Mental Equilibrium and Rational Emotions*

Eyal Winter, Luciano Méndez-Naya, Ignacio García-Jurado

October 28, 2013

Abstract

We model emotions as part of an equilibrium notion. In a mental equilibrium each player "selects" an emotional state that 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 to the emotional states of the others (in the sense of maximazing material payoffs). We discuss the concept behind the definition of mental equilibrium and examine it in the context of some of the most popular games discussed in the experimental economics literature. We shall demonstrate the role of mental equilibrium in incentive mechanisms and 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 behavioral economists. There are two sources of rationality inperfectness 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

*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. Ignacio García-Jurado acknowledges the financial support of Ministerio de Ciencia e Innovación through projects ECO2008-03484-C02-02 and MTM2011-27731-C03, and of *Xunta de Galicia* through project INCITE09-207-064-PR.

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 past two decades several interesting and important models have 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 re-evaluate 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 to 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 anger, 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[1] 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,

[1] 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.

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 of 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 not unreasonable to assume that human emotions like fairness, anger, envy, and revenge, which play a role in many game situations, have been partially developed through an evolutionary process to increase individuals' fitness to the social environment in which they live. Our equilibrium concept can be viewed as a theoretical foundation for this feature. 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 process. 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 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[2]. To further explain this point we suggest a thought experiment that demonstrates how emotions are triggered by incentives. Imagine that you are informed at the airport that your flight has been cancelled 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

[2]A considerable 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.

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 that are global and robust, while under the rational emotions interpretation they can be specific and fragile. However, we shall be subscribing to both interpretations and will not argue in favour 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 games vaguely defined and 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 rational emotions interpretation 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 made from a self-centered point of view, the equilibrium choice of preferences may give rise to non-trivial 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. An important part of our analysis would be to identify "non-emotional" games, i.e., games in which all mental equilibria can be sustained by material preferences only. We show that all zero-sum games are non-emotional and we characterize the entire class of non-emotional games using the concept of "Stackleberg strategies" that appear in the literature on repeated games and reputation (e.g., Mailath and Samuelson 2006).

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 (just like the Nash equilibrium). Nevertheless, there are two grounds on which this assumption can be justified. Firstly, a player's choice of mental states

involves some sort of pre-play communication game that we intentionally leave unspecified. A players signals his 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 while our ability to identify the mental state of the other is inperfect, our ability to decieve is imperfect as well. In Section 11 we bring some empirical evidence to this effect and analyse a model of noisy detection of mental states. 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 signalling models are beyond the scope of this paper.

The relevance 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 discuss some 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. 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 terms of motivation, interpretation, and formal modelling. 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 modelling 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's (1990) 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 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 cooperation in a class of cooperation games that include the Prisoner's Dilemma. We show that reciprocity seeking preference are both necessary and sufficient to sustain cooperation in a mental equilibrium for these games. Section 4 deals with "non-emotional" games. In this section we provide a characterization

of those two-person games in which all mental equilibria can be supported by material preferences. Sections 5 and 6 are dedicated to discussing the role of mental equilibrium in two games that have been prominently discussed in the experimental economics literature: the Trust game and the Ultimatum game. In Section 5 we derive the mental equilibria for these games without restricting the set of preferences. By contrast, in Section 6 we restrict the domain of mental states to include only preferences of inequality aversion, and show how the level of aversion is determined endogenously.

We devote Section 7 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 arises from the threat of retaliation by peers, when transparency is available, is internalized at the level of mental states even when no transparency is available.

Sections 8 and 9 deal with a model of mental equilibrium in which mental states can use mixed strategies. This model is motivated in Section 8 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 equilibria. 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 mixed 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 10 we discuss collective emotions. These emotions emerge when a group of players coordinate their mental state 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 unlike anything else in the plethora of game-theoretic solution concepts. Section 11 studies a variant of mental equilibrium that builds on the idea that the detection of others' emotions is imperfect. We discuss this concept in the context of the famous TV game "Split or Steal", which is itself a variant of the Prisoner's Dilemma. We refer to the empirical observation of "mind reading" according to which pre-play comunication results with correlated actions in the Prisoner's Dilemma, and show that our concept of mental equilibrium can explain this correlation.

We conclude in Section 12.

2 Basic Definitions

In the first part of this paper we shall assume that 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 1 *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 do not exist a 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 j 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 of the set of mental equilibria in two-person games which will prove to be useful for various applications. In any Nash equilibrium each player attains at least his minmax value. Proposition 1 asserts that this property is both a necessary and sufficient condition for mental equilibria in two-person games. For the formal result let $m_i = \max_{s_i} \min_{s_j} U_i(s_i, s_j)$, where $i, j \in \{1, 2\}$, $i \neq j$, be the minmax value of player i, which we assume to exist (the existence is guaranteed for finite strategy sets).

Proposition 1 *Let G be a two-person game; then $s \in S$ is a mental equilibrium if and only if $U_i(s) \geq m_i$.*[3]

Proof. Let v_1 and v_2 be the minmax values of players 1 and 2 respectively, with s_1 and s_2 being minmax strategies.[4] 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 minmax 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 minmax 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 minmax 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 minmax 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

[3] In the Appendix we show that Proposition 1 does not apply to three-person games; indeed it is shown there that, neither of the two directions of the proposition holds true.

[4] s_i is said to be a minmax strategy of player i if $m_i = \min_{s_j} U_i(s_i, s_j)$

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 be 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 equilibria 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 minmax value. To show this, let s_1' be a minmax 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 strategy to s_1'. Clearly, (s_1', s_2') is the desired profile. Player 1 gets paid at least his minmax value 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 value. □

Our definition of mental equilibrium relied on the assumption that players are "optimistic" when contemplating deviations as it is enough that there exists at least one equilibrium in the new mental game (after player i deviates) that player i prefers to the original (putative) equilibrium in order to trigger him to deviate. A more stringent condition on deviations would require that player i deviate only if all equilibria of the new 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 2 *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 do not exist a player i and a mental state u_i' such that $NE(N, S, u_i', u_{-i}) \neq \varnothing$ and for every equilibrium $s' \in NE(N, S, u_i', u_{-i})$ it holds that $U_i(s') > U_i(s)$.*

Clearly, every mental equilibrium is a weak mental equilibrium, but we shall argue that for two-player games the two solution concepts coincide.

Proposition 2 *In two-person games the set of mental equilibria and the set of weak mental equilibria coincide.*

Proof. 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 minmax value, and that s^* is a weak mental equilibrium supported by the mental states $u = (u_1, u_2)$. Let s_1 be a minmax 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 minmax 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. □

Our objective now is to investigate the mental states that support mental equilibria. We will start with the prominent game of the Prisoner's Dilemma.

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	$1, 1$	$5, -4$
C	$-4, 5$	$4, 4$

Observation 2 *There are two mental equilibria in the Prisoner's Dilemma game,* (C, C) *and* (D, D).

Proof. It is easy to check that the minmax vector in this game is $v = (1, 1)$. By Proposition 1, (C, C) and (D, D) are mental equilibria but (D, C) and (C, D) are not, as they pay less than the minmax value to one of the players. □

It can be easily verified that the outcome (C, C) can be supported as a mental equilibrium through the following mental states: $u_1(C, D) = u_2(C, D) = u_1(D, C) = u_2(D, C) = -10$, and $u_i = U_i$ otherwise. Note that these mental preferences represent reciprocity seeking individual; 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 of a larger class of games which we call *Cooperation games*. A Cooperation game is a two-person game with two strategies $\{D, C\}$ (defection and cooperation) for each player such that (1) each player's best response to cooperation by the other player is to deffect and (2) cooperation by both players dominates defection by both players. More formally: $U_1(D, C) > U_1(C, C)$, $U_2(C, D) > U_2(C, C)$, and $U_i(C, C) > U_i(D, D)$,

$i = 1, 2$. Every Prisoner's Dilemma game is a Cooperation game but the set of Cooperation games includes also all Chicken games.

In Observation 3 we restrict ourselves to generic mental states. A mental state is generic if the corresponding player never displays indifferences.

Observation 3 *Let $G = (N, S, U_1 U_2)$ be a Cooperation game. Then (C, C) is a mental equilibrium. Furthermore, a necessary and a sufficient condition for the generic mental states (u_1, u_2) to sustain (C, C) as mental equilibrium in G is: $u_1(C, C) > u_1(D, C)$, $u_1(D, D) > u_1(C, D)$, and $u_2(C, C) > u_2(C, D)$, $u_2(D, D) > u_2(D, C)$.*

Proof. Consider first mental states (u_1, u_2) that satisfy the conditions above. First note that (C, C) is a Nash equilibrium in the mental game defined by the mental preferences. Consider different mental states for player 1. In order for 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 equilibria under the payoff functions (u_1', u_2); but this is impossible because $u_2(D, D) > u_2(D, C)$. Since an analogous argument can be developed for player 2, we conclude that (C, C) is a mental equilibrium supported by (u_1, u_2). Consider now any profile of mental preferences (u_1, u_2) which sustains (C, C). 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 inequality failed to hold and player 2 would deviates if the third one failed). Suppose that the second inequality is violated; then the mental state of player 2 is not optimal, as player 2 is better off (in terms of material preferences) with the mental state u_2',which satisfies $u_2'(C, D) > u_2'(C, C)$, as under (u_1, u_2') the outcome (C, D) is a Nash equilibrium. Likewise if the fourth inequality failed to hold, then player 1 would be better off deviating to u_1' with $u_1'(D, C) > u_1'(C, C)$ and increasing 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 (generically) 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 at least for some Prisoner's Dilemma games.

Starting with altruism, consider the Prisoner's Dilemma given by:

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

We argue that mental preferences that sustain 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 preferences would result in the following mental game:

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

For (C, C) to be an equilibrium in this mental game we need to have $4(\alpha_2 + \beta_2) \geq 5\alpha_2$. This inequlity implies, $5\beta_2 \geq \alpha_2 + \beta_2$. But this means that player 1 is better off in terms of material payoffs if he adopts the mental state with $u_1 = U_1$. With such a mental state he will be able to sustain (D, C) as an equilibrium, which is the best pssible outcome in terms of materail payoffs.

Note the difference between the preference given by $u_i = \alpha_i U_i + \beta_i U_j$ and the one we used in Observation 3. 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 for reciprocity seeking behavior. These mental preferences 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) 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 a greater (material) payoff for player 1, and involves a greater equality than the outcome (C, C). Hence, given our discussion above, there exists no profile of (inequality averse) mental preferences which supports (C, C) as a mental equilibrium in this Prisoner's Dilemma game.

We conclude that reciprocity-seeking preferences can explain cooperation in every Prisoner's Dilemma game, but altruism, spitefulness, or inequality aversion cannot.

4 Non-Emotional Games

As we have seen, cooperation in the Prisoner's Dilemma is sustained through mental states that represent reciprocity. Players are therefore required to depart from their material preferences in order to sustain cooperation as a mental equilibrium. In this sense the Prisoner's Dilemma induces emotional behavior. Do all two-person games induce emotional behavior? Clearly

games in which all mental equilibria can be supported by material preferences do not induce emotional behavior. We refer to such games as *non-emotional games*. In a non-emotional game all players can play according to their selfish and material payoffs in every mental equilibrium. It implies in particular that commitment plays no role in such games. In this section we shall characterize this class of games.

Let $G = (N, S, U)$ be a two-person game for which the following two vectors $m, M \in \mathbb{R}^2$ are well defined:

- m is the minmax vector; i.e., $m_i = \max_{s_i \in S_i} \min_{s_j \in S_j} U_i(s_i, s_j)$, where $i, j \in \{1, 2\}$, $i \neq j$.
- M is the vector of Stackleberg values for the two players, i.e., it pays each player the maximal payoff under the assumption that the other player will best respond to his action. Formally, for $i, j \in \{1, 2\}$, $i \neq j$, and all $s_i \in S_i$, define

$$B_j(s_i) = \{s_j \in S_j \mid U_j(s_i, s_j) \geq U_j(s_i, \tilde{s}_j),\ \forall \tilde{s}_j \in S_j\}$$

  and
  $M_i = \max_{s_i \in S_i} \max_{s_j \in B_j(s_i)} U_i(s_i, s_j)$, where $i, j \in \{1, 2\}$, $i \neq j$.

Notice that m and M are well defined in games with finite sets of strategies, and $M \geq m$. Furthermore, $M = m$ whenever the game is zero-sum. We can now establish the following result.

Proposition 3 *Let $G = (N, S, U)$ be a two-person game for which the two vectors $m, M \in \mathbb{R}^2$ are well defined. Then G is a non-emotional game if and only if the following condition holds for every $s \in S$:*

$$U(s) \geq m \ \Rightarrow\ U(s) \geq M \text{ and } s \text{ is a Nash equilibrium of } G. \qquad (1)$$

Proof. Let s be a mental equilibrium. By Proposition 1 s satisfies $U(s) \geq m$, and thus by (1) s is a Nash equilibrium with respect to the mental preferences u given by $u = U$. To show that $u = U$ supports s as a mental equilibrium consider a deviation by one player to an alternative mental state say u_i'. Let s' be an equilibrium of the resulting mental game. By assumption $U_i(s) \geq M_i \geq U_i(s')$. So $u = U$ satisfies the second condition of the definition of mental equilibrium. Hence, all mental equilibria of G are supported by material preferences. Conversely, assume that (1) does not hold. Then, there exists $s \in S$ with $U(s) \geq m$ and such that $U(s) \ngeq M$ or s is not a Nash equilibrium of G. Since $U(s) \geq m$, then by Proposition 1 s is a mental equilibrium. If it is not a Nash equilibrium then it obviously cannot be supported by material preferences. Assume by way of contradiction that s is a Nash equilibrium; then $U(s) \ngeq M$, which means that $U_i(s) < M_i$ for an $i \in \{1, 2\}$. Thus, s cannot be supported by material preferences; to prove it notice that i can choose mental state u_i' given by:

- $u_i'(\hat{s}_i, s_j') = 1$, for all $s_j' \in S_j$, if $\max_{s_j \in B_j(\hat{s}_i)} U_i(\hat{s}_i, s_j) = M_i$,
- $u_i'(s_i', s_j') = 0$, for all $s_j' \in S_j$, if $\max_{s_j \in B_j(s_i')} U_i(s_i', s_j) < M_i$.

Clearly $(N, S, (u_i', U_j))$ has a Nash equilibrium offering i a payoff $M_i > U_i(s)$. □

Note that the condition specified in Proposition 3 applies for zero-sum games. Hence all zero-sum games are non-emotional. This is a rather intuitive observation. In zero-sum games commitments play no role whatsoever. If by committing himself to a certain mental state player 1 can get more than his minmax value, this means that player 2 cannot guarantee his minmax value, which is a contradiction.

5 Mental Equilibrium in Trust and Ultimatum Games

We shall now discuss the role of mental equilibrium in two games that are prominently discussed in the experimental economics literature: the Trust game and the Ultimatum game.

Example 2 *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 $x - y + z$ and the payoff for player 2 is $3y - z$.*

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

Proof. Consider such an outcome (a_1, a_2). Since $a_1 \geq x$ player 2 can guarantee that player 1 gets no more than a_1. This can be done by transferring no money back to player 1 if player 2 received any money from player 1. Furthermore, 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 minmax value, which contradicts Proposition 1. □

We note that the Trust game has a unique Nash equilibrium in which player 1 makes a zero transfer to player 2. Observation 4 suggests that any level of trust displayed by player 1 coupled with a level of trustworthiness

that compensates player 1 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.

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

Example 3 *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* $(0.5, 0.5)$.

To discuss the concept of mental equilibrium for this game we first need to discuss the subgame perfect version of mental equilibrium. We shall show that this concept has little bite if the set of mental states is allowed to include all possible utility functions. This 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 2-person extensive form game $G = (T, U)$ with perfect information, where T is the game form defined by a tree, and $U = (U_1, U_2)$ are payoff functions for players $1, 2$ assigning a payoff vector to each to terminal node of the game. We denote by $SPE(G)$ the set of subgame perfect equilibria of the game G.

Definition 3 *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. $s \in SPE(T, u)$.
2. *There exist no player* i, *mental state* u_i', *and strategy profile* $s' \in SPE(T, u_i', u_{-i})$ *with* $U_i(s') > U_i(s)$.

As mentioned above, when the set of mental states from which players can choose is not restricted, the concept of mental subgame perfect equilibrium does not have much predictive power.

Observation 5 *Take any two-person extensive form game with perfect information G. Every Nash equilibrium outcome of G is a mental subgame perfect equilibrium outcome of G.*

Proof. Let s be a Nash equilibrium of the game. We construct the following mental (extensive form) game. For player i ($i = 1, 2$) 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 player $j \neq i$ must have subgame perfect equilibrium strategies that are consistent with their s_j. Let s_i' be the subgame perfect equilibrium 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 5 and of the fact that a mental subgame perfect equilibrium is not an effective refinement of Nash equilibria. 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 subgame perfect equilibrium 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.

6 Restricting the Set of Mental States

Observation 5 implies that without restricting the set of mental states all allocations of the unit of goods between the two players are sustainable as a mental SPE of the Ultimatum game, since the set of Nash equilibrium outcomes covers the entire set of allocations. We now wish to confine our attention to mental states that display inequality aversion as characterized by Fher and Schmidt's (1999) model. We 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 $\alpha_i \geq \beta_i$. 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^* and satisfying that $\alpha_i^* \geq 1$, 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 6 *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 for player 1 an optimal mental state must satisfy $\beta_1 = 0$. Furthermore, player 1 cannot gain by setting $\alpha_1 > 0$ either. This is due to the fact that player 1 as first mover in the game cannot benefit from a commitment, and thus without loss of generality this player's best mental state must correspond to the material preferences. Consider now player 2 and suppose that player 1 offers the mental state of player 2 a payoff of less than 1/2. 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 $x_2 \geq \frac{\alpha_2}{1+2\alpha_2}$ (notice that $x_1 + x_2 = 1$); the fact that the right-hand side is increasing in α_2 and that the mental state of 1 can be assumed to be 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 subgame perfect

equilibrium 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 present 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, that 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 7 *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, being the first mover cannot benefit from a mental state that is different from his material preferences because (as argued earlier) first movers cannot gain from a commitment device. This implies again, $\alpha_1 = \beta_1 = 0$. The receiver's optimal mental state should have an inequality aversion parameter β_2 large enough to incentivize a rational sender (driven by material preferences) to transfer his entire endowment to the receiver. Suppose that under such a large β_2 the receiver returns r to the sender. Suppose that the sender indeed transfers $3x$ to the receiver. Since $\beta_2 \leq 1$ the receiver after repaying back to the sender will still hold at least as much as the sender sender. The receiver's utility will be $3x - r - \beta_2(3x - 2r)^+$ (α_2 does not appear in the expression because the receiver's payoff is larger than the sender's payoff). If $\beta_2 < 1/2$, the receiver's optimal payback is $r = 0$, and the sender has no incentive to transfer anything to the receiver. On the other hand, if $1/2 \leq \beta < 1$, the receiver's optimal payback would equalize the payoffs of the two players, i.e., $r = \frac{3}{2}x$. Hence, under $\beta_2 < 1/2$ the sender transfers his entire endowment to the receiver and gets back half of the tripled revenue. Since this is the best possible outcome for the receiver it is also the subgame perfect equilibrium outcome. □

Interestingly, Observation 7 shows how the level of inequality aversion is determined endogenously. In equilibrium the receiver's mental state must have β_2 between 1/2 and 1.

7 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 an incentive mechanism may wish to use mental equilibrium (rather then the standard Nash equilibrium) as the underlying 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 he observe the success of individual tasks. However, he 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 an incentive 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., he 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.

[5] 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 is, the more agents have internal information about effort, the larger 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 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 8 *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. The simultaneous move game is the following two-person game.

	1	0
1	$v_1 - c, v_2 - c$	$v_1\alpha - c, v_2\alpha$
0	$v_1\alpha, v_2\alpha - c$	$v_1\alpha^2, v_2\alpha^2$

It is easy to see that the minmax vector is

$$m = (\max\{v_1\alpha - c, v_1\alpha^2\}, \max\{v_2\alpha - c, v_2\alpha^2\}).$$

Clearly, $v_i - c \geq v_i\alpha - c$. Besides, $v_i - c \geq v_i\alpha^2$ if and only if $v_i \geq \frac{c}{1-\alpha^2}$. Hence, in view of Proposition 1, it is clear that $(1, 1)$ is a mental equilibrium under the mechanism $(\frac{c}{1-\alpha^2}, \frac{c}{1-\alpha^2})$. Now, if $v_i < \frac{c}{1-\alpha^2}$, then $v_i - c < v_i\alpha^2 \leq m_i$, so $(1, 1)$ is not a mental equilibrium (again because of Proposition 1). □

8 n-Person Games

[5] 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.

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 deviate unilaterally 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 a 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 the set of equilibria expands to the extent that it loses its predictive power. This is demonstrated in Proposition 4 below. In Section 9 we shall therefore amend the definition of mental equilibrium to allow players' mental states to play mixed strategies. Before we move to the alternative model let us return to the benchmark model to establish two results for games with more than two players.

Proposition 4 *For every normal form game G with $n \geq 4$, every strategy profile is a mental equilibrium.*

Proof. For each player i we select one strategy and denote it by 0. We denote by T_i the set of the remaining strategies so that $S_i = T_i \cup \{0\}$. We 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. [6] 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 0 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

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

mental game will have no equilibrium different from $(0, ..., 0)$. Regardless 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 5 *Every three-person game has a mental equilibrium.*

Proof. 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 other 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. □

Corollary 1 and Propositions 4 and 5 imply that:

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

9 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.[7] We will consider here only normal form games with finite sets of strategies for each player. For each player i, we denote by Δ_i the set of mixed strategies of player i.

Definition 4 *A mixed mental equilibrium is a profile of mixed strategies*[8] $x \in \Pi_{i \in N} \Delta_i$ *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, mental state u_i', and 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, the following example shows that a mixed Nash equilibrium may not be a mixed mental equilibrium.

Example 4 *Consider the following two-person game:*

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

The unique Nash equilibrium of this game is $((1/2, 1/2), (1/2, 1/2))$, which is a fully mixed profile . It provides both players the payoff of $1/2$. Suppose by way of contradiction that it is a mixed mental equilibrium and let the following game be a mental game supporting it:

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

[7] Allowing the choice of mental state to be a mixed strategy will render the model intractable, as it will assume probability distributions over the continuum set of all utility functions. It will also unnecessarily expand the set of equilibria.

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

For the strategy profile $((1/2, 1/2), (1/2, 1/2))$ *to be a Nash equilibrium in the mental game the following two conditions must hold:*

1. $a_1 = a_3$ *and* $a_2 = a_4$, *or* $a_1 > a_3$ *and* $a_2 < a_4$, *or* $a_1 < a_3$ *and* $a_2 > a_4$.
2. $b_1 = b_2$ *and* $b_3 = b_4$, *or* $b_1 > b_2$ *and* $b_3 < b_4$, *or* $b_1 < b_2$ *and* $b_3 > b_4$.

Now consider all the possible cases.

- *If* $a_1 = a_3$ *and* $a_2 = a_4$, $b_1 \geq b_2$ *and* $b_3 \leq b_4$, *then* $(top, left)$ *is a Nash equilibrium in the mental game providing a payoff vector of* $(1, 1)$ *in the original game. This is impossible (it contradicts that the mental game supports* $((1/2, 1/2), (1/2, 1/2))$*).*
- *If* $a_1 = a_3$ *and* $a_2 = a_4$, $b_1 < b_2$ *and* $b_3 > b_4$, *then* $(top, right)$ *is a Nash equilibrium in the mental game providing a payoff of* 2 *to player two in the original game. This is impossible.*
- *If* $a_1 > a_3$ *and* $a_2 < a_4$, $b_1 \geq b_2$ *and* $b_3 \leq b_4$, *then* $(top, left)$ *is a Nash equilibrium in the mental game providing a payoff vector of* $(1, 1)$ *in the original game. This is impossible.*
- *If* $a_1 > a_3$ *and* $a_2 < a_4$, $b_1 < b_2$ *and* $b_3 > b_4$, *then player two is better off replacing his mental state with one in which the left strategy is dominant because then* $(top, left)$ *is a Nash equilibrium in the new mental game providing player two a material utility of* 1. *This is impossible.*
- *If* $a_1 < a_3$ *and* $a_2 > a_4$, $b_1 \leq b_2$ *and* $b_3 \geq b_4$, *then* $(top, right)$ *is a Nash equilibrium in the mental game providing a payoff of* 2 *to player two in the original game. This is impossible.*
- *If* $a_1 < a_3$ *and* $a_2 > a_4$, $b_1 > b_2$ *and* $b_3 < b_4$, *then player one is better off replacing his mental state with one in which the top strategy is dominant because then* $(top, left)$ *is a Nash equilibrium in the new mental game providing player one a material utility of* 1. *This is impossible.*

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

1,0	*1,0*
1,0	*2,0*

where (top,left) is an equilibrium. Clearly, player 1 has no incentive to

change his mental state since 1 is the highest (material) payoff he can get. Consider the other player. Suppose player 2 chooses a different mental state. Since bottom is a weakly dominant strategy for player 1, any new mental preferences that would make player 2 play right with positive probability will trigger (the mental state of) player 1 to play bottom with probability 1 and player 2's material payoff will be less than 1. Hence, player 2 cannot deviate to a different mental state and increase his material payoff.

Note that Proposition 1 does not apply for the concept of mixed mental equilibrium. The minmax value 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 equilibria does not contain the set of Nash equilibria 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.[9] 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 5 *The Public Good game (Social Dilemma game). $n > 1$ 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 9 *A strategy profile in the Public Good game is a mixed 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.

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

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 in which choosing 0 is 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 0. If the proportion of contributors is less than $\frac{1}{k}$, then $w > \frac{krw}{n}$ and player i is better off deviating. If the equilibrium is not $(0, ..., 0)$, then with positive probability some players contribute in the equilibrium of the new mental game and the expected equilibrium payoff of player i is greater than 0, which makes deviation even more attractive. We now show that a profile with a proportion of contributors $p \geq \frac{1}{k}$ is a mental equilibrium. Consider such a profile and denote by T the set of players who choose 0 and by $N-T$ the players who choose 1. To show that this profile is a mental equilibrium we assign the following mental states to players. For each player in $N-T$ we assign a mental state that prefers to choose 1 if and only if the proportion of agents who choose 1 is at least p (otherwise he prefers to choose 0). For each player in T we assign a mental state whose preferences are identical to those of the other players (i.e., choosing 0 is a dominant strategy). Given this set of mental states it is clear that the underlying strategy profile is an equilibrium of the mental game. It therefore remains to show that condition (2) in the definition of mixed 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 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 0, and then the payoff to i is less than or equal to zero, which contradicts condition (2). 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 is better off by assigning a different mental state. □

The attractive property of mixed 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.

In the n-person public good game, strategy profiles with a small group of contributors are nnot mental equilibira. The reason why $(1, 0, ..., 0)$ for example cannot be a mental equilibrium outcome is that it pays player 1 less than his minmax value. This observation can be generalized.

Let G be a finite n-person game. For a mixed strategy profile x, denote by $U_i(x)$ 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 $a_i = \max_{x_i \in \Delta_i} \min_{x_{-i} \in \Pi_{j \in N \setminus \{i\}} \Delta_i} U_i(x_1, ..., x_n)$.

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

Proof. Suppose that $x = (x_1, ..., x_n)$ is a mixed mental equilibrium with $U_i(x) < a_i$. 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 that 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 $x_i^0 = \arg\max_{x_i \in \Delta_i} \min_{x_{-i} \in \Pi_{j \in N \setminus \{i\}} \Delta_i} U_i(x_1, ..., x_n)$, and let $G^*_{x_i^0}$ be the game defined on the set of players $N \setminus \{i\}$ such that $U_j^{x_i^0}(x_{N \setminus \{i\}}) = U_j(x_i^0, x_{N \setminus \{i\}})$. 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 $N \setminus \{i\}$ 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 $U_i(x_i^0, z) \geq a_i$, which contradicts the assumption that x is a mixed mental equilibrium. □

Note that Example 4 implies that the converse of Proposition 6 is not true. The Nash equilibrium of the game (which is not a mental equilibrium) yields a payoff vector of $(\frac{1}{2}, \frac{1}{2})$, which exceeds the minmax vector $(1/2, 0)$.

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

Proof. Follows directly from the proposition above. □

We conclude with another useful property of mental equilibrium.

Proposition 7 *Let G be a finite 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 $C = (C_1, C_2, ..., C_n)$ be a profile of mental states supporting s as a mental equilibrium. By supposition s is a Nash equilibrium of (N, S, C). Since both s and s' are pure strategy profiles we can rename strategies for each player so that C' is isomorphic to C up to strategy names and such that s' is a Nash equilibrium of (N', S', 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 $(N', S', (C_i^*, C_{-i}^{'}))$ there exists another Nash equilibrium s^* with $C_i(s^*) > C_i(s^{'})$. But the isomorphism between C and C' implies that there is a mental state of player i $\overline{C}_i$ such that s^* is a Nash equilibrium of the game $(N, S, (\overline{C}_i, C_{-i}))$ with $C_i(s^*) > C_i(s^{'}) \geq C_i(s)$, which contradicts the fact that s is a mental equilibrium. □

10 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 that 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.[10] 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.

Definition 5 *Let $G = (N, S, U)$ be a normal form game. A strategy profile s is a* strong mental equilibrium*, 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 Nash equilibrium $s' \in NE(N, S, u'_T, u_{N \setminus T})$ we have $U_j(s') > U_j(s)$ for all $j \in T$.*

Note that condition 2 requires that a joint deviation by a group of players to a different profile of mental states cannot produce a Nash equilibrium in which all the members of the group are better off. The reason for referring to *Nash* instead of *strong* equilibrium here is twofold. Firstly, from a conceptual point of view once a coalition of players deviates it is less reasonable to assume that players will continue to coordinate their actions in the new mental game. Secondly, and more importantly, by allowing only for deviations that improve the players' payoffs through a strong equilibrium we are limiting the scope of deviations and, thus, expanding the set of equilibria to a point where the concept becomes uninformative. It can be shown, for example, that if we replace Nash equilibrium by strong equilibrium in condition 2 of the definition we will be able to sustain even outcomes that are individually not as rational as strong mental equilibrium.

It is clear that every strong mental equilibrium is a mental equilibrium; it follows from Definitions 1 and 5 and from the fact that every strong equilibrium is a Nash equilibrium. We can also prove the following result.

Observation 10 *A strong equilibrium of a game is a strong mental equilibrium.*

Proof. Let $G = (N, S, U)$ with a strong equilibrium s. To show that s is an strong mental equilibrium consider a profile of mental states $u = (u_1, u_2, ..., u_n)$, that satisfies the following conditions: 1) s_i is a strictly dominant strategy for player i [11] 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 $s' \in NE(N, S, u'_T, u_{N \setminus T})$ we

[10] Formally, $s \in S$ is a strong equilibrium of $G = (N, S, U)$ if there do not exist a non-empty $M \subset N$ and $s'_M \in S_M$ with $U_i(s'_M, s_{-M}) > U_i(s)$ for all $i \in M$.

[11] In the sense that $u_i(s_i, \hat{s}_{-i}) > u_i(s'_i, \hat{s}_{-i})$ for all $\hat{s}_{-i} \in S_{-i}$ and all $s'_i \in S_i$.

have $U_j(s') > U_j(s)$ for all $j \in T$. Under $u_{N\setminus T}$ each player has a dominant strategy that is s_i. Hence if s' is a Nash 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. □

In two-person games we can prove the following characterization of the strong mental equilibrium.

Observation 11 *In two-person games a strategy profile s is an strong mental equilibrium if and only if it is a Pareto undominated mental equilibrium.*[12]

Proof. Assume that s is a Pareto undominated mental equilibrium. Consider the mental preferences (u_1, u_2) that support s as a mental equilibrium as built in the proof of Proposition 1. It is clear that s is a strong equilibrium in the mental game. Moreover, 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 there exists a new equilibrium paying both of them a higher material payoff. This implies that s is a strong mental equilibrium. For the converse, if s^* is an strong mental equilibrium, then as argued before it is also a mental equilibrium. Furthermore, s^* is Pareto undominated. Otherwise, if s' dominates s^*, then the two players can deviate to a joint mental state $u' = (u'_1, u'_2)$ with $u'(s) < u'(s^*)$ for all s. □

Reflecting on the Prisoner's Dilemma again, we recall that the set of strong equilibria of the game is empty. The set of Nash equilibria includes only the outcome (D, D) while the set of mental equilibria contains both (D, D) and (C, C). Interestingly,

Observation 12 *(C,C) is the unique strong mental equilibrium of the Prisoner's Dilemma.*

Proof. (C, C) is the unique Pareto undominated mental equilibrium and hence by Observation 11 it is the unique strong mental equilibrium. □

11 Mental Equilibrium and Mind Reading

"Split or Steal" is a popular TV game in the UK ("Friend or Foe" in the US version), whose final stage involves a simplified Prisoner's Dilemma game. The two competing individuals who aquired jointly a substantial sum of money (sometimes more than 100,000 pounds) in a preliminary stage (by

[12] A strategy profile $s \in S$ is Pareto undominated in $G = (N, S, U)$ if there does not exist $s' \in S$ with $U_i(s') > U_i(s)$ for all $i \in N$.

solving trivia problems) are called to play the following game. Each of the two participants faces two balls. One ball has "split" written inside and the other has "steal" written inside. Each of the participants has to choose one of the balls (after observing privately which is the "split" ball and which is the "steal" ball). If both players choose the "split" ball they share the underlying amount of money 50:50. If one chooses the "split" ball and the other chooses the "steal" ball, the one choosing steal obtains all the money while the other obtains nothing. If they both choose "steal" they both obtain nothing. Before the players make their decision they engage in a 30 second face to face discussion that is completely non-binding. Data based on the "Split or Steal" game and "Friend or Foe" have been studied by several authors. One remarkable finding concerns the high correlation between players decisions. This correlation is generated by pre-play communications. Generating empirical distributions over the four strategy profiles of the game from the collected data, Kalay et al (2003) discover a significant correlation in the choices of players in the Friend or Foe game. In this section we develop a variant of mental equilibrium to explain this correlation. The main feature of this variant is the assumption that the detection of others' mental states is imperfect. We shall again assume that players play the Prisoner's Dilemma game given in Example 1. We restrict the set of mental states in this section to include only two elements: $\{Ra, Re\}$. Ra (Rational) refers to a mental state that represents self-interest. Under Ra the player chooses D regardless of the signal he receives. A player endowed with Re (Reciprocal) chooses C if and only if he received a signal of Re from the other player.

The interaction involves the following three stages:

- In stage 1 players choose a mental state out of two possible mental states. These choices are potentially mixed.
- In stage 2 a signal, which reveals the mental state of player i to player j is generated. The signal involves a probability $1-p$ of error; i.e., if player i choses Re player j receives the signal Re with probability p and the signal Ra with probability $1-p$. Likewise, if he chooses Ra the signal is Ra only with probability p.
- In stage 3 the players choose an action C or D in the Prisoner's Dilemma game.

Equilibrium is now defined as follows via the following two conditions.

1. Actions taken in the mental game (after the choice of the mental states) form a Bayesian equilibrium.
2. Given the equilibrium expected in the mental game the choice of mental states forms a Nash equilibrium with respect to players' material preferences.

To avoid occurrence of multiple equilibria that arises from higher-order beliefs, we shall assume that an *Re* type who believes he is facing an *Re* type chooses C.

Under these conditions, the normal form game played by the players is the following one:

	Ra	Re
Ra	$1, 1$	$p + 5(1-p), p - 4(1-p)$
Re	$p - 4(1-p)$, $p + 5(1-p)$	$4p^2 + 5p(1-p) - 4p(1-p) + (1-p)^2$, $4p^2 + 5p(1-p) - 4p(1-p) + (1-p)^2$

Simplifying the expressions in the payoff matrix we get:.

	Ra	Re
Ra	$1, 1$	$5 - 4p, 5 - 4p$
Re	$5 - 4p, 5 - 4p$	$4p^2 - p + 1, 4p^2 - p + 1$

It is easy to check that this game has a symmetric and totally mixed equilibrium given by

$$(\frac{4p^2 + 3p - 4}{4p^2 - 2p + 1}, \frac{5(1-p)}{4p^2 - 2p + 1})$$

provided that $p \in (\frac{-3+\sqrt{73}}{8}, 1)$. This equilibrium induces a probability distribution on the strategy profiles of the Prisoner's Dilemm. This distriibution generically involve correlation between the players' actions. If, for example, $p = 4/5$, then the equilibrium is $(\frac{24}{49}, \frac{25}{49})$ and the corresponding distribution is

	D	C
D	0.167	0.0916
C	0.0916	0.65

which is not a product distribution.

12 Discussion

In his treatise *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 although we have also proposed the concept of mental subgame perfect equilibrium. However, it would be interesting to investigate the role of a new concept in which players can change their mental state during the course of the game. To capture the idea that emotions have a certain degree of persistence one can, for example, require 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 "globally"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 research directions pursued, possibly by a different set of authors.

13 Appendix

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

Example 6 *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 minmax vector of this game is $(1,0,0)$. *Hence,* (U,R,L) *does not pay player 1 at least his minmax 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 7 *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 minmax vector of this game is $(0, 0, 0)$ and all strategy profiles of the game pay each player at least his minmax 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] Kalay A., A. Kalay, and A. Kalay (2003). "Friends or Foes? Empirical Tes of a Simple One-Period Nash Equilibrium," mimeo.

[22] Mailath, G. and L. Samuelson "Repeated Games and Reputations: Long-Run Relationships" Oxford University Press, 2006.

[23] 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.

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

[25] 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.

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

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

[28] 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.

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

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

[31] 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.

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

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

[34] Winter, Eyal (2009). "Incentive Reversal," *American Economic Journal: Microeconomics*, forthcoming.

[35] 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.