' \ u_2))$. All Nash equilibria of this game involve player 1 playing s_1 . Hence, player 1 gets at least his/her maxmin value (in the game $G = (N, S, U)$), but this contradicts the fact that s is a weak mental equilibrium since player 1 is better off deviating under the condition imposed by the definit ion of weak mental equilibrium.

A large body of experimental results has been obtained for two-person games. Proposition 1 serves us with a very useful tool for identifying the set of mental equilibria for such games. We will now discuss some of the most prominent example of these games.

Example 1 The prisoner's dilemma

We consider the game given by the matrix below. This is the prisoner's dilemma game with a unique Nash equilibrium using dominant strategies (D,D).

	D	C
D	$1,1$	$5,0$
C	$0,5$	$4,4$

Observation 2: There are two mental equilibria in the prisoner's dilemma game, (C,C) and (D,D).

Proof: Players 1 and 2 can guarantee that the other player gets no more than 1 by playing the strategy D. Using Proposition 1 this means that (1,1) is a mental equilibrium. Since (4,4) dominates (1,1) it is also a mental equilibrium. To show that (5,0) and (0,5) are not a mental equilibrium note that no player can guarantee a reduction of the payoff of his opponent to zero because the dominant strategy of a player always guarantees a payoff of 1.

An example of mental states that sustain the cooperative outcome is $u_1(C,D) = u_2(C,D) = u_1(D,C) = u_2(D,C) = -1$, and $u_i = U_i$ otherwise. These mental preferences represent aversion to lack of reciprocal behavior.

It is easy to verify that a necessary and sufficient conditions for the mental states (u_1, u_2) to sustain (c, c) as mental equilibrium in the prisoner's dilemma is : $u_1(C,C) \geq u_1(D,C)$, $u_1(D,D) \geq u_1(C,D)$ and $u_2(C,C) \geq u_2(C,D)$, $u_2(D,D) \geq u_2(D,C)$.

It is worthwhile mentioning that the mental preferences sustaining the cooperative outcome cannot be of the form $u_i = \alpha U_i + \beta U_j$. Such a utility function would give rise to the following mental game:

	D	C
D	$\alpha+\beta$,$\alpha+\beta$	$\alpha 5, \beta 5$
C	$\beta 5, \alpha 5$	$4(\alpha+\beta), 4(\alpha+\beta)$

For (C,C) to be an equilibrium in this mental game we need to have $4(\alpha+\beta) \geq 5\alpha$. But this means that player 1 by sending a different mental state with $u_1 = U_1$ will be able to sustain (D,C) as an equilibrium since $5\beta \geq \alpha + \beta$.

Note the difference between the social preferences given by $u_i = \alpha U_i + \beta U_j$ and the one we used in Observation 2. The former represents a mental state with some degree of altruism (if $\beta > 0$) or spitefulness (if $\beta < 0$). Indeed, other cardinal representation of the prisoner's dilemma may admit α and β such that cooperation is sustainable as a mental equilibrium with players' mental states being $u_i = \alpha U_i + \beta U_j$, but there are other representations for which no such α and β exist. In contrast, the mental preferences that we used to sustain (C,C) represent mental states with aversion to lack of reciprocity and they sustain (C,C) regardless of the cardinal representation of the prisoner's dilemma game. We conclude that aversion to lack of reciprocity can explain cooperation in every prisoner's dilemma game but altruism or spitefulness cannot.

Example 2: The chicken game

Consider the following two-person game:

	retreat	fight
retreat	$1,1$	$-2,2$
fight	$2,-2$	$-10,-10$

Observation 3: The game has three mental equilibria: the two Nash equilibria with the outcomes (-2,2) and (2,-2) and another one which is the outcome (1,1). This can be easily verified by using Proposition 1 and noting that the maxmin value for both players is -2.

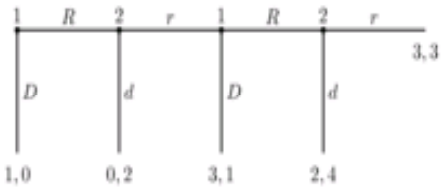

Figure 1:

Rapoport, Guyer and Gordon (1976) have established experimental results for the chicken game by varying the payoff from (fight, fight). For this particular game they observe an 87% probability of retreat and 13% probability of fight. So the mental equilibria that is not a Nash equilibrium is played with probability 75.6% more than the frequency of the two Nash equilibria together. Substantial proportions of retreats have also been established for much lower disutility levels from (fight, fight).

Example 3: The trust game

Player 1 has an endowment of x Euros. He can make a transfer $0 \leq y \leq x$ to player 2. If player 1 makes the transfer y, player 2 receives $3y$. Player 2 can now reward player 1 with a transfer of $z \leq 3y$. Finally, the payoff for player 1 is $x - y + 3z$ and the payoff for player 2 is $3y - z$.

Observation 4: An outcome (a_1, a_2) is a mental equilibrium outcome if and only if $a_1 \geq x$ and $a_2 \geq 0$.

Proof: Consider such an outcome (a_1, a_2). Since $a_1 \geq x$ player 2 can guarantee that player 1 gets no more than a_1. This can be done by transferring no money back to player 1 if player 2 received any money from player 1. Furthermore, clearly player 1 can guarantee that player 2 receives no more than zero by simply making a zero transfer to player 2. In view of Proposition 1, (a_1, a_2) is an mental equilibrium outcome. Consider a mental equilibrium outcome (a_1, a_2) such that either $a_1 < x$ or $a_2 < 0$. Then either player 1 or player 2 get less than their maxmin value, which contradicts Proposition 1.

We note that the trust game has a unique Nash equilibrium in which player 1 makes a zero transfer to player 2. Observation 4 suggests that any level of trust displayed by player 1 coupled with a level of trustworthiness that compensates player 2 to at least the level of his initial endowment can be supported by mental equilibria. We point out that experimental results support a considerable level of trust by player 1 and a considerable reciprocity by player 2 (see for example Cox et al. (1995)). We shall return to this example by restricting the set of mental states to include only Fehr and Schmidt (1999)-type utility functions representing inequality aversion.

Example 4: The centipede game

Consider the extensive form game presented in Figure 1, which is an example of the centipede game.

We recall that the centipede game has a unique Nash equilibrium that results in player

1 choosing D in his first decision node. The set of mental equilibria is, however, larger:

Observation 5: Consider the game presented in Figure 1. All strategy profiles but the one leading to the payoff outcome (0,2) are mental equilibria.

Proof: The maxmin value of player 1 is 1 (achieved by choosing D at the first node) and it is zero for player 2 (player 1, by choosing D at his first node, can prevent player 2 from getting more than zero). By Proposition 1, (0,2) cannot be a mental equilibrium outcome because player 1 is getting less than his maxmin value. All other outcomes pay both players at least their maxmin value and are therefore mental equilibrium outcomes.

The above observation can be easily extended to a general centipede game. One might find it intriguing that the second branch of the centipede game can never be a mental equilibrium. The intuition is however quite straightforward. If players assign mental states for which this outcome is an equilibrium, then player 1 can deviate by assigning a different mental state with a utility function yielding an arbitrary large payoff for choosing D in the first round by which he will guarantee a higher payoff of 1 (instead of zero). The fact that mental equilibrium allows for outcomes in which players trust each other to move into the game instead of opting out immediately is consistent with experimental results (see McKelvey and Palfrey (1992)). Indeed, in these experimental results the second terminal node is also reached with some propensity; however, in a different study by Nagel and Fang Tang (1998) where the centipede game was played in its normal form, the strategy profile with the lowest propensity is either to exit at the second mode or to exist at the first node. Furthermore, in one out of the five sessions the propensity of the second terminal node is substantially lower than that of all other nodes. The second lowest is the first node, which is almost twice as frequent as the second.

We now discuss our concept of mental equilibrium in the context of another prominent game, the ultimatum bargaining game.

Example 5: The ultimatum game involves two players: Player 1 has an endowment 1 from which he has to make an offer to player 2. An offer is a number $0 \leq y \leq 1$.

Player 2 can either accept the offer or reject it. If player 2 accepts the offer player 1 receives $1 - y$ and player 2 receives y. If player 2 rejects the offer both players receive a payoff of zero. The subgame perfect equilibrium of the game predicts a zero offer by player 1, which is accepted by player 2. Experimental results (see Gueth et al. (1982)) have however shown substantial offers made by player 1 with the mode of the distribution of offers being 50:50. To discuss the concept of mental equilibrium for this game we first need to amend the concept to the framework of subgame perfection in extensive form games. The following is the most natural amendment:

Consider an n-person extensive form game $G = (N, T, U)$ with perfect information, where N is the set of players, T is the game form defined by a tree (using the standard definition of extensive form games), and $U = U_1, ..., U_n$ are payoff functions for players 1,2,...,n assigned to terminal nodes of the game. We denote by $SPE(G)$ the set of subgame perfect equilibria of the game G. We define the notion of mental subgame perfect equilibrium:

Definition[5]**:** A *mental subgame perfect equilibrium* of the game G is a strategy profile s of G such that for some profile of mental states u the following two conditions are satisfied:

(1) $s \in SPE(N, T, u)$.

(2) There exists no player i, a mental state u_i', and a strategy profile $s' \in SPE(N, T, u_i', u_{-i})$ with $U_i(s') > U_i(s)$.

Unfortunately, when the set of mental states from which players can choose is not restricted, mental SPE loses much of its predictive power:

Observation 6: Take any extensive form game with perfect information G. Every Nash equilibrium of G is a mental SPE of G.

Proof: Consider the normal form version of G, and let s be the Nash equilibrium. For each player i choose a mental state u_i for which the strategy s_i is a strictly dominant strategy in the game. Clearly s is a subgame perfect equilibrium in the mental game. To see that condition 2 is satisfied assume that some player j can assign a different mental state u_j' and generate an SPE in the new mental game in which his payoff is higher. Since the Nash strategies are dominant for all mental states, no mental state other than j will choose any other strategy in the new mental game. Let s_j' be the strategy taken by $j's$ new mental state. We have assumed that $U_j(s_j', s_{-j}) > U_j(s)$ but this cannot happen since s is a Nash equilibrium.

4. Restricting the Set of Mental States

Observation 6 implies that without restricting the set of mental states all allocations of the unit of goods between the two players are sustainable as a mental SPE of the ultimatum game, since the set of Nash equilibrium outcomes covers the entire set of allocations. We now wish to confine our attention to mental states that display inequality aversion as characterized by Fher and Schmidt's (1999) model. We will start with the ultimatum game and then explore mental equilibria in this framework for other games. This analysis will contribute to the heated debate conducted in the early nineties over the role of fairness in ultimatum games and games in general. To recall, in a two-person game each mental state of player i has a utility function $u_i(x_i, x_j)$ over the allocations (x_i, x_j), which is of the following form: $u_i(x_i, x_j) = x_i - \alpha_i(x_j - x_i)^+ - \beta_i(x_i - x_j)^+$, where $z^+ = \max(z, 0)$, $0 \leq \beta_i < 1$, and $\alpha_i > \beta_i$. α_i represents the disutility from my opponent earning more than me while β_i stands for the disutility due to me getting more than my opponent.. We will introduce a bound on the value of α_i denoted by α_i^* so that (α_i, β_i) belong to the trapezoid with the vertices $(0, 0), (1, 1)(\alpha_i^*, 1)$ and $(\alpha_i^*, 0)$.

Observation 7: There exists a unique mental subgame perfect equilibrium outcome for the ultimatum bargaining game which is $(\frac{1+\alpha_2^*}{1+2\alpha_2^*}, \frac{\alpha_2^*}{1+2\alpha_2^*})$. Furthermore, as the bound α_2^* goes to infinity the unique equilibrium outcome goes to $(1/2, 1/2)$, which is the mode of the distribution of accepted offers in experimental results on the ultimatum game.

[5] An alternative way to define a mental SPE is to require that the strategy profile yields a mental equilibrium on each subgame of the game. It is easy to show that under this definition a strategy profile is a mental SPE if and only if it is an SPE.

Proof: It is clear that player 1 will be no better off if he selects a mental state different from the one with zero inequality aversion. Suppose that the mental state of player 1 offers the mental state of player 2 a payoff of less than 1/2, and assume α_2, β_2 are the parameters of inequality aversion of player 2. Then the mental state of 2 will accept the offer if and only if $x_2 - \alpha_2(x_1 - x_2) \geq 0$ or $x_2 \geq \frac{\alpha_2}{1+2\alpha_2}$; the fact that the right-hand side is increasing in α_2 and that the mental state of 1 can be assumed to be perfectly rational (has preferences identical to the material preferences) implies that player 2 should be assigned a mental state with maximal α, i.e., α_2^*. This in turn implies that among the mental states in the game the equilibrium outcome is $(\frac{1+\alpha_2^*}{1+2\alpha_2^*}, \frac{\alpha_2^*}{1+2\alpha_2^*})$ and furthermore no player by changing his mental state can generate a better SPE from his point of view. Finally, as α_2^* approaches infinity the allocation approaches $(1/2, 1/2)$.

We conclude this section by revisiting the trust game in the current framework where the set of mental states includes only Fher and Schmidt's (1999)-type of utility functions. We saw earlier that if we allow the set of mental states to include all utility functions, then any outcome in which the sender makes some transfer (possibly zero) and the receiver reimburses the sender for at least his cost can be supported by a mental equilibrium and nothing else. In our framework here, as we will show, there exists a unique mental equilibrium, which yields the socially optimal outcome. In this equilibrium the sender sends his entire bundle to the receiver and the receiver shares the amplified amount equally with the sender.

Observation 8: Assuming that the set of mental states includes all inequality averse-type utility functions, there exists a unique mental equilibrium in the trust game. In this equilibrium the sender sends x to the receiver and the receiver pays back $\frac{3}{2}x$ to the sender.

Proof: Clearly the sender cannot do better by having a mental state with a positive inequality aversion because what counts is not the preferences of the sender but his action. The receiver's best response to the sender's mental state is to have a mental state with an inequality aversion parameter β large enough so that it would make sense for the sender's mental state (whose preferences are identical to those of the sender) to transfer a positive amount and thus induce the mental state of player 1 to transfer the entire bundle to player 2. Note that if $\beta < 1/2$ the sender's mental state will make no transfer. On the other hand, if $1/2 < \beta < 1$, the receiver's mental state will attempt to equalize his own payoff to that of the sender's mental state. Hence, the sender's mental state is better off when he sends his entire endowment and gets back $\frac{3}{2}x$.

Interestingly, Observation 8 shows how the level of inequality aversion is determined endogenously. In equilibrium the receiver's mental state must have β between 1/2 and 1.

5. Imlementing Effort with Mental Equilibrium: Example

Two individuals operate on a project. Each individual is responsible for a single task. For the project to succeed it reqires that both tasks are successful. Playrs can choose to exert effort towards the performance of their task at a cost c which is identical for all agents. Effort increases the probability that the task succeeds from $\alpha < 1$ to 1. The principal cannot monitor the agents for their effort nor can he observe the success of individual tasks. However,

he is informed about the success of the entire project. An incentive mechansim is therefore given by a vector $v = (v_1, v_2)$ with agent i getting the payoff v_i if the project succeeds and zero otherwise (limited liability). Given a mechanism the two agents face a Noramal form game $G(v)$ with two strategies for each player : 0 for shirking and 1 for effort. The principal wishes to implement effort by both players at a minimal expense, i.e., he is looking for a mechanism under which there exists an equilibrium with both agents exerting effort. In Winter (2004) it is shown that the optimal mechanism pays each player $c/(1-\alpha)$ when agents' effort decision are taken simultaneously. If agents move sequentially (assuming that the second player observes the effort decision of the first) then the optimal incentive mechanism is pays $\frac{c}{1-\alpha^2}$ to the first player and $\frac{c}{1-\alpha}$ to the second. Under this mechanism player 2 will exert effort if and only if player 1 does so. This generates a implicit incentive on part of player 1 that allows the principal to pay him less than he pays in the simultaneous case (and less than the payoff of player 2 in the sequential case, see Winter (2006)). To model an environment in which the two agents can monitor each other effort we would need to split agents' task to many small sub-tasks and introduce a game of alternating effort decision (i.e., player 1 decides on the effort of the first sub-task, then player 2 decides on the first sub-task, the player 1 decide on the second sub-task etc.). It can be shown that in this environment the optimal mechanism pays each player $\frac{c}{1-\alpha^2}$, which is what player 1 (the player whose effort is observable) gets in the standard sequential case. We now show that under mental equilibrium this is also the optimal mechanism in the simultaneous case: Roughly, instead of player 1 affecting the effort decision of player 2 through his own effort choice, in mental equilibrium players affect each other mental state through their own mental state which enhances the prospect of cooperation. Our claim below is also extendable to environments with more than 2 agents.

Claim 1: The optimal mecahnsim for effort under mental equilibrium is $(\frac{c}{1-\alpha^2}, \frac{c}{1-\alpha^2})$.

Proof: Consider any pair of mental states (u_1, u_2) for the two agents such that given the action of player i player $j \neq i$ best response is to imitate the action of player i (i.e., j exerts effort iff i does so). We will show that under $v = (\frac{c}{1-\alpha^2}, \frac{c}{1-\alpha^2})$ effort by both players in a mental equilibrium (note however that it is not a Nash equilibrium). Indeed under the mental states specified earlier effort by both players is a Nash equilibrium. We therefore need to check only the second equilibrium condition. Assume w.l.o.g. that player 1 changes his mental state and by doing so he generates a Nash equilibrium which he prefers more with respect to his material preferences denote this mental state by u_1' . It must be the case that under u_1' taking the same action as player 2 cannot be the best response. Hence, either (1) $u_1'(1,1) < u_1'(0,1)$ or (2) $u_1'(0,0) < u_1'(1,0)$ or both. Furthermore since the only strategy profile in which player 1 material payoff improves is the one in which player 1 shirks and player 2 exerts effort, this profile must be a Nash equilibrium under the new mental state. This means that (1) must hold. But if (1) holds player 2 cannot exert effort in equilibrium. Under the payoff $\frac{c}{1-\alpha^2}$ player 2 is better off exerting effort only if 1 exerts effort. This contradiction rules out that player 1 or player 2 can be made better off by changing their mental state and shows that $(1,1)$ is a Mental equilibrium. To show that $v = (\frac{c}{1-\alpha^2}, \frac{c}{1-\alpha^2})$ is the optimal mechanism under Mental equilibrium we have to establish that if the principal pays, say, player 1 less, then the corresponding game has no mental

equilibrium in which both agents exert effort. Indeed if, say, player 1 is paid less than $\frac{c}{1-\alpha^2}$, then player 1 has a dominant strategy in the game, which is shirking. If effort by both players is sustainable by a mental equilibrium it must be that player 1 has a mental state under which he exerts effort. But then player 1 is better off changing his mental state to one in which shirking is a dominant strategy. This contradiction shows that effort by both players is not implementable by mental equilibrium with smaller payoffs.

6. n-Person Games

Our definition of mental equilibria requires that no player be able unilaterally to replace his mental state and improve his equilibrium outcome. This implies that if the mental state with whom player i deviates give rise to a game with no pure Nash equilibrium, then this deviation is not profitable. This turns out to expand the set of mental equilibria to the extent that it loses its predictive power for games with four or more players. We will later fix this drawback by introducing mixed strategies.

Proposition 4: For every normal form game G with $n \geq 4$ every strategy profile is a delegation equilibrium.

Proof: For each player i we select one strategy and denote it by 0. We denote by T_i the set of the remaining strategies so that $S_i = T_i \cup \{0\}$. We will show that the profile $(0, 0, ..., 0)$ is a mental equilibrium. Since the strategy was selected arbitrarily it will show that every profile is a mental equilibrium.

For a strategy profile $s \in S$ we denote $d(s) = \#\{j \in N \; s.t. \; s_j \in T_j\}$, i.e., the number of players choosing a strategy different from 0. For each integer k we denote by $p(k)$ the parity of k (i.e., whether k is odd or even). Consider now the following vector of mental states $(u_1, ...u_n)$ where $u_i : S \to \{0, 1\}$: $u_i(0, ..., 0) = 1$ for all i. For any strategy profile s different from $(0, ..., 0)$ we set $u_i(s) = 0$ if and only if $p(d(s)) = p(i)$. Otherwise $u_i(s) = 1$. We show that for any profile $s \neq 0$, half of the players can profit by deviating [6]. Indeed, each player who receives 0 can increase his payoff by changing his strategy from playing 0 to playing in T_i or the other way around. By doing so he will trigger a new profile s' for which $p(d(s')) \neq p(i)$ and he will raise his payoff from zero to 1. To show that $(0, 0, ..., 0)$ is a mental equilibrium first note that it is a Nash equilibrium with respect to the chosen mental states $(u_1, ...u_n)$. Furthermore, if player i deviates and sends a different mental state u_i' he will not be able to sustain a better equilibrium with respect to his basic preferences because for any other strategy profile there will be at least one mental state $j \neq i$ that will deviate, and hence the new mental game will have no pure Nash equilibrium.

We have shown that every two-person game has a mental equilibrium and that every game with at least four players admits all strategy profiles as mental equilibria. To complete the proof of existence we need a separate argument for three-person games.

Proposition 5: Every three-person game has a mental equilibrium.

Proposition 5 implies that:

[6] This holds when n is even; if the number of players is odd, then at least $\frac{n-1}{2}$ players will choose to deviate.

Corollary 2: Every n-person game has a mental equilibrium.

Proof of Proposition 5: We denote by s^* the strategy profile in which player 2 attains his highest payoff. If there is more than one such profile we select one of these arbitrarily. We will show that s^* is a mental equilibrium. We define the mental game to be $u_i(s^*) = 1$ for all players. We set again $S_i = T_i \cup \{s_i^*\}$ and $d(s) = \#\{j \in N \; s.t. \; s_j \in T_j\}$. For any other strategy, $u_i(s) = 0$ if and only if $p(d(s)) = p(i)$. Otherwise, $u_i(s) = 1$. We first note that s^* is a Nash equilibrium in the mental game. Furthermore, for any strategy profile of the mental game either players 1 and 3 want to deviate or player 2 alone does. To show that s^* is a mental equilibrium we need to show that no player can assign a different mental state and generate a new equilibrium that he prefers more. Clearly such a player cannot be player 2 as he has already attained his highest payoff. Suppose now that player 1 is better off assigning a different mental state and let s' be the new equilibrium that arises in the mental game that player i prefers to s^*. If $p(d(s'))$ is odd, then player 3 would deviate from s' in the mental game. If instead $p(d(s'))$ is even, then player 2 would deviate. Both consequences contradict that s' is an equilibrium in the mental game, which shows that s^* is a mental equilibrium.

Note that because we can rename the player an immediate corollary of Proposition 5 is that any strategy profile in which at least one player attains his maximal payoff is a mental equilibrium.

7. Mixed Strategies

We have seen that when we confine ourselves to pure strategies we get a multiplicity of equilibria, to the extent that every profile of strategies is a mental equilibrium when $n \geq 4$. To reduce the set of mental equilibria and increase the predictive power of our concept two tracks are possible. The first is to restrict the sets from which players may choose mental states. We used this approach in an earlier section when we restricted the set of mental states to include only utility functions representing inequality aversion. The second track is to introduce mixed strategies. At first this may sound puzzling: how can the introduction of mixed strategies shrink the set of equilibria? Indeed, in our equilibrium concept with pure strategies mental equilibria can arise simply due to the fact that players' deviations in choosing mental states lead to (mental) games that fail to have pure strategy equilibria. In such a case the conditions defining a mental equilibrium vacuously apply. By allowing mixed strategies we can guarantee that no matter what deviation a player undertakes, there will always be a Nash equilibrium in the new mental game. This expands the prospects of profitable deviation and can reduce the set of mental equilibria. Indeed, we will show that if we allow for mixed strategy equilibria in the mental games, then mental equilibria have a predictive power also for a large number of players. In our new solution concept the choice of mental states is pure but players' mental states can play a mixed strategy. For each player i, we denote by Δ_i the set of mixed strategies of player i. A mixed mental equilibrium is a profile of mixed strategies[7] $x \in \Pi_{i \in N} \; \Delta_i$ such that the following two conditions are

[7] Note that the set of mixed strategies also includes all the pure strategies.

satisfied:

(1) x is a mixed strategy equilibrium of the game (N, S, u).

(2) There exists no player i, a mental state u_i', and a mixed strategy equilibrium π' of the game (N, S, u_i', u_{-i}) with $U_i(\pi') > U_i(\pi)$.

Unlike the pure case where every pure Nash equilibrium is a mental equilibrium in the mixed version, we have:

Example 6: A Nash equilibrium of the game may not be a mixed mental equilibrium:

Consider the following two-person game:

1,1	0,2
0,0	1,-1

The game has a unique Nash equilibrium which is fully mixed. In this equilibrium both players assign a probability 1/2 to each of their strategies. Suppose by way of contradiction that this is a mental equilibrium and let the following game be the mental game supporting it:

a_1, b_1	a_2, b_2
a_3, b_3	a_4, b_4

For the strategy profile $[(\frac{1}{2}, \frac{1}{2}); (\frac{1}{2}, \frac{1}{2})]$ to be an equilibrium in the mental game we must have one of the following sets of inequalities:

$b_1 \leq b_2$ and $b_4 \leq b_3$ and $a_3 \leq a_1$ and $a_2 \leq a_4$ or

$b_1 \geq b_2$ and $b_4 \geq b_3$ and $a_3 \geq a_1$ and $a_2 \geq a_4$

In the first case player 2 is better off replacing his mental state with one in which the left strategy is dominant, and in the other case player 1 is better off changing his mental state to one in which his top strategy is dominant. In both cases a new equilibrium of (top,left) arises, yielding both players a payoff of 1 (in the original game), which is higher than the payoff of 1/2 that they both get under the putative mental equilibrium. Thus arises a contradiction showing that the Nash equilibrium is not a mixed mental equilibrium.

We further show that the game has a mental equilibrium of (1,1). For this we take the mental state game $\begin{matrix} 1,1 & 0,0 \\ 0,0 & 1,1 \end{matrix}$, where (top,left) is an equilibrium. Clearly player 1 has no incentive to change his mental state since 1 is the highest payoff he can get. Consider the other player. Player 2 can be better off with a different mental state if either (top,right) can be made an equilibrium or there is some mixed strategy equilibrium yielding more than 1 to player 2. The former case is impossible since 1's mental state will deviate from (top,right) to play bottom. Concerning the latter case if there is a mixed equilibrium 2's mental state's strategy must be (1/2,1/2) (to make player 1 indifferent between his two strategies), and so the only possible deviation for player 2 is a mental state that assigns a higher payoff for (top,right). If this payoff is greater than 1 then the new game has a unique equilibrium which is again (top,left); if the payoff is less than 1, then player 1's mental state must assign a higher probability to bottom (in order to make player 2's mental state indifferent). This means that the mixed strategy equilibrium will yield an expected payoff of less than 1/2 in the original game for player 2. Hence, player 2 cannot profitably deviate in choosing his mental state.

We now go back to the famous public good game that we discussed in the Introduction to show that the notion of mixed mental equilibrium is rather instructive for this game no

matter how large it is.

Example 7: The social dilemma game/The public good game.

n players hold an endowment of $w > 0$ each. Each player has to decide whether to contribute to the endowment (choose 1) or not (choose 0). The total endowment contributed is multiplied by a factor $1 < k < n$ and divided equally among all players. Thus supposing that r players contribute, the payoff for a player who chooses 1 is $\frac{krw}{n} - w$ and the payoff for a player who chooses 0 is $\frac{krw}{n}$. Note that the unique equilibrium in the game is $(0, ..., 0)$, but the profile that maximizes social welfare is $(1, ..., 1)$.

Observation 8: A strategy profile in the public good game is a mental equilibrium if and only if either no one contributes or the number of contributors is at least $\frac{n}{k}$.

Proof: We first show that any profile in which the number of contributors is positive but with a proportion of less than $\frac{1}{k}$ cannot be a mental equilibrium. Suppose by way of contradiction that such an equilibrium exists. Consider a player i whose mental state contributes. Player i's payoff in such an equilibrium is $\frac{krw}{n} - w$. Suppose that this player assigns a different mental state than choosing 0 as a dominant strategy. The new mental game must have an equilibrium (in pure or mixed strategies). In the worst-case scenario (for player i) this equilibrium is $(0, ..., 0)$, in which case player i's payoff will be w. If the proportion of contributors is less than $\frac{1}{k}$, then $w > \frac{krw}{n}$ and player i is better off deviating. If the equilibrium is not $(0, ..., 0)$, then with positive probability some players contribute in the equilibrium of the new mental game and the expected equilibrium payoff of player i is greater than 0, which makes deviation even more attractive. We now show that a profile with a proportion of contributors $p \geq \frac{1}{k}$ is a mental equilibrium. Consider such a profile and denote by T the set of players who choose 0 and by $N - T$ the players who choose 1. To show that this profile is a mental equilibrium we assign the following mental states to players. For each player in $N - T$ we assign a mental state that prefers to choose 1 if and only if the proportion of agents who choose 1 is at least p (otherwise he prefers to choose 0). For each player in T we assign a mental state whose preferences are identical to those of the other players (i.e., choosing 0 is a dominant strategy). Given this set of mental states it is clear that the underlying strategy profile is an equilibrium of the mental game. It therefore remains to show that condition (2) in the definition of mental equilibrium applies. Clearly no player in T can be better off deviating. Assigning a different mental state will trigger no one else to contribute in the mental game. Consider now a player i in $N - T$. Suppose i nominates a different mental state and assume by way of contradiction that π' is the new equilibrium with respect to which player i is made better off. If the mental state of player i chooses 1 with probability 1 in π', then player i is neither better off nor worse off when deviating and π' is identical to the original profile. Suppose therefore that the mental state of player i chooses 0 with positive probability in π'. Since each mental state whose player is in T has a dominant strategy to choose 0 the expected proportion of mental states that choose 1 in π' is less than p. But this means that each mental state whose player is in $N - T$ has a best response to π', which is choosing zero, which contradicts π' being an equilibrium. To complete the proof of the proposition it remains to show that $(0, ..., 0)$ is a mental equilibrium. This is done by assigning to each player i a mental state with preferences identical to those of player i. Since choosing 0 is a dominant strategy for each player, $(0, ..., 0)$ is a Nash equilibrium in the mental game and no player can be made

better off by assigning a different mental state.

The attractive property of mental equilibria when applied to the public good game is that in contrast to the concept of Nash equilibrium where the set of equilibria is invariant to the value of k (i.e., the extent to which joint contribution is socially beneficial), the set of mental equilibria strongly depends on k in a very intuitive way. As k grows the social benefit from joint contribution become substantial even when the number of contributors is low; this allows for more strategy profiles with a small number of contributors to be sustainable as equilibria.

The observation made in the example above that $(1, 0, ..., 0)$ cannot be a mental equilibrium can be generalized:

Let G be an n-person game. For a mixed strategy profile x denote by $f_i(x)$ the expected payoff for player i under the profile x. Let the payoff that player i can guarantee himself regardless of what the other players are doing be denoted by $a_i = \max_{x_i \in \Delta_i} \min_{x_{-i} \in \Delta i} f_i(x_1, ..., x_n)$.

Proposition 6: Any mixed mental equilibrium must yield each player i a payoff of at least a_i.

Proof: Suppose that $x = (x_1, ..., x_n)$ is a mental equilibrium with $f_i(x) < a_i$. Suppose that G^* is the mental game sustaining this equilibrium. We denote by 0_i the payoff function of player i that assigns a zero payoff for all strategy profiles. Consider player i changing his mental state by choosing the mental state 0_i (if 0_i is the original mental state, then player i will choose any other mental state which is indifferent between all the strategy profiles) and denote by $G^*_{0_i}$ the game obtained by replacing the mental state of player i with 0_i. Define $x_i^0 = \arg\max_{x_i \in \Delta_i} \min_{x_{-i} \in \Delta i} f_i(x_1, ..., x_n)$, and let $G^*_{x_i^0}$ be the game defined on the set of players $N \backslash \{i\}$ such that $f_j^{x_i^0}(x_{N\backslash\{i\}}) = f_j(x_i^0, x_{N\backslash\{i\}})$. Let z be a Nash equilibrium of the game $G^*_{x_i^0}$. We claim that (x_i^0, z) is a Nash equilibrium of the game $G^*_{0_i}$. Indeed the fact that no player in $N \backslash \{i\}$ can do better by deviating follows from the fact that z is a Nash equilibrium of $G^*_{x_i^0}$. The fact that i cannot do better is a consequence of i being indifferent between all his strategies. By the definition of x_i^0 we have that $f_i(x_i^0, z) \geq a_i$, which contradicts the assumption that x is a mixed mental equilibrium.

Note that Example 7 implies that the converse of Proposition 6 is not true. The Nash equilibrium of the game (which is not a mental equilibrium) yields a payoff vector of $(\frac{1}{2}, \frac{1}{2})$, which exceeds the maxmin vector $(0, 0)$.

Corollary 3: In two-person zero-sum games there is a unique mixed mental equilibrium. This equilibrium yields the value of the game.

Proof: Follows directly from the proposition above.

We conclude with another useful property of mental equilibrium:

Proposition 7: Let G be an n-person game and let s and s' be two pure strategy profiles yielding the payoff vectors $u = (u_1, ..., u_n)$ and $v = (v_1, ..., v_n)$ respectively and such that v dominates u ($v_i \geq u_i$). If s is a mixed mental equilibrium, then s' must be a mixed mental equilibrium as well.

Proof: Let $s = (s_1, ...s_n)$ be the pure strategy profile that sustains u and let $s' = (s'_1, ...s'_n)$ be the strategy profile that sustains v. Let $C = (C_1, C_2, ..., C_n)$ be the mental game supporting u as a mental equilibrium (C_i is a payoff function of the mental state of

player i in the mental game). By supposition s is a Nash equilibrium of C. Since both s and s' are pure strategy profiles we can rename strategies for each player so that the new game C' is isomorphic to C up to strategy names and such that s' is an equilibrium of C'. Suppose by way of contradiction that s' is not a mixed mental equilibrium. Then it must be the case that some player i can change his mental state from C_i' to C_i^* in such a way that in the new mental game $(C_i^*, C_{-i}^{'})$ there exists another equilibrium s^* with $G_i(s^*) > G_i(s^{'})$. But the isomorphism between C and C' implies that there is a mental state of player i $\overline{C}_i$ such that $s^*$is an equilibrium of the game $(\overline{C}_i, C_{-i})$ with $G_i(s^*) > G_i(s^{'}) \geq G_i(s)$, which contradicts the fact that s is a mental equilibrium.

As a corollary of Proposition 7 we obtain that the cartel behavior in an oligopoly/Cournot game is supported by a mental equilibrium. This follows from the fact that being a pure Nash equilibrium, the Cournot equilibrium is a mental equilibrium. Since cartel behavior yields a higher payoff for each player, Proposition 7 implies that it must also be a mental equilibrium.

8. Appendix

We provide two examples showing that neither of the two sides of Proposition 1 applies to three-person games:

Example 8: Consider the following three-person game:

	L	R
U	1,1,1	0,0,0
D	1,2,3	1,3,0

	L	R
U	1,1,2	2,0,0
D	2,0,0	1,1,1

The maxmin vector of this game is $(1, 0, 0)$. Hence, (U, R, L) does not pay player 1 at least his maxmin value in this game. However, it is a mental equilibrium. To verify the claim consider the following profile of mental states:

	L	R
U	1,0,0	1,1,1
D	0,0,1	0,1,0

	L	R
U	0,0,1	1,1,0
D	1,1,0	0,0,1

Notice that (U, R, L) is a Nash equilibrium of this game. Suppose now that one player unilaterally deviates to a different mental state; then the only possible Nash equilibria different from (U, R, L) are (D, L, R) and (U, R, R). However, these can be Nash equilibria only if player 3 is the deviating player. Since $U_3(U, R, L) = U_3(D, L, R) = U_3(U, R, R)$, we must have that (U, R, L) is a mental equilibrium of this game.

Example 9: Consider the following three-person game:

	L	R
U	0,0,0	1,1,1
D	1,1,1	1,1,1

	L	R
U	1,1,1	1,0,1
D	0,1,1	1,1,0

The maxmin vector of this game is $(0, 0, 0)$ and all strategy profiles of the game pay each player at least his maxmin value, in particular the profile (U, L, L). However, this profile is not a mental equilibrium. Suppose by way of contradiction that it is. Then, there must

exist a profile of mental states satisfying the conditions of mental equilibrium. The second condition of the definition (i.e., no player is better off changing his mental state) implies the following: (A) for the strategy profiles (U, R, L), (D, L, L), (D, R, L), (U, L, R), at least two players are willing to deviate, and (B) in (D, L, R) either player 1 wants to deviate or players 2 and 3 want to deviate, and in (U, R, R) either player 2 wants to deviate or players 1 and 3 want to deviate, and in (D, R, R) either player 3 wants to deviate or players 1 and 2 want to deviate. It is easy to verify that (A) and (B) cannot be simultaneously consistent.

9. References

[1] Bergman, Nittai, and Bergman, Yaacov Z (2000). "Ecologies of Preferences with Envy as an Antidote to Risk-aversion in Bargaining," mimeo, The Hebrew University of Jerusalem.

[2] Bolton, Gary E., and Ockenfels, Axel (2000). "A Theory of Equity, Reciprocity and Competition," *American Economic Review* 90, 166-193.

[3] Tang, Fang-Fang, and Nagel, Rosemarie (1998). "Experimental Results on the Centipede Game in Normal Form: An Investigation on Learning," *Journal of Mathematical Psychology* 42, 356-384.

[4] Fershtman, Chaim, and Heifetz, Aviad (2006). "Read My Lips, Watch for Leaps: Preference Equilibrium and Political Instability," *The Economic Journal* 116, 246-265.

[5] Fher, Ernst, and Schmidt, Klaus (1999). "A Theory of Fairness, Competition and Cooperation," *The Quarterly Journal of Economics* 1, 817-868.

[6] Gueth, Werner, Schmittbereger, Rolf, and Schwarze, Bernd (1982). "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3, 367-388.

[7] Gueth, Werner, and Yaari, Menahem (1992). "An Evolutionary Approach to Explaining Reciprocal Behavior in a Simple Strategic Game," in *Explaining Process and Change*, Witt, Ulrich (ed.), Ann Arbor, MI: The University of Michigan Press.

[8] Gueth, Werner, and Kliemt, Hartmut (1998). "The Indirect Evolutionary Approach: Bridging Between Rationality and Adaptation," *Rationality and Society* 10, 377-399.

[9] Gueth Werner, and Ockenfels, Axel (2001). "The Coevolution of Morality and Legal Institutions: An Indirect Evolutionary Approach," mimeo, Max Planck Institute for Research into Economic Systems.

[10] Olschewski, Guido and Lukasz Swiatczak (2008) "Existence of Mental Equilibria in 2x2 Games. mimeo Handelshochschule Leipzig.

[11] Rabin, Matthew (1993). "Incorporating Fairness into Game Theory and Economics," *American Economic Review* 83, 1281-1302.

[12] Rapoport, Anatol, Guyer, Melvin J., and Gordon, David G. (1976). *The 2 X 2 Game*, Ann Arbor, MI: The University of Michigan Press.

[13] Roth, Alvin A., Prasnikar, Vesna, Okuno-Fujiwara, Masahiro, and Zamir, Shmuel (1991). "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review* 81, 1068-1095.

[14] Winter, Eyal (2004) "Incentives and Discrimination" *American Economic Review* 94 (3) 764-773

[15] Winter, Eyal (2006) ”Optimal Incentives for Sequential Production” *Rand Journal of Economics* 37 (2) 376-390

[16] Winter, Eyal., Gershon Ben Shahar, Itzhak Aharon, Meir Meshulam (2008) ”Rational Emotions” Preliminary Notes, The Center for the Study of Rationality, The Hebrew University of Jerusalem.