Mental Equilibrium and Rational Emotions[1]

Eyal Winter, Ignacio Garcia-Jurado, Luciano Mendez-Naya

Jan 20, 2012

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 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 experimental economics literature. We shall demonstrate the role of mental equilibrium in incentive mechaisms and will discuss the concept of collective emotions, which is based on the idea that players can coordinate their emotional states.
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 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. 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

[1] The authors wish to thank Itai Arieli, Ken Binmore, Werner Gueth, Sergiu Hart, Eric Maskin, Assaf Rom, Reinhard Selten, and Jean Tirole for their comments and suggestions on an earlier draft of this paper. We also thank audiences at Boconni, Copenhagen, Harvard, Johns Hopkins, Max Planck in Jenna, Northwestern, Michigan, Minnesota, Paris School of Economics, Tel Aviv, UBC, UCLA, Wisconsin, The Behavioral Game Theory Workshop at Stony Brook, and the Fifth International Meeting on Experimental and Behavioral Economics in Granada for numerous suggestions and comments.

been developed that try to reconcile the discrepancy between experimental results and game-theoretic predictions, without neglecting the idea that players behave strategically. The common objective of these papers is to reevaluate the outcomes of the game for each player, while taking into account emotional factors such as 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 the 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) and that these immaterial preferences are rational in the sense that they promote players' material interests. We shall use the term *mental state* to represent these emotions[2] as part of an equilibrium concept called *mental equilibrium*, which seems to organize the experimental evidence for some of the most prominent examples quite well. Much of the focus of our analysis will be on deriving players' behavioral preferences endogenously through the equilibrium conditions.

The concept of mental equilibrium can be described as follows. Each player, who we assume seeks to maximize only 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 with respect to players' mental states. Secondly, each player's mental state is a best response to the mental states of the other players, given his material and selfish preferences. We offer two valid interpretations for our equilibrium concept. The first 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 reasonable to assume that fairness, anger, envy, and revenge, which play a role in many game situations, have been developed through evolutionary forces to increase individuals' fitness to the social environment in which they live. Our equilibrium concept can be viewed as a theoretical foundation for the emergence of such emotions. We are not proposing any specific evolutionary model to this effect, but conceptually mental equilibrium can be viewed as a stability concept arising from an evolutionary proccess. Evolutionary selection reinforces different mental states in different strategic environments, and material payoffs in the game can be viewed as a measure of fitness. 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

[2] We use "emotions" or "mental states" when we refer to social/immaterial preferences although emotions or mental states are in fact the mechanism by which these preferences arise.

their interest. Emotional states are often induced through cognitive reasoning whether 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[3]. 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 to 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 that day. 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 in which you would have been in scenario A. You are likely to exhibit signs of anger quite quickly in scenario B; in fact, these won't be mere signs, you will actually be angry. You have been offered incentives to be angry and as a consequence you "choose" to be angry. The example above suggests 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. (2010), which shows that the objective emotional reactions of receivers in a dictator game strongly depend on the presence of incentives. Under the interpretation of rational emotions 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. The two interpretations we propose are very distinct. The evolutionary interpretation fit emotions which are global and robust, while under the rational emotion interpretation they can be specific and fragile. However, we shall be subscribing to both interpretations and will not argue in favor of one of them as we believe that the appeal of each of these interpretations is context-dependent . In particular, in explaining the foundation of emotional conventions and norms in vaguely defined games and that are robust to whether players can see each other or not, the evolutionary approach seems more appropriate (most "blind" experiments fall under this category). On the other hand, the interpretation of rational emotions might be more relevant to situations that rely on mutual eye contact and are strongly responsive to incentives. We point out that the distinction between the two interpretations is akin to the recent distinction made by Aumann (2009) between rule rationality and act rationality. In both interpretations, 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 shall restrict the set of mental states to include only preferences of inequality aversion as in Fher and Schmidt (1999), and we shall be able to endogenously derive conditions on the parameters of inequality aversion that mental states must exhibit in equilibrium.

[3] A considearble body of recent papers in the psychology literature discusses the conscious control and regulation of emotions (see Demasio et al (2000), Ochsner and Gross (2005)). Tice and Bratslavsky (2000) suggest specific types of emotion control tasks (such as "getting into" and "getting out of" emotions) and discuss their regulation strategies.

An implicit assumption that is built into the definition of mental equilibrium is that players must have correct beliefs regarding other players' mental states when playing a game. This is a critical issue 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 (like the Nash equilibrium). Nevertheless, there are two grounds on which this assumption can be justified. Firstly, players' choice of mental states involves some sort of pre-play comunication game that we intentionally leave unspecified. Players signal their mental state in this game through body movement, facial expression, voice intonation, and other actions. One cannot exclude deception, but it makes sense to assume that the longer and the more elaborate this pre-play game is, the less likely and the more costly it is for players to manage a successful deception. But even without direct eye contact players may still form consistent beliefs about the mental states of their counterparts. Just as with the learning literature that explains how consistent beliefs leading to Nash equilibrium emerge, it is conceivable that one can come up with a dynamic model that converges to consistent beliefs about mental states. Such a model can rely on the intuition that by experiencing identical or similar strategic environments over and over again players can learn quite a bit about the function that maps strategic environments onto mental states. While interesting and important in themselves, these learning and signaling models are beyond the scope of this paper.

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. To this end we shall introduce a battery of well-known games about which considerable experimental data has been collected and we shall 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 the experimental results of the games considered here very often correspond to mental equilibria that are not Nash equilibria). In doing so we shall identify the mental states that support various prominent experimental results as mental equilibria.

In addition to its relation to the literature on social preferences that we have discussed above, our work is related and inspired by two other strands of literature. The first is the literature on delegation pioneered by Fershtman, Judd, and Kalai (1990). This paper discusses strategic environments in which players can choose delegates to play a game on their behalf. By setting up the incentives to delegates properly, players can support strategic outcomes that are not standard Nash equilibria (see also Fershtman and Kalai (1997) and Bester and Sakovic (2001)). The second strand of literature concerns papers that discuss the evolutionary foundation of preferences. Gueth and Yaari (1992) introduced a game of cooperation between two players and showed how preferences for cooperation (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). 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 Fershtman and Heifetz (2006) in the context of elections and political competition. Our paper departs from the two strands of literature discussed above in motivation, interpretation, and formal modeling. Our objective is to study the role of emotions in strategic decision-making. Accordingly, much of our attention will be given to identifying the mental states that support specific strategic outcomes. We shall compare our model with experimental observations and argue that it well organizes laboratory evidence from several well-known experiments. In doing so we shall specify the mental states that support various prominent experimental outcomes. In terms of formal modeling our model differs from those used in the literatures discussed above. It is more general in that it deals with the class of all games and with an arbitrary number of players. Mental states in our model differ from delegates in the Fershtman et al. paper in the sense that they induce no costs to the players (although one can think of a framework in which they can). Motivated by the idea of rational emotions we do not specify evolutionary conditions for stability. Instead, our model involves two levels of equilibrium conditions. One level involves the mental game in which the payoffs are derived from players' mental states (emotions) and the other level involves the selection of players' mental states to maximize material preferences. 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 selection of emotions are less stringent than Dekel et al.'s (2007) evolutionary conditions, our set of mental equilibria is typically larger than the set of stable outcomes à la Dekel et al. (2007) and other related papers, and our model admits a (pure) mental equilibria for any game. Finally, we expand the scope of applications by defining mental equilibrium variants to other solution concepts (beyond Nash equilibrium) including subgame perfect equilibrium and strong Nash equilibrium.

In Section 2 we continue with the formal definition of mental equilibrium. We start with the simplest model where mental states are assumed to play only pure strategies. In Section 3 we provide a useful characterization of mental equilibria in two-person games, which we later use to study mental equilibria in some prominent games for which experimental results have been accumulated. We then reflect on the mental states that support outcomes that are observed in the laboratory. In Section 4 we discuss the consequence of restricting the set of mental states, using the set of preferences of inequality aversion as our domain of mental states. We then show how mental equilibrium can endogenously derive parameters of inequality aversion for some prominent games.

We devote Section 5 to a discussion of the role of mental equilibrium in the context of contracting and incentive mechanisms, using a simple model of moral hazard in teams. Our main observation here is that the cost of implementing effort under mental equilibrium is much less than the cost under Nash equilibrium and is in fact equivalent to the cost of implementing effort in a sequential mechanism where players operate under full transparency regarding peers' effort. This is due to the fact that the extra incentive to exert effort that emerges from the threat of retaliation by peers, when transparancy is available, is internalized at the level of mental states even when no transparency is available.

Sections 6 and 7 deal with a model of mental equilibrium in which mental states can use

mixed strategies. This model is motivated in Section 6 by showing that for games with four or more players the standard concept of mental equilibria (based on pure strategies) 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. We study properties of this amended concept of *mixed mental equilibrium* and apply it to the game of voluntary contributions (the n-person Prisoner's Dilemma). We show that in a mental equilibrium either no one contributes or the set of contributors is sufficiently large. These equilibria are supported by very intuitive mental states in which players experience substantial disutility when they contribute alone or together with a small group of contributors. In Section 8 we discuss collective emotions. These emotions emerge through coordination within a group to enhance the rational role of emotions as a commitment device. Our definition and analysis here builds on Aumann's (1959) notion of strong equilibrium. Strong mental equilibrium, which is our main concept here, uniquely selects cooperation in the Prisoner's Dilemma, quite differently than anything else in the plethora of game-theoretic solution concepts. We conclude in Section 9.

2 Basic Definitions

In the first part of this paper we shall assume players (in all their mental states) play only pure strategies. Later we shall expand the model by allowing the mental game to involve also mixed strategies. As we shall see, these two models are not nested. The pure strategy model, while simpler to use for applications, is more limited in its predictive power for games with more than two players.

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 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 Nash equilibria of the game G.

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

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

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

Condition 1 in the definition of mental equilibrium requires that once the mental states have been determined, the players' interaction will result in a Nash equilibrium. Condition 2 requires that the players' mental states be chosen rationally with respect to their material preferences. We proceed with the following basic observation:

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. It is interesting to note (as we shall show later) that observation 1 does not hold for mixed strategy Nash equilibria.

3 Two-Person Games

In this section we offer a simple characterization for the set of mental equilibria in two-person games, which will prove useful for various applications. In any Nash equilibrium each player attains at least his maxmin value. Proposition 1 asserts that this property is both a necessary and sufficient condition for (pure) mental equilibria in two-person games.

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 : Let v_1 and v_2 be the maxmin values of players 1 and 2 respectively with s_1 and s_2 being the maxmin strategies. 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 to player 2 and we conclude that s must be a mental equilibrium.

Proposition 1 almost immediately implies the existence of mental equilibrium for two-person games:

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

Proof: Proposition 1 implies that it is sufficient to show that in any two-person game there exists a strategy profile that pays each player at least his/her maxmin value. To show this, let s_1' be a maxmin (pure) strategy for player 1, i.e., $s_1' = \arg\max_{s_1} \min_{s_2} U_1(s_1, s_2)$ and let s_2' be a best response (pure) strategy to s_1'. Clearly, (s_1', s_2') is the desired profile. Player 1 gets paid at least his/her maxmin payoff per definition and for player 2 this holds because a best response to any of player 1's strategies must yield player 2 at least his maxmin payoff.

Our definition of mental equilibrium relied on the assumption that players are optimistic when contemplating deviations as it is enough that there exist 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 mental game yield a higher utility level. 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 exist no player i, and a mental state u_i' such that $NE(N, S, u_i', u_{-i}) \neq \varnothing$ and 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 2: 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 that 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 definition of weak mental equilibrium.

A large body of experimental results has been obtained for two-person games. Proposition 1 serves as a very useful tool for identifying the set of mental equilibria for such games. We shall now discuss some of the most prominent examples 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 each 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 mental equilibrium outcomes, note that a payoff of zero is less than the maxmin level (which is 1). By Proposition 1, (D, C) and (D, C) are not mental equilibria.

It can be easily verified that the outcome (C, C) can be supported as a mental equilibrium through the following mental states: $u_1(C, D) = u_2(C, D) = u_1(D, C) = u_2(D, C) = -1$, and $u_i = U_i$ otherwise. Note that these mental preferences represent aversion to lack of irreciprocity, i.e., both players suffer when one of them cooperates and the other one defects.

It is instructive to characterize the set of mental states that support the cooperative outcome as a mental equilibrium in a general Prisoner's Dilemma game. In fact we shall characterize the set of mental states supporting cooperation for a larger class of games which we call Cooperation Games. A two-person cooperation game is a game with two strategies {D,C} for each player such that $U_1(D, C) > U_1(C, C)$, $U_2(C, D) > U_1(C, C)$ and $U_i(C, C) > U_i(D, D)$ for $i = 1, 2$. Every Prisoner's Dilemma game is a cooperation game but the set of Cooperation Games includes also all chicken games.

Observation 3: Let $G = (N, S, U_1U_2)$ be a Cooperation Game. Then (C, C) is a mental equilibrium. Furthermore, a necessary and a sufficient condition for the mental states (u_1, u_2) to sustain (C, C) as mental equilibrium in *G* is : $u_1(C, C) \geq u_1(D, C)$, $u_1(D, D) \geq u_1(C, D)$ and $u_2(C, C) \geq u_2(C, D)$, $u_2(D, D) \geq u_2(D, C)$.

Proof: Consider first mental states (u_1, u_2) that satisfy the conditions above. It is easily verified that (C, C) is a Nash equilibrium under these mental states. Furthermore, in order for, say, player 1 to increase his material payoff, this player needs to deviate to a mental state u_1' for which (D, C) is a Nash equilibrium under the payoff functions (u_1', u_2). But $u_2(D, D) \geq u_2(D, C)$, which is a contradiction. So (C, C) is a mental equilibrium supported by (u_1, u_2). Consider now any profile of mental preferences (u_1, u_2). First, both the first and the third inequalities must hold. Otherwise, (C, C) cannot be a Nash equilibrium under (u_1, u_2) (as player 1 would deviate if the first 1 fails and player 2 deviates if the third one fails). Suppose that second inequality is violated, then the mental state of player 2 is not optimal, as player 2 can be made better off (in terms of material preferences) with the mental state u_2' satisfying $u_2'(C, D) \geq u_2'(C, C)$ as under (u_1, u_2') the outcome (C, D) is a Nash equilibrium. Likewise if the fourth inequality fails to hold, then player 1 is better off deviating to u_1' with $u_1'(D, C) \geq u_1'(C, C)$ and increase his material payoff as under (u_1', u_2) the outcome $(D, C)$is a Nash equilibrium .

Observation 3 has the important implication that players' mental states *must* have the reciprocity-seeking property to sustain cooperation in any Prisoner's Dilemma game. This is an important insight that cannot be derived from standard game-theoretic solution concepts. To elaborate on this point, we shall consider here two alternative types of mental preferences—the first one involving altruism and the second based on inequality aversion—to demonstrate that none of these can explain cooperation in the Prisoner's Dilemma.

Starting with altruism we argue that mental preferences sustaining the cooperative outcome cannot be of the form $u_i = \alpha_i U_i + \beta_i U_j$. Based on the payoff function in our example above, these mental pereferences would result in the following mental game:

	D	C
D	$\alpha_1 + \beta_1$,$\alpha_2 + \beta_2$	$\alpha_1 5, \beta_2 5$
C	$\beta_1 5, \alpha_2 5$	$4(\alpha_1 + \beta_1), 4(\alpha_2 + \beta_2)$

For (C, C) to be an equilibrium in this mental game we need to have $4(\alpha_2 + \beta_2) \geq 5\alpha_2$. 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 $5\beta_2 \geq \alpha_2 + \beta_2$.

Note the difference between the social preferences given by $u_i = \alpha_i U_i + \beta_i U_j$ and the one we used in Observation 2. The former represents a mental state with some degree of altruism (if $\beta_i > 0$) or spitefulness (if $\beta_i < 0$). 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 next discuss inequality aversion (à la Fehr and Schmidt (1999) or à la Bolton and Ochenfels (2000)) and consider the following Prisoner's Dilemma game:

	D	C
D	$35, 50$	$45, 45$
C	$30, 65$	$40, 60$

We point out that an inequality-averse mental state of player 1 must satisfy $u_1(D, C) > u_1(C, C)$. This is because (D, C) generates more (material) payoff for player 1, and involves more equality than the outcome (C, C). Hence, given our discussion above, there exists no profile of mental preferences that will support (C, C) as a mental equilibrium in this Prisoner's Dilemma game.

We conclude that aversion to lack of reciprocity can explain cooperation in every Prisoner's Dilemma game, but altruism, spitefulness, or inequality aversion cannot.

Example 2: The Chicken Game

Consider the following two-person game:

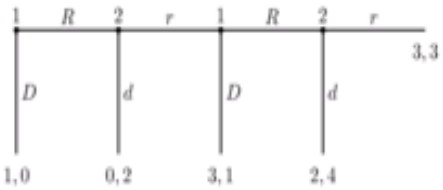
	retreat	fight
retreat	$1, 1$	$-2, 2$
fight	$2, -2$	$-10, -10$

Observation 4: 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.

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 a 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). It is interesting to note that the mental preferences that sustain (retreat, retreat) as a mental equililibrium have the same form as those we constructed for the Prisoners' Dilemma, i.e., with players' experiencing major disutility whenever one of them fights and the other retreats.

Example 3: The Trust Game

Massive experimental data have been accumulating on the Trust game since Berg,



Dickhaut, and McCabe (1995). In its most standard form the game can be described as follows: Player 1 has an endowment of x. 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 5: 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, it is clear that 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 a 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 gets less than the 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 5 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, e.g., Berg, Dickhaut, and McCabe (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

The Centipede game is a sequential version of the trust game. The extensive form game, presented in Figure 1, is a simple version 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 6: In 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 (material) payoff of 1 (instead of zero). This argument does not apply for the outcome (3,1). This outcome can be supported by mental states that assign the following payoffs to the terminal nodes of the game (from left): (0,1), (0,2), (7,7), (0,0), and (0,0). Roughly speaking, any mental equilibrium outcome of the Centipede game is supported by mental states in which players' emotions are coordinated to achieve cooperation up to a certain level of depth but not beyond. 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 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 exit 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

The 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. Massive experimental evidence starting with Gueth et al. (1982) has however shown that player 1 makes substantial offers, with the mode of the distribution being 50:50. To discuss the concept of mental equilibrium for this game we first need to discuss the sequential/subgame perfect version of mental equilibrium. We shall show that this concept has little bite if one is allowed to consider all mental states that will motivate our interest in restricting the set of mental states in the next section. The following is a natural definition of Mental Subgame Perfect Equilibrium:

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

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

Observation 7: Take any extensive form game with perfect information G. Every Nash equilibrium outcome of G is a mental SPE outcome of G.

Proof: Let s be a Nash equilibrium of the game. We construct the following mental (extensive form) game. For each player i choose a mental state u_i in the following manner: For each terminal node d of the game, $u_i(d) = 1$ if and only if the path leading to d is consistent with player i playing the strategy s_i, i.e., at each decision node along this path player i's action is as specified by s_i (note that d does not have to be part of the equilibrium path of s). If the path leading to d is not consistent with s_i, we take $u_i(d) = 0$. The construction described above can intuitively be thought of as making s_i a dominant strategy in the normal form version of the game. Unfortunately, we cannot work directly with the normal form game as it has more strategies than the number of terminal nodes in the extensive form game. We first argue that s constitutes a subgame perfect equilibrium in the mental game based on a profile of mental states u. This is done by backward induction by noting that at each decision node if i has not yet deviated from s, then choosing to stay with s would yield i (using the induction hypothesis) a payoff of 1,while deviating from s would get him zero. It is left to show that no player can unilaterally change his mental state in such a way that the new mental game will possess a subgame perfect equilibrium with a higher material payoff for this player. Suppose by way of contradiction that such a mental state u_i' exists, and consider the mental game based on $(u_i'\ u_{-i})$.Again, by backward induction all players j other than i must have SPE strategies that are consistent with their s_j. Let s_i' be the SPE strategy of player i in the new mental game. By assumption we have $U_i(s_i', s_{-i}) > U_i(s)$.But this cannot happen since s is a Nash equilibrium.

Subgame perfect equilibria are often described as Nash equilibria with credible threats. Mental equilibria that are based on commitment can turn non-credible threats into credible ones. This is the basic insight of Observation 7 and of the fact that a mental subgame perfect equilibrium is not an effective refinement of mental equilibria. An alternative definition would be to require that mental subgame perfect equilibrium induce a mental equilibrium on every subgame[4]. Using a backward induction argument one can easily verify that in games of perfect information the set of equilibria under this definition will coincide with the set of subgame perfect equilibria (with the standard definition)[5]. In the Discussion section of this paper we suggest an intermediate concept that allows players to change their mental state during the course of the game, but the formal analysis is beyond the scope of this paper. As we have seen, without restricting the set of mental states a mental SPE does not have sufficient teeth in games of perfect information. Going back to the Ultimatum game we shall show that restricting the set of mental states offers a much better insight into the game.

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

[4] Parallel to the definition of SPE which requires that it induce a Nash equilibrium on every subgame.

[5] This is not the case for games with imperfect information. Consider for example the two-player game in which player 1's first choice is between "exit" and "enter." If he chooses "exit" both players get 3. If he chooses enter, they play the Prisoner's Dilemma with the payoffs specified in Example 1. The game has a unique SPE outcome, which is player 1 choosing exit. On the other hand a mental SPE admits two equilibria: one in which player 1 chooses "exit" (expecting (D,D) to be played if he "enters") and another equilibrium in which player 1 enters and (C,C) is played in the subgame.

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 shall start with the Ultimatum game and then explore mental equilibria in this framework for other games. This analysis will contribute to the debate conducted in the early nineties over the role of fairness in Ultimatum games and games in general. Our objective in the analysis below is also methodological, as it will show how standard models of social preferences can be incorporated into the framework of mental equilibrium to offer further insight into experimental results.

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 one's opponent earning more than one, while β_i stands for the disutility arising from one getting more than one's opponent. We shall 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 8: 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.

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 that 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 (1999)-type 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 shall 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 9: Assuming that the set of mental states includes all inequality averse-type utility functions, there exists a unique mental subgame perfect 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 9 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 Implementing Effort with Mental Equilibrium

The concept of mental equilibrium has interesting implications in the context of contracting and incentive mechanisms. This section attempts to demonstrate this in a simple model of moral hazard. If emotions play a role in contractual environments, then a principal who attempts to implement a desirable outcome through a contract or a mechanism may wish to use mental equilibrium (rather then the standard Nash equilibrium) as the uderlying solution concept. To demonstrate the consequences of this approach, we shall use the following two-agent model that builds on Winter (2004), Winter (2006), and Winter (2009).

Two individuals cooperate on a project. Each individual is responsible for a single task. For the project to succeed, both individuals must succeed at their task. Players can choose to exert effort towards the performance of their task at a cost c which is identical for both 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 she observe the success of individual tasks. However, she is informed about the success of the entire project. An incentive mechanism 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 normal 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., she is looking for the least expensive 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 decisions are taken simultaneously. If agents move sequentially (assuming that the second player observes the effort decision of the first), then the optimal incentive mechanism pays $\frac{c}{1-\alpha}$to the second player, but the first player gets $\frac{c}{1-\alpha^2}$, which is less. Under this mechanism player 2 will exert effort if and only if player 1 does so. This generates an implicit incentive on the 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's effort, we would need to split each agent's task to n 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, then player 1 decides on the second sub-task, etc.) To keep the accounting in line, we have to set the cost of effort on each sub-task to be c/n, and the probability of success for each task (when no effort is exerted) to be $\alpha^{1/n}$. It can be shown that in this environment, when the number of sub-tasks (the value

of n) goes to infinity the optimal mechanism pays *both* players $\frac{c}{1-\alpha^2}$, which is what player 1 (the player whose effort is observable) gets in the standard sequential case [6]. In fact, the principal expenditure monotonically declines with the number of sub-tasks with the limit being $\frac{c}{1-\alpha^2}$. Intuitively, the larger n, the more agents have internal information about effort, the larger is the implicit incentive to exert effort and the less the principal has to expend to sustain effort. The equilibrium through which effort is being implemented with the optimal mechanism is one based on reciprocity. Each player continues to exert effort as long as his peer has done so as well. We shall now show that mental equilibrium implements effort with the same limit mechanism (i.e., a payoff of $\frac{c}{1-\alpha^2}$ to each agent) even when agents move simultaneously and have no feedback at all about each other's effort.

Roughly, the reciprocity that builds up in the sequential mechanism (with multiple sub-tasks) through the threat of shirking is sustained with a mental equilibrium in the simultaneous case through mental states under which players experience aversion to situations without reciprocity. Substantial experimental and empirical evidence points out that workers in real organizations are very much endowed with these kinds of mental states. They tend to tend to exert effort in response to effort by their peers also when such effort does not pay off, but are reluctant to exert effort when detecting shirking by their peers (see Ichino and Maggi (2000), Fischbacher, Gaechter, and Fehr (2001), Fehr and Falk (2002), Falk and Ichino (2006)).

Observation 10: The optimal mechanism for sustaining effort under mental equilibrium (in the Simultaneous Move game) 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 $j \neq i$'s best response is to imitate the action of player i (i.e., j prefers to exert effort if and only if i does so). We shall show that under $v = (\frac{c}{1-\alpha^2}, \frac{c}{1-\alpha^2})$, effort by both players is 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, thereby generating a Nash equilibrium that he prefers with respect to his material preferences, and 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's material payoff improves is the one in which player 1 shirks and player 2 exerts effort; this profile must be a Nash equilibrium in the new mental game with the preferences (u_1', u_2). This means that (1) must hold. But if (1) holds, player 2's mental state u_2 must be such that exerting effort is a dominant strategy. But this contradicts the property of u_2 as specified at the beginning of the proof. 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. Note first that for any reward $v < \frac{c}{1-\alpha^2}$, we have $v - c < v\alpha^2$. This means that the payoff for each agent is higher when both agents

[6] More specifically, for a fixed n the optimal mechanism pays $\frac{c}{1-\alpha^2}$ to the first mover and $\frac{c}{1-\alpha^{1+(n-1)/n}}$ to the other agent.

shirk than when both agents exert effort. This also implies that for $v < \frac{c}{1-\alpha^2}$, a player's maxmin value of the game is $v\alpha^2$. But this means that the effort outcome pays each player less than his maxmin value. Hence this outcome cannot be sustained as a mental equilibrium for such value of reward and the optimal mechanism must pay $\frac{c}{1-\alpha^2}$ to both agents. Note that under this mechanism both agents shirking and both agents exerting effort generate the same level of (material) payoff in the game.

6 n-Person Games

We started in Section 2 with a model in which players use only pure strategies (in and out of equilibrium). Mental equilibrium under this restriction has a simple structure and for many of the applications, particularly those which involve two-person games, such a model is adequate. This definition requires that no player be able to unilaterally deviate to a mental state under which his equilibrium outcome is improved. However, this implies that if the mental state with which player i deviates results in a game with no pure Nash equilibrium, then such deviation is not profitable. This in turn can give rise to artificial equilibria in games with more than two players that are sustainable by the mere fact that the resulting mental game has no pure-strategy Nash equilibrium. For games with four or more players it the set of equilibria expands to the extent that it loses its predictive power. This is demonstrated in Proposition 3 below. We shall therefore amend the definition of mental equilibrium to allow players' mental states to play mixed strategies. The choice of mental states will however remain pure. Before we move to the alternative model we remain with the benchmark model to establish two results for games with more than two players:

Proposition 3: For every normal form game G with $n \geq 4$, every strategy profile is a mental equilibrium (in the model of pure strategies).

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 shall 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 [7]. Indeed, each player who receives 0 can increase his payoff by changing his strategy from playing 0 to playing something else in T_i or if he is already playing a strategy in T_i he should switch to playing 0. By so switching the deviator will trigger a new profile s' for which $p(d(s')) \neq p(i)$ and he will raise his own 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 $(u_1, ...u_n)$ as it globally maximizes the payoff to all players. Furthermore, if player i deviates and sends a different mental state u_i' he will not be able to sustain a better equilibrium because the corresponding mental game will have no equilibrium. Regardless

[7] 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.

of what mental state player i plays, there will be at least one other mental state $j \neq i$ that deviates.

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 establish existence for all games in the benchmark model we need a separate argument for three-person games.

Proposition 4: Every three-person game has a mental equilibrium.

Proposition 4 implies that:

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

Proof of Proposition 4: 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 shall 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, $u_i(s) = 1$. Clearly, 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 will deviate from s' in the mental game. If instead $p(d(s'))$ is even, then player 2 will 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 players 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.

Mixed Strategies

We have seen that the model of mental equilibrium that restricts players to use only pure strategies breeds a plethora of equilibria to the extent that the concept can be uninformative. 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? As we noted earlier, 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 shall 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[8]. For each player

[8] Allowing the choice of mental state to be a mixed strategy will render the model

i, we denote by Δ_i the set of mixed strategies of player i.

Definition: A *mixed mental equilibrium* is a profile of mixed strategies[9] $x \in \Pi_{i \in N} \Delta_i$ such that the following two conditions are satisfied:

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

(2) There exist 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, 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 that is fully mixed. In this equilibrium both players assign a probability of 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 a contradiction arises 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 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. In the latter case, if there is a mixed equilibrium, player 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

intractable, as it will assume probability distributions over the continuum set of all utility functions. It will also unnecessarily expand the set of equilibria.

[9] The set of mixed strategies also includes all the pure strategies.

the original game for player 2. Hence, player 2 cannot profitably deviate in choosing his mental state.

Note that Proposition 2 does not apply for the concept of mixed mental equilibrium. The minmax payoff of player 1 is $1/2$ and it is 0 for player 2. Yet a strategy profile that pays each player $1/2$ is not a (mixed) mental equilibrium.

The fact that the set of mixed mental equilibrium does not contain the set of Nash equilibrium leaves the question of a general existence result for this concept open. It turns out that none of the standard fixed-point theorems are helpful because of non-convexities that arise from the flexibility of the choice of mental states[10]. Recently, Olschewski and Swiatczak (2009) used brute-force techniques to prove existence for all 2x2 games.

To demonstrate the advantage of the revised concept over the original one for large games we discuss the famous Public Good game, which is also the n-person version of the Prisoner's Dilemma. We shall show that the notion of mixed mental equilibrium is rather instructive for this game, no matter how large it is.

Example 7: The Public Good Game (Social Dilemma 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 Nash equilibrium in the game is $(0, ..., 0)$, but the profile that maximizes social welfare is $(1, ..., 1)$. Contrary to the Nash prediction, experimental evidence clearly shows a substantial contribution in the game, which depends on the number of players and the value of k (see Isaac, Walker, and Arlington (1994)).

Observation 11: 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

[10] We are grateful to Sergiu Hart for helping us clarify some technical issues on this.

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 is endowed with 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 5: 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 5 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 6: 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 6 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.

7 Collective Emotions

Some emotions tend to intensify when experienced within a group. When watching a comedy in a group people tend to laugh more than they would do when viewing it alone. Violent mob behavior is often a result of a collective rage which is experienced at a level that exeeds individual rage. In many strategic environments the benefits of emotional reactions, and in particular its usage as a commitment device, are enhanced when they are generated collectively with a group (often vis-à-vis outside players). We refer to this framewok as collective emotions. Wars, riots, and political campaigns are driven to a large extent by collective emotions. Collective mental states are generated through rituals, mass media, and education, all of which facilitate coordination among group members to improve the effectiveness and deterrence of a joint commitment. We point out that our approach here does not view the group as a unitary player. Players still "select" their own mental states. However, in contrast to our standard framework, where we assumed players' choices of

mental states and actions (as well as deviations) to be individual and independent, in our new framework we allow these choices to be collective and coordinated. The benchmark solution concept here (substituting for Nash equilibrium) is strong equilibrium à la Aumann (1959). We recall that a strong equilibrium is a Nash equilibrium in which no group of players can coordinate a joint deviation that would make all its members better off. This leads us to the concept of strong mental equilibrium.

For a normal form game G we denote by $SE(G)$ the set of strong Nash equilibria (à la Aumann (1959)) of the game G.

Let $G = (N, S, U)$ be a normal form game. A (pure) strategy profile s is a ***Strong Mental Equilibrium*** (SME), if there exists a vector of mental preferences $(u_1, u_2, ..., u_n)$ such that the following conditions are satisfied:

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

(2) There exist no coalition $T \subset N$ and mental preferences for the members of T denoted $u'_T = \{u'_j\}_{j \in T}$ such that for some strong equilibrium $s' \in SE(N, S, u'_T, u_{N \setminus T})$ we have $U_j(s') > U_j(s)$ for all $j \in T$.

It is easy to verify that:

Observation 12: A strong equilibrium of a game is a strong mental equilibrium and every strong mental equilibrium is a mental equilibrium.

Proof: Let $G = (N, S, U)$ with a strong equilibrium s. To show that s is an SME consider a profile of mental states: $u = (u_1, u_2, ..., u_n)$, which satisfies the following conditions: 1). s_i is a dominant strategy for player i and 2) $u_i(s) > u_i(s')$ for every player i and for every strategy profile $s' \neq s$. Clearly, s is a strong equilibrium in (N, S, u): any deviation by a coalition $T \subset N$ to a different profile will make all players worse off. Suppose now that a group of players T can choose an alternative profile of mental states u'_T such that for some $s' \in SE(N, S, u'_T, u_{N \setminus T})$ we have $U_j(s') > U_j(s)$ for all $j \in T$. Under $u_{N \setminus T}$ each player has a dominant strategy which is s_i. Hence if s' is a strong equilibrium of the new mental game it must be the case that $s_{N \setminus T} = s'_{N \setminus T}$. But then $U_j(s') > U_j(s)$ for all $j \in T$ contradicts the fact that s is a strong equilibrium in G.

The fact that a strong mental equilibrium is a mental equilibrium follows directly from the definitions of these concepts.

Observation 13 below implies that our concept above selects cooperation as the unique equilibrium outcome for the Prisoners' Dilemma.

Observation 13: In two-person games a strategy profile s is a strong mental equilibrium if and only if it is a Pareto undominated mental equilibrium.

Proof: Clearly, if s is not a mental equilibrium it cannot be a strong mental equilibrium as it implies that a singleton coalition can deviate to a different mental state and be made better off. Furthermore, if s is Pareto dominated by, say, s' it cannot be a strong mental equilibrium either. This is because by deviating to mental states for which s' are dominant strategies players 1 and 2 are both made better off relative to material preferences as the new mental game has a unique strong equilibrium, which is s'.Assume now that s is a Pareto undominated mental equilibrium; then consider the mental preferences (u_1, u_2) that support s as a mental equilibrium. It can be seen that no player can unilaterally deviate to a different mental state u'_i and be made better off. Also, since s is Pareto undominated (with respect to material preferences), there exist no joint deviations for players 1 and 2 to different mental states for which a new equilibrium will arise that pays both of them more than what they get

in s, which implies that s is a strong mental equilibrium.

Reflecting on the Prisoner's Dilemma again, we recall that the set of strong equilibria of the game is empty, while the set of mental equilibria selects both (D,D) and (C,C). Only the concept of strong mental equilibrium, which is the hybrid of the other two, uniquely selects cooperation as the equilibrium outcome of the game.

We can now revisit the effort model described in Section 5 and obtain full implementation of effort with respect to the new solution concept.

Observation 14: The optimal mechanism for sustaining effort as a *unique* strong mental equilibrium (in the Simultaneous Move game) is $v = (\frac{c}{1-\alpha^2}, \frac{c}{1-\alpha^2})$

Proof: To show that under v the outcome $(1,1)$ is a strong mental equilibrium we take the same mental preferences that sustain $(1,1)$ as a mental equilibrium. With these mental preferences it is enough to show that the two players cannot be simultaneously better off by jointly chosing different mental states. The only relevant pair of such mental states are those leading to the outcome $(0,0)$. This follows from the fact that $\frac{c}{1-\alpha^2}\alpha^2 \leq \frac{c}{1-\alpha^2} - c$. This also shows that $(0,0)$ cannot be a strong mental equilibrium as both players are better off deviating to mental states that will sustain the outcome $(1,1)$. Also $(0,1)$ and $(1,0)$ are not Strong mental equilibria as they do not guarantee players their individually rational payoffs (since $\frac{c}{1-\alpha^2}\alpha - c < \frac{c}{1-\alpha^2}\alpha^2$) So v uniquely implement $(1,1)$. To argue that this cannot be achieved by lower rewards simply recall that lower rewards cannot sustain $(1,1)$ as a mental equililibrium and therefore they cannot sustain it as a strong mental equilibrium.

8 Discussion

In his book *Politics* Aristotle makes the following observation about the emotion of anger: "Anyone can become angry—that is easy. But to be angry with the right person, to the right degree, at the right time, for the right purpose, and in the right way; this is not easy."

Anger, just like many other emotions, is an important component of strategic decision-making. In this paper we attempted to introduce a formal framework to discuss the role of emotions in strategic interactions using the concept of mental equilibrium. Two promising directions seem to suggest themselves at this stage:

1. The role of emotions in sequential interactions. Our concept was mainly applied to normal form games. We have proposed two polar versions of mental SPEs: one in which players commit to the same mental state for the entire game, and the other one which is based on allowing players to change their mental states from one stage to another. It would be interesting to investigate the role of a hybrid concept in which players can change their mental state during the course of the game but are not as flexible in changing it at each decision node. To capture the idea that emotions have a certain degree of persistence one can, for example, impose that players must commit to a mental state for a duration of k periods, or alternatively, that players must have the same mental states in every two subgames that are isomorphic. It would indeed be interesting to examine such a model in the context of sequential bargaining.

2. Emotions often trigger values and norms. One can often think of social norms as mental states that apply to a class of games. Put differently, norms may arise by having players commit themselves to the same mental state over multiple, possibly similar, games. This brings us back to Aumann's (2008) insight about the difference between "rule

rationality" and "act rationality." Our concept of mental equilibrium can lend itself to a formal model of rule rationality. Roughly, in a rule-rational equilibrium players are restricted to a small number of mental states, but they allocate these mental states to different games in a way that is "globaly" optimal relative to some distribution of the occurrence of these games through the course of life. The fact that players cannot freely change their mental state from one game to another facilitates the commitment device that can work in favor of their own material interests.

We hope to see both directions pursued, possibly by a different set of authors.

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

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

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

R	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.

References

[1] Aumann, Robert J., (1959). "Acceptable Points in General Cooperative n-Person

Games," in *Contributions to the Theory of Games IV*, Annals of Mathematics Study 40, Tucker, A. W. and Luce, R. D. (eds.), Princeton: Princeton University Press, pp. 287–324.

[2] Aumann, Robert J., (2008). "Rule Rationality vs. Act Rationality," Discussion Paper #497, Dec. 2008. The Center for the Study of Rationality, The Hebrew University.

[3] Berg, Joyce, Dickhaut, John, and McCabe, Kevin (1995). "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, 10, 122–142.

[4] 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.

[5] Bester, Helmut, and Sakovics, Jozsef (2001). "Delegated Bargaining and Renegotiation," *Journal of Economic Behavior & Organization*, 45(4), 459-473.

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

[7] Dekel, Eddie, Ely, Jeffrey C., and Yilankaya, Okan (2007). "Evolution of Preferences," *Review of Economic Studies*, 74(3), 685–704.

[8] Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L., Parvizi, J., Hichwa, R.D. (2000) "Subcortical and cortical brain activity during the feeling of self-generated emotions." *Nature Neuroscience* 3, 1049–1056.

[9] Falk, Armin, and Ichino, Andrea (2006). "Clean Evidence on Peer Effects," *Journal of Labor Economics*, 24(1), 39–57.

[10] Fershtman, Chaim, Judd, Kenneth L., and Kalai, Ehud (1991). "Observable Contracts: Strategic Delegation and Cooperation," *International Economic Review*, 32(3), 551–559.

[11] Fershtman, Chaim, and Kalai, Ehud, (1997). "Unobserved Delegation," *International Economic Review*, 38(4), 763–774.

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

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

[14] Fehr, Ernst, and Falk, Armin (2002). "Psychological Foundations of Incentives," *European Economic Review*, 46, 687–724.

[15] Fischbacher, Urs, Gaechter, Simon, and Fehr, Ernst (2001). "Are People Conditionally Cooperative? Evidence from a Public Good Experiment," *Economic Letters*, 71, 397–404.

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

[17] 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, pp. 23–34.

[18] Gueth, Werner, and Kliemt, Hartmut (1998). “The Indirect Evolutionary Approach: Bridging between Rationality and Adaptation,” *Rationality and Society*, 10, 377–399.

[19] 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.

[20] Ichino, Andrea, and Maggi, Giovanni (2000). “Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm,” *The Quarterly Journal of Economics*, 115(3), 1057–1090.

[21] Issac, R. Mark, Walker, James M., and Williams, Arlington W. (1994). “Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Very Large Groups,” *Journal of Public Economics*, 54, 1–36.

[22] McKelvey, Richard D, and Palfrey, Thomas R. (1992). “An Experimental Study of the Centipede Game,” *Econometrica*, 60 (4), 803–836.

[23] Nagel, R. and Tang, F. F. (1998). “An Experimental Study on the Centipede Game in Normal Form: An Investigation on Learning,” *Journal of Mathematical Psychology*, 42, 356–384.

[24] Olschewski, Guido and Swiatczak, Lukasz (2009). “Existence of Mental Equilibria in 2x2 Games,” mimeo, Handelshochschule Leipzig.

[25] Ochsner, K.N. and Gross, J.J. (2005). "The cognitive control of emotion." *Trends in Cognitive Science* 9 (5), 242–249.

[26] Phan, K. Luan, Tor Wager, Stephan F. Taylor, and Israel Liberzon (2002) "Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion
Activation Studies in PET and fMRI" *NeuroImage*, 16, 331-348.

[27] Rabin, Matthew (1993). “Incorporating Fairness into Game Theory and Economics,” *American Economic Review*, 83, 1281–1302.

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

[29] Tice, D.M. and Bratslavsky, E. (2000). "Giving in to Feel Good: The Place of Emotion Regulation in the Context of General Self-Control."
Psychological Inquiry 11, 149–159.

[30] Winter, Eyal (2004). “Incentives and Discrimination,” *American Economic Review*, 94(3) 764–773.

[31] Winter, Eyal (2006). “Optimal Incentives for Sequential Production,” *Rand Journal of Economics*, 37(2), 376–390.

[32] Winter, Eyal (2009). “Incentive Reversal,” *American Economic Journal: Microeconomics*, forthcoming.

[33] Winter, Eyal, Ben Shahar, Gershon, Aharon, Itzhak, Meshulam, Meir (2009). “Rational Emotions in the Lab,” Preliminary Notes, The Center for the Study of Rationality, The Hebrew University of Jerusalem.