Mental Equilibrium and Strategic Emotions*

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

Abstract

We model mental states 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 mental state must be a best response to the mental 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. In particular our approach allows us to identify the mental states (the psychology) that lead players to play various prominent experimental outcomes. We provide necessary and sufficient conditions for mental equilibria to be sustained by material preferences. Finally, we discuss the concept of collective emotions, which is based on the idea that players can coordinate their mental states.

Keywords: Games, Equilibrium, Behavioral Economics, Emotions

1 Introduction

Over the past three decades several interesting and important models have been developed to reconcile the discrepancy between experimental economic results and game-theoretic predictions, without neglecting the idea that

*The authors wish to thank Itai Arieli, Ken Binmore, Werner Gueth, Sergiu Hart, Eric Maskin, Assaf Romm, Reinhard Selten, and Jean Tirole for their comments and suggestions on an earlier draft of this paper. We also thank audiences at Bocconi, 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 Economía y Competitividad through projects MTM2011-27731-C03 and MTM2014-53395-C3-1-P.

players behave strategically. The common objective of these models is to re-evaluate the outcomes of the game for each player, while taking into account non-pecuniary 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. However, the measure of fairness depends on the players' actions and beliefs, which are determined in equilibrium.

In this paper we attempt to take an endogenous approach to non-monetary preferences by allowing these preferences to arise endogenously from equilibrium conditions. Immaterial preferences often tend to be contextual and even tailored to the underlying strategic environment. Rustichini and Villeval (2014) demonstrate that individuals' fairness judgments can be strongly affected by their bargaining positions. Meshulam, Winter, and Ben Shachar (2012) show how anger can be "synthesized" when agents have monetary incentives to become angry. Gneezy and Imas (2013) further show that sometimes subjects are aware of the effects of emotions, such as anger, on strategic interactions and try to gain a strategic advantage by manipulating these emotions. In this paper we attempt to investigate theoretically the role of such strategic emotions. Our main focus here is on the role of commitment in strategic emotions.

We shall use the term *mental state* to represent these non-monetary preferences and embed them in an equilibrium concept called *mental equilibrium*. Much of the focus of our analysis will be on deriving players' behavioral preferences through the equilibrium conditions. The linkage we have here between mental preferences and emotions requires some discussion. Mental preferences can arise from a variety of social and psychological generators. We use "emotions" in this paper as a general term for the mechanisms that generate mental preferences, including attitudes like fairness and reciprocity[1]. The term emotion is also meant to capture two additional im-

[1]Substantial evidence in neuroscience (starting with Damasio 1994) indicates that at-

portant aspects of mental preferences in our model: their role in securing commitment and the fact that they need to be transparent (at least to some degree). As we shall argue later, an emotional reaction, say, anger, allows players who experience it to credibly commit to retaliate against the cause of anger even when such retaliation is costly and therefore materially irrational. Furthermore, emotional reactions transmit social cues and signals to others and they are therefore, to a certain extent, transparent. Hence, they can be used as a commitment device.

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) that is typically different from the material utility function. A strategy profile s of the game is said to be a *mental equilibrium* if two conditions hold: firstly, s has to be a Nash equilibrium with respect to players' mental states. Secondly, each player's mental state is a best response to the mental states of the other players, given his material and selfish preferences. We offer two valid interpretations 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 anger, envy, and revenge, which play a role in many interactive situations, have been 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.[2] 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 be endowed with a

titudes such as fairness and reciprocity can be related to emotional reactions that are processed in the pre-frontal cortex.

[2]Our equilibrium conditions are necessary but not sufficient conditions for evolutionary stability.

certain emotional state that serves their interest. Emotional states are often induced through cognitive reasoning whether in full or in partial awareness and they can serve as a commitment device. In order for the commitment to be credible, the emotional state has to be genuine and not feigned.[3] To illustrate 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 canceled and that you should report to the airline desk the next day. Consider the following two scenarios: in scenario A you observe most of the passengers leaving the terminal quietly. In scenario B you run across an acquaintance who tells you that he was rerouted to a different flight after explaining to the airline employees, in a very assertive and determined manner, that he had to arrive at his destination that day. In scenario B 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. (2009), which shows that the objective emotional reactions of receivers in a dictator game strongly depend on the presence of incentives. Under the interpretation of rational emotions one can think of mental equilibrium as an equilibrium in an amended game of credible commitments. The material payoffs here are standard payoffs in a game and not a measure of evolutionary fitness. The two interpretations we propose are very distinct. The evolutionary interpretation fits 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 shall not argue in favor of one or the other as we believe that the appeal of each of these interpretations is context-dependent. In particular, in explaining the foundation of conventions and norms in games vaguely defined and robust to whether players can see each other or not, the evolutionary approach seems more appropri-

[3] A considerable body of recent work 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.

ate (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 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. Because we wish to fully endogenize non-monetary preferences we adopt a framework that puts no constraints on the set of potential preferences. An important part of our analysis is to identify "material" games, i.e., games in which all mental equilibria can be sustained by material preferences only. We show that all zero-sum games are material and we characterize the entire class of material games using the concept of "Stackelberg strategies" that appear in the literature on repeated games and reputation (e.g., Mailath and Samuelson 2006).

An equilibrium outcome of a game in our model involves a combination of a vector of mental preferences and a strategy profile. Most of our attention in this paper will be given to the endogenous preferences rather than to the strategy profiles. While the set of strategy profiles supported by a mental equilibrium is very large, the set of mental preferences supporting a given profile are much more structured and informative. Hence our approach is particularly useful as a contribution to understanding the underlying motives of players' strategies. Our approach allows us to take a strategy profile that is played frequently in laboratory or fields experiments and identify the mental preferences that support it. Put differently, it helps us reveal the psychology behind various prominent outcomes that emerge from the empirical data. This, we believe, is the most important advantage of our concept that other behavioral concepts lack as they treat preferences exogenously or assume specific functional forms. We demonstrate this point in Section 3 by

showing that for the class of all conflict games (which includes the Prisoner's Dilemma) only "reciprocal" mental preferences can support cooperation as a mental equilibrium. This result rules out the possibility that cooperation in the Prisoner's Dilemma is driven by altruism or inequality aversion.

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.[4] 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 player signals his mental state in this game through body language, facial expression, 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 imperfect, our ability to deceive is imperfect as well. In Section 8 we bring some empirical evidence to this effect and analyze 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 various dynamic models exist that converges to consistent beliefs about mental states. While interesting and important in themselves, these learning models are beyond the scope of this paper.

In addition to its relation to the literature on social preferences discussed above, our work is related to 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

[4]If mental preferences and actions are chosen simultaneously, so that players have strategic uncertainty about the mental states, the set of mental equilibria will boil down to be the set of Nash equilibria of the game. This is because mental preferences would lose their role as a commitment devise if the other players cannot monitor them. However, as we show in Section 8 it is enough to have informative but noisy signals about the mental preferences to facilitate the role of mental states as a commitment device.

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 being 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, was also present in Dekel et al. (2007), who develop a more general model than that in Gueth and Yaari (1992). Dekel et al. (2007) study the evolution of preferences using a notion of evolutionary stability that is much more stringent than Nash equilibrium. Both our model and Dekel et al.'s deal with endogenous preferences and both make the distinction between objective/material preferences and subjective/mental preferences. Moreover, both models adopt the approach by which the stability conditions imposed on subjective preferences intend to maximize objective outcomes. Thus, the two models are very similar. However, the two papers differ in many respects, both conceptually and technically. First, Dekel et al.'s results deal only with two-person symmetric games because their model relies on pairwise random matching within a single population. Since our main interest lies in the study of strategic preferences (and not in evolution), we impose simple Nash equilibrium conditions on preferences. This allow us to treat the entire class of normal form games for any finite number of players. However, the main difference between Dekel et al. (2007) and our work lies in the results and their implications. While their main interest is to characterize the equilibrium outcomes that survive their evolutionary process, our focus is on the endogenous preferences. Most our results are directed at identifying the mental preferences that support specific equilibrium outcomes, a direction on which Dekel et al. (2007) are almost silent. 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 modeling. Our objective is to study the role of mental states in strategic decision-making. Accordingly, much of our attention will be given to identifying the mental states that support specific strategic outcomes, mainly those which arise in

laboratory experiments. We shall show that our results are consistent with some prominent experimental results that cannot be explained by standard game-theoretic concepts. In terms of formal modeling our model differs from those used in the literature discussed above. It is formulated as a general equilibrium concept of normal form games with an arbitrary number of players. In contrast to other papers using the indirect evolutionary approach we do not impose 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, 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 mental states 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 equilibrium 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. 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 preferences are both necessary and sufficient to sustain cooperation in a mental equilibrium for these games. Section 4 deals with "material" games. In this section we provide a characterization of those two-person games in which all mental equilibria can be supported by material preferences. Section 5 is devoted 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 6 we deal with an alternative definition of mental equilibrium for n-person games. This amendment is motivated by showing that for games with four or more players the standard concept of mental equilibria loses its predictive power, since any strategy profile in such games is a mental

equilibrium. This follows from the fact that for some choices of mental states by the players the corresponding mental game may possess no pure Nash equilibria. We study properties of this amended concept and apply it to the game of voluntary contributions (the n-person Prisoner's Dilemma). We show that in a mental equilibrium (according to the new definition) 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 7 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 8 studies a variant of mental equilibrium that builds on the idea that the detection of others' mental states is imperfect. We discuss this concept in the context of the famous TV game "Split or Steal," which is a variant of the Prisoner's Dilemma. We refer to the empirical observation of "mind-reading" according to which pre-play communication results in correlated actions in the Prisoner's Dilemma, and show that our concept of mental equilibrium can explain this correlation. Section 9 concludes.

2 Basic Definitions

Let $G = (N, S, U)$ be a normal form game where N is the set of players, $S = S_1 \times S_2 \times \ldots \times S_n$ 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. Our focusing on pure strategies has to do with our desire to keep the set of mental preferences unrestricted. A mixture over the set of unrestricted mental preferences would mean mixing over an uncountable set of utility functions, which would complicate remarkably the corresponding analysis. Furthermore, in view of the existence results we provide in this paper, introducing mixed strategies seems unnecessary. Nevertheless, in Section 8 we provide an example where players' sets of mental states are finite and where mixed strategies are considered. We proceed now with the following basic observations.

Observation 1 *Any Nash equilibrium s of a game is also a mental equilibrium.*

Proof: *To see that this is the case, take 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 chooses a different mental state. Clearly, in the new mental game all other players will stick to their strictly dominant strategies. 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 choosing a different mental state.*

The converse is not true, as we shall see in many examples in this paper. Observation 2 provides a useful sufficient and necessary condition for a mental equilibrium to be a Nash equilibrium.

Observation 2 *A mental equilibrium s is a Nash equilibrium if and only if it can be supported by a profile of mental preferences in which each player's payoff is only a function of his own strategy.*

Proof: Consider first a mental equilibrium s that is a Nash equilibrium. For each player i, define a mental state u_i that depends only on player i's strategy

and is such that s_i is a strictly dominant strategy for i. Clearly, u defined in this way supports s. Suppose now that s is a mental equilibrium that is not a Nash equilibrium and let u be the profile of mental preferences supporting s. Assume by way of contradiction that u_i is only a function of i's strategy for each player i. Then for each i, s_i satisfies that $u_i(s_i, \bar{s}_{-i}) \geq u_i(\bar{s}_i, \bar{s}_{-i})$ for all $\bar{s}_i$ and for all $\bar{s}_{-i}$. Since s is not a Nash equilibrium then there is a player j and a strategy s_j^* that yields player j a higher material payoff in response to s. Consider now a mental state u_j' for player j under which s_j^* is a dominant strategy. Clearly, player j can guarantee himself a higher material payoff in equilibrium in the new mental game by moving to u_j'; this contradicts the fact that s is a mental equilibrium supported by u.

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 maxmin 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 maxmin value of player i, which we assume to exist (the existence is guaranteed for wide classes of strategic games, such as finite games).

Proposition 1 *Let G be a two-person game; then $s \in S$ is a mental equilibrium if and only if $U_i(s) \geq m_i$ for all $i \in \{1, 2\}$.*[5]

Proof. Let v_1 and v_2 be the maxmin values of players 1 and 2, respectively, with s_1 and s_2 being maxmin strategies.[6] 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 by 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

[5] In Section 6 and in the Appendix we show that Proposition 1 does not apply to n-person games with $n \geq 3$.

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

exists a pure Nash equilibrium yielding a payoff of at least v_1 to player 1. This contradicts the assumption that s^* is a mental equilibrium, and proves one direction. We next argue that every profile yielding at least the maxmin value for the two players is a mental equilibrium. For this we construct the following mental game: let $s = (s_1, s_2)$ be a profile that yields each of the two players at least his maxmin value. For the mental state of player 1 we set $u_1(s) = 1$, and $u_1(s'_1, s_2) = 0$ for all $s'_1 \neq s_1$. Furthermore, for every $s'_2 \neq s_2$ there exists s'_1 such that $U_2(s'_1, s'_2) \leq U_2(s)$; otherwise the maxmin value of player 2 is greater than $U_2(s)$, which contradicts the definition of s. We now set $u_1(s'_1, s'_2) = 1$ and $u_1(s^*_1, s'_2) = 0$ for all $s^*_1 \neq s'_1$. We now define the mental state of player 2 in a similar manner: $u_2(s) = 1$, and $u_2(s_1, s'_2) = 0$ for all $s'_2 \neq s_2$. Furthermore, for every $s'_1 \neq s_1$ there exists s'_2 with $U_1(s'_1, s'_2) \leq U_1(s)$; otherwise the maxmin value of player 1 must be greater than $U_1(s)$, which is impossible. We now have $u_2(s'_1, s'_2) = 1$ and $u_2(s'_1, s^*_2) = 0$ for all $s^*_2 \neq s'_2$. We can now show that s is a mental equilibrium of the game supported by u_1 and u_2. Indeed, s is clearly a Nash equilibrium under u_1 and u_2, as the mental game never has a payoff of more than 1 to either player. To show that condition (2) in the definition of mental equilibrium applies, note that if, say, player 1 changes his mental state to u'_1, then a Nash equilibrium of the new mental game (u'_1, u_2) must involve a strategy profile s' such that $u_2(s') = 1$. Otherwise the mental state of player 2 will deviate. But for such s' we must have $U_1(s') \leq U_1(s)$, which implies that player 1 cannot be better off 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 maxmin value. To show this, let s be a profile of maxmin strategies. Clearly, it pays each player 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 deviates only if all equilibria of the new mental game yielded 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 maxmin value, and that s^* is a weak mental equilibrium supported by the mental states $u = (u_1, u_2)$. Let s_1 be a maxmin strategy of player 1. Consider a mental state u_1' under which s_1 is a strictly 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 maxmin value (in the game $G = (N, S, U)$), but this contradicts the fact that s^* is a weak mental equilibrium since player 1 is better off deviating under the condition imposed by the definition of weak mental equilibrium. □

The two propositions above show that the set of mental equilibrium outcomes is typically quite large. This follows from the fact that unlike

models in the literature about the indirect evolutionary approach, which use refinements in the form of evolutionary (dynamic) conditions, we prefer to maintain all outcomes that pass the two levels of (Nash) equilibrium conditions. At the level of strategy profiles our equilibrium conditions can be thought of as necessary but not sufficient conditions for stability. Our weak equilibrium conditions are used not merely for the sake of simplicity, but primarily because our main concern is not with equilibrium outcomes but with the mental states that support these outcomes. Identifying the set of mental states supporting mental equilibria is more interesting and more challenging with milder equilibrium conditions. We shall start with the 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 3 *There are two mental equilibria in the Prisoner's Dilemma game,* (C,C) *and* (D,D).

Proof. It is easy to check that the maxmin 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 maxmin value to one of the players. □

It can easily be 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 a 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 defect, 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 4 we restrict ourselves to generic mental states. A mental state is generic if the corresponding player never displays indifference.

Observation 4 *Let $G = (N, S, U_1, U_2)$ be a Cooperation game. Then (C, C) is a mental equilibrium. Furthermore, a necessary and 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 the 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 equilibrium 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) that 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 deviate if the third inequality 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 4 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 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 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 inequality 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 $u_1 = U_1$. With such a mental state he will be able to sustain (D, C) as an equilibrium, which is the best possible outcome in terms of material 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 4. The former represents a mental state with some degree of altruism (if $\beta_i > 0$) or spitefulness (if $\beta_i < 0$). By 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 to player 1, and involves a greater equality than the outcome (C, C). Hence, in light of our discussion above, there exists no profile of (inequality-averse) mental preferences that 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.

Example 2 *Consider the finitely repeated Prisoner's Dilemma of the following one-shot game:*

	D	C
D	$1, 1$	$10, 0$
C	$0, 10$	$4, 4$

In this example we consider the one-shot Prisoner's Dilemma above as well as the finitely repeated game. In the case of repetition we shall be referring to the game in its normal form and to the Nash equilibria and the mental equilibria of this normal form game.

Observation 5 *It is well known that (D,D) is the only profile played in a Nash equilibrium in both the one-shot game and the finitely repeated game. Furthermore, as established earlier, only (C,C) and (D,D) are the mental equilibria in the one-shot game. However, mental equilibria of the finitely repeated game (in its normal form) admit plays in which both (C,D) and (D,C) are played. One such mental equilibrium is for players to alternate between these two outcomes. Such an equilibrium exists in any repeated Prisoner's Dilemma game where the total material payoff to the two players exceeds that of (D,D).*

Proof. The mental preferences supporting this equilibrium in our game above can be described as follows: (1) for the outcomes in which the two

players alternate between (C,D) and (D,C) the mental utility is identical to the material utility (and is $10k/2$, with k being the even number of periods); (2) if only (C,C) or (D,D) are played along the path the mental and material preferences are again identical; (3) for any other profile (i.e., in which the two players choose different actions in the same period) the mental utility function yields a payoff of -1 to both players. □

Interestingly, the mental preferences described above allow players in the repeated game to trade reciprocity within periods with reciprocity between periods. Substantial experimental evidence shows that players reciprocate by taking turns on their preferred outcomes in a variety of repeated interactions, including the repeated Prisoner's Dilemma. Kaplan and Ruffle (2011) study a class of two-person entry games and show that when payoffs are sufficiently symmetric players tend to cooperate by means of turn taking and alternate between (C,D) and (D,C). Sibly, Tisdel, and Evans (2014) have also documented substantial turn-taking behavior in laboratory experiments of the repeated Prisoner's Dilemma with and without cheap talks. Finally, Cason, Lau, and Mui (2013) also investigate the dynamics behind turn-taking behavior and show how it spreads through learning.

4 Material 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 mental preferences that affect players' behavior. Do all two-person games induce non-trivial mental preferences? Clearly games in which all mental equilibria can be supported by material preferences do not induce mental preferences that affect players' behavior. We refer to such games as *material games*. In a material 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 maxmin 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 Stackelberg 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 material game if and only if the following condition holds for every $s \in S$:*

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

Note that the condition specified in Proposition 3 applies to zero-sum games. Hence all zero-sum games are material. 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 maxmin value, this means that player 2 cannot guarantee his maxmin value, which is a contradiction. Let us see now the proof of Proposition 3.

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 $U(s) \geq m$, it follows from Proposition 1 that 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 of $M_i > U_i(s)$. □

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 3 *The Trust game. A large body of experimental data has helped us understand the Trust game since the publication of 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 transfer $0 \leq y \leq x$ to player 2. If player 1 transfers y, player 2 receives $3y$. Player 2 can now reward player 1 with a transfer of $z \leq 3y$. Finally, the payoff to player 1 is $x - y + z$ and the payoff to player 2 is $3y - z$.*

Observation 6 *An outcome (a_1, a_2) is a mental equilibrium outcome if and only if $a_1 \geq x$ and $a_2 \geq 0$.*

Proof. Consider such an outcome (a_1, a_2). Since $a_1 \geq x$ player 2 can guarantee that player 1 gets no more than a_1. This can be done by transferring no money back to player 1 if player 2 received any money from player 1. Furthermore, it is clear that player 1 can guarantee that player 2 receives no more than zero by simply making a zero transfer to player 2. In view of

Proposition 1, (a_1, a_2) is a mental equilibrium outcome. Consider a mental equilibrium outcome (a_1, a_2) such that either $a_1 < x$ or $a_2 < 0$. Then either player 1 or player 2 gets less than the maxmin value, which contradicts Proposition 1. □

We note that the Trust game has a unique Nash equilibrium in which player 1 makes a zero transfer to player 2. Observation 6 suggests that any level of trust displayed by player 1 coupled with a level of trustworthiness that compensates player 1 for 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 now discuss our concept of mental equilibrium in the context of another prominent game, the Ultimatum game.

Example 4 *The Ultimatum game. The game involves two players. Player 1 has an endowment of* 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. A massive amount of experimental evidence starting with Gueth et al. (1982) has shown, however, 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 introduce the subgame perfect version of mental equilibrium. Consider an extensive form game $G = (T, U)$, where T is the game form defined by a tree, and $U = (U_1, \ldots, U_n)$ are the payoff functions for players in $N = \{1, \ldots, n\}$ assigning a payoff vector to each 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 $U_i(s') > U_i(s)$.*

Observation 7 *Let G be an extensive form game. Every Nash equilibrium outcome of G is a mental subgame perfect equilibrium outcome of G.*

Proof. Let s be a Nash equilibrium of G. We construct the following mental (extensive form) game. For player $i \in N$ 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 includes a decision node where player i plays according to the strategy s_i; in any other case, $u_i(d) = 0$. It is clear that s constitutes a subgame perfect equilibrium in the mental game based on the profile of mental states u. 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 take $s' \in SPE(T, u_i', u_{-i})$ with $U_i(s') > U_i(s)$. From the definition of u it is clear that $s_j' = s_j$ in all the information sets in the path of s' (for all $j \neq i$). Hence, $U_i(s') = U_i(s_i', s_{-i})$. But this cannot happen since s is a Nash equilibrium of G. □

Subgame perfect equilibria are often described as Nash equilibria with credible threats. Mental equilibria are, in some sense, based on commitment and, thus, can turn non-credible threats into credible ones. This is the basic insight of Observation 7.

Returning to the Ultimatum game, it is easy to check that every distribution $(y, 1-y)$ can be supported by a Nash equilibrium. Hence, in view of Observation 7, it holds that all allocations of an Ultimatum game are supported by a mental subgame perfect equilibrium. Let us characterize the corresponding mental states.

Observation 8 *Consider the outcome in which player 1 gets x and player 2 gets $(1-x)$. Then the following conditions are both necessary and sufficient for a pair of mental states $u = (u_1, u_2)$ to sustain this outcome as a mental subgame perfect equilibrium.*

1. *If $u_1(y, 1-y) \geq u_1(0,0)$ then $y \geq x$, and if $u_2(y, 1-y) \geq u_2(0,0)$ then $x \geq y$.*

2. $u_1(x, 1-x) \geq u_1(0,0)$ *and* $u_2(x, 1-x) \geq u_2(0,0)$.

Proof. First we claim that under any pair of mental states satisfying the above conditions $(x, 1-x)$ is supported by a subgame perfect equilibrium of the mental game. Note that, by condition 1, any offer y with $y > x$ will satisfy that $u_2(0,0) > u_2(y, 1-y)$ and will be rejected by player 2. Adding condition 2 we get that $(x, 1-x)$ is the most preferred offer for player 1 among those which are acceptable for player 2. Otherwise, if for some $(z, 1-z)$ with $z < x$, $u_1(z, 1-z) \geq u_1(x, 1-x)$, then we have $u_1(0,0) > u_1(z, 1-z) \geq u_1(x, 1-x) \geq u_1(0,0)$, which is a contradiction. Since both players prefer the outcome $(x, 1-x)$ to the outcome $(0,0)$ this must be a unique subgame perfect equilibrium outcome under (u_1, u_2). Note now that no player can deviate to a different mental state and increase his material payoff in subgame perfect equilibrium since no such an outcome is acceptable for the other player on the basis of the mental states (u_1, u_2). Hence, the two conditions are sufficient. We now show that they are also necessary. Suppose that condition 1 fails to hold with respect to player 1; then there exists an outcome $(y, 1-y)$ with $y < x$ such that $u_1(y, 1-y) \geq u_1(0,0)$. Consider the mental state u_2' of player 2 such that $u_2'(y, 1-y) > u_2'(0,0)$ and $u_2'(z, 1-z) < u_2'(0,0)$ for $z \neq y$ (i.e., player 2 is willing to accept only $(y, 1-y)$). Clearly $(y, 1-y)$ is a subgame perfect equilibrium outcome of the mental game based on (u_1, u_2') and player 2's material payoff has increased, which is a contradiction. Suppose now that condition 1 is violated with respect to player 2, i.e. that for some $(y, 1-y)$ with $y > x$ it holds that $u_2(y, 1-y) \geq u_2(0,0)$. Consider now a deviation of player 1 to the mental state u_1' such that $u_1'(y, 1-y) \geq u_1'(z, 1-z)$ for any distribution $(z, 1-z)$. In that case $(y, 1-y)$ is a subgame perfect equilibrium outcome of the mental game based on (u_1', u_2). Since this outcome increases player 1's material payoff we again reach a contradiction. Finally, note that the necessity of condition 2 is immediate; if this condition is violated $(x, 1-x)$ cannot be a is a subgame perfect equilibrium outcome under the mental states (u_1, u_2). □

The insight provided by the observation above is that the mental preferences that sustain a given mental subgame perfect equilibrium outcome are characterized by the allocations that player doom as unacceptable; the way they compare two acceptable outcomes is irrelevant.

6 n-Person Games

Our model and results in this paper are based on pure strategies. [7] The restriction to pure strategies implies that when a player deviates to a different mental state the resulting mental game may not possess (pure) Nash equilibria. This would rule out such deviations and may sustain a large set of artificial mental equilibria. We start this section by showing that the set of mental equilibria expands to the entire set of strategy profiles when the number of players is at least four. In order to restore the predictive power of the mental equilibrium concept we screen out these artificial mental equilibria. We shall thus introduce a minor amendment to the definition that allows for more deviations and hence for fewer equilibria. This amended definition coincides with our original definition for two-person games. We use the amended definition for our subsequent analysis in this section. We start by demonstrating the drawback of the original definition.

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 the set of the remaining strategies by T_i such 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 the parity of k (i.e., whether k is odd or even) by $p(k)$. 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 $s \neq (0, ..., 0)$, half of the players can profit by deviating.[8] Indeed, each player who receives 0 can increase his payoff by changing his

[7] It turns out that mixed strategies pose a serious technical problem to our analysis, mainly for two reasons: (1) the space of mental states is a continuum and (2) f, the best-response correspondence at the level of the mental states, is highly non-convex, which makes existence hard to deal with. Olschewski and Swiatczak (2009), who build on our paper, address this issue for the case of 2×2 bimatrix games.

[8] 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.

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. In so doing 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 $(0, ..., 0)$ is a mental equilibrium, first note that it is a Nash equilibrium with respect to the chosen mental states $(u_1, ..., u_n)$ as it globally maximizes the payoff to all players. Furthermore, if player i deviates and sends a different mental state u_i' he will not be able to sustain a better equilibrium because the corresponding mental game will have no equilibrium 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. □

As pointed out earlier, the main reason why a mental equilibrium supports all strategy profiles in games of four or more players is that it does not permit deviations that induce mental games without pure Nash equilibria. This requirement seems rather demanding. If a player can deviate to a mental state that guarantees him a higher material payoff regardless of what other players are doing, then the lack of a pure Nash equilibrium is not essential. Indeed, while the deviating player cannot have consistent beliefs about what the other players will do, he can realize that no matter what the others do he will be better off. Building on this insight we propose a minor amendment to the original definition, which will resolve the excessive multiplicity described in Proposition 4.

Definition 4 *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 and a mental state u_i' such that (a) for a strategy profile $s' \in NE(N, S, u_i', u_{-i})$ it holds that $U_i(s') > U_i(s)$, or (b) i has a dominant strategy s_i' in the game (N, S, u_i', u_{-i}) with $U_i(s_i', s_{-i}') > U_i(s)$ for all s_{-i}'.*

In the sequel we shall use the amended definition presented here. In the Appendix we show that (1) a mental equilibrium (as defined here) always

exists, (2) any mental equilibrium pays each player at least his maxmin payoff, and (3) the converse of 2 is not true, i.e., there are profiles yielding at least maxmin payoffs to all players that are not mental equilibria. This also implies that the new definition is a refinement of the original concept and that the set of mental equilibria is now generically smaller than the set of all profiles regardless of the number of players. In the Appendix we also show that the amended definition coincides with the original definition for two-person games.

To illustrate the advantage of the amended definition for games with more than two players we discuss the famous *Public Good game*, which is also the n-person version of the Prisoner's Dilemma.

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 public endowment (choose 1) or not (choose 0). The total public endowment contributed is multiplied by a factor of $1 < k < n$ and divided equally among all players. Thus, supposing that r players contribute, the payoff to a player who chooses 1 is $\frac{krw}{n} - w$ and the payoff to a player who chooses 0 is $\frac{krw}{n}$. Note that the unique Nash equilibrium (with dominant strategies) 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 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 less than $\frac{n}{k}$ cannot be a mental equilibrium. Suppose by way of contradiction that such an equilibrium exists. Consider a player i whose mental state contributes. Player i's payoff in such an equilibrium is $\frac{krw}{n} - w < 0$. Suppose that this player selects a different mental state in which choosing 0 is a dominant strategy. Then, in the new mental game player i has a dominant strategy that guarantees him a higher payoff: if no player contributes then player i receives a zero payoff, and in all other

cases he receives a strictly positive payoff. This contradicts the equilibrium conditions. We now show that a profile with a number of contributors $r \geq \frac{n}{k}$ 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 number of other agents who choose 1 is $|N-T|-1$ (otherwise he prefers to choose 0). For each player in T we assign a mental state in which choosing 0 is a strictly dominant strategy. Given this set of mental states it is clear that the underlying strategy profile is an equilibrium of the mental game. It therefore remains to show that condition (2) in the definition of mental equilibrium applies. Clearly, no player in T is better off deviating. Selecting a different mental state will not trigger any 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 first that the resulting mental game has a pure Nash equilibrium. Suppose by way of contradiction that this equilibrium yields a higher material payoff to player i. Clearly in the new equilibrium players in T will still all choose 0. Hence, player i can only be made better off if in his new mental game his equilibrium strategy is 0. But in such an equilibrium all the remaining players in $N-T$ must choose 0 as well, which would make player i strictly worse off. Assume now that the new mental game has no pure Nash equilibrium; then we have to show that player i has no strategy that guarantees him a material payoff of more than $\frac{krw}{n} - w > 0$. This is clearly the case since whenever all other players choose 0 player i gets zero or less. 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 under which 0 is a strictly dominant strategy. This would make $(0, ..., 0)$ a Nash equilibrium in the mental game and no player is better off by selecting a different mental state, as he would get zero or less. □

The attractive property of mental equilibria when applied to the Public Good game (with the amended definition) 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 becomes 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.

7 Collective Mental States

Some emotions tend to intensify when experienced within a group. People tend to laugh more when watching a comedy as a group than they would when viewing it alone. Violent mob behavior is often a result of a collective rage that is experienced at a level that exceeds individual rage. In many strategic environments the benefits of emotional reactions, and in particular their usage as a commitment device, are enhanced when they are generated collectively by a group (often vis-à-vis outside players). We refer to this framework as collective mental states. 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. Let us 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.[9] 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:*

[9]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 $i \in M$.

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

2. *There exist no coalition* $T \subset N$ *and mental preferences for the members of* T *denoted by* $u'_T = \{u'_j\}_{j \in T}$ *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 not be able to 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 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 a 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 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 have $U_j(s') > U_j(s)$ for all $j \in T$. Under $u_{N \setminus T}$ each player has a strictly 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 the fact that $U_j(s') > U_j(s)$ for all $j \in T$ contradicts the fact that s is a strong equilibrium in G. □

At this point, one may think of amending the definition of strong mental equilibrium in a similar way as we amended the definition of mental equilibrium. Although we might do it, it is not really necessary since the set of strong mental equilibrium as defined in Definition 5 does not expand to the entire set of strategy profiles when the number of players is at least four. We prove it in the next observation.

Observation 11 *If s is a strong mental equilibrium of a game $G = (N, S, U)$, then it is a Pareto undominated strategy profile.*[10]

Proof. Let s be a strong mental equilibrium of G supported by the mental profile u. Suppose by way of contradiction that there exists $s' \in S$ with $U_i(s') > U_i(s)$ for all $i \in N$. It is clear that there exists a mental profile u' such that $s' \in NE(N, S, u')$; it contradicts the fact that s is a strong mental equilibrium of G. □

Moreover, in two-person games we can prove the following characterization of the strong mental equilibria.

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

Proof. 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 presented 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 that yields both of them a higher material payoff. This implies that s is a strong mental equilibrium. For the converse, if s^* is a strong mental equilibrium, then as argued before it is also a mental equilibrium. Furthermore, in view of Observation 11, s^* is Pareto undominated. □

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

[10] 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 $i \in N$.

only the outcome (D, D) while the set of mental equilibria contains both (D, D) and (C, C). Interestingly,

Observation 13 *(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 12 it is the unique strong mental equilibrium. □

8 Mental Equilibrium and Mind-Reading

"Split or Steal" is a popular TV game in the UK ("Friend or Foe" is the US version), whose final stage involves a simplified Prisoner's Dilemma game. The two competing individuals who acquired jointly a substantial sum of money (sometimes more than 100,000 pounds) in a preliminary stage (by solving trivial 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 jointly acquired 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" and "Friend or Foe" games 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. To model the strategic environment of the TV game we shall consider a variant of

mental equilibrium. Bachi, Ghosh, and Neeman (2013) introduce a related model that is based on a direct commitment on actions and not on mental states. 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 that 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 chooses 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 Dilemma. This distribution generically involves 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

and note that similarly to the empirical results in Kalay et al. (2003) this distribution over strategy profiles displays a significant correlation in players' actions.

9 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 in which 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 mental states 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 mental states have a certain degree of persistence one can, for example, require that players commit to a mental state for a duration of k periods or, alternatively, that players 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 from 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 over the course of a 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.

10 Appendix

In order to prevent ambiguities, in this appendix we call *r-mental equilibrium* to the version of mental equilibrium in Definition 4.

Proposition 5 *Let $G = (N, S, U)$ be an n-person game and take $s \in S$ to be an r-mental equilibrium of G. Then $U_i(s) \geq m_i$ for all $i \in N$.*

Proof. Take $s \in S$ with $U_i(s) < m_i$ for some $i \in N$ and assume that there exists a profile u of mental states supporting s as an r-mental equilibrium. Take now $s_i' \in S_i$ to be a maxmin strategy of i and u_i' to be a mental state according to which playing s_i' is a dominant strategy for i. Then, since

$$U_i(s_i', s_{-i}') \geq m_i > U_i(s)$$

for all s'_{-i}, u does not support s as an r-mental equilibrium, which is a contradiction. □

Proposition 6 *Let $G = (N, S, U)$ be an n-person game and take $s \in S$ to be a mental equilibrium of G such that $U_i(s) \geq m_i$ for all $i \in N$. Then s is an r-mental equilibrium of G.*

Proof. Take a profile u of mental states supporting s as a mental equilibrium. If u does not support s as an r-mental equilibrium then there must exist $i \in N$ and $s'_i \in S_i$ with

$$U_i(s'_i, s'_{-i}) > U_i(s)$$

for all s'_{-i}, but this clearly implies that $U_i(s) < m_i$, which is impossible. □

Corollary 2 *Let $G = (N, S, U)$ be an n-person game with $n = 2$. Then, $s \in S$ is an r-mental equilibrium if and only if $U_i(s) \geq m_i$ for all $i \in N$.*

Proof. Proposition 5 shows the *only if* part. To prove the *if* statement, take into account that in two-person games every profile whose associated payoff vector is greater than or equal to the maxmin vector is a mental equilibrium; using this and Proposition 6, the *if* statement follows. □

Corollary 2 and Proposition 1 imply that in two-person games mental equilibrium and r-mental equilibrium are equivalent concepts.

Corollary 3 *Let $G = (N, S, U)$ be an n-person game with $n \geq 4$. Then, $s \in S$ is an r-mental equilibrium if and only if $U_i(s) \geq m_i$ for all $i \in N$.*

Proof. Proposition 5 shows the *only if* part. To prove the *if* statement take into account that in n-person games with $n \geq 4$ every profile is a mental equilibrium; using this and Proposition 6, the *if* statement follows. □

The case $n = 3$ is a singular case. The following example shows a three-person games with a profile that is not an r-mental equilibrium in spite of the fact that its associated payoff vector is greater than or equal to the maxmin vector.

Example 6 *Consider the following three-person game.*

a	a	b
a	*0,0,0*	*1,1,1*
b	*1,1,1*	*1,1,1*

b	a	b
a	*1,1,1*	*1,0,1*
b	*0,1,1*	*1,1,0*

Clearly $m = (0, 0, 0)$ *and thus* $U(a, a, a) \geq m$. *However,* (a, a, a) *is not a mental equilibrium (and hence not r-mental). Suppose that it is. Then, there must exist a mental game satisfying that (i) for the strategy profiles* (a, b, a), (b, a, a), (b, b, a), (a, a, b) *there exist at least two players willing to deviate, and (ii) in* (b, a, b) *either player* 1 *wants to deviate or players* 2 *and* 3 *want to deviate, and in* (a, b, b) *either player* 2 *wants to deviate or players* 1 *and* 3 *want to deviate, and in* (b, b, b) *either player* 3 *wants to deviate or players* 1 *and* 2 *want to deviate. It is an easy exercise to check that conditions (i) and (ii) cannot be simultaneously satisfied.*

In spite of this negative result, the following proposition shows that in three-person games there always exists at least an r-mental equilibrium.

Proposition 7 *Every three-person game* $G = (N, S, U)$ *has an r-mental equilibrium.*

Proof. The proof is constructive. First we define a strategy profile s^* and then we prove that it is an r-mental equilibrium. For every $s_3 \in S_3$ define $S_2(s_3)$ as the set of maxmin strategies of player 2 in $G(s_3)$, the two-person game resulting from G when we fix s_3. Denote by $m_2^{s_3}$ the maxmin payoff to player 2 in $G(s_3)$. Now, choose a strategy of player 2 in $S_2(s_3)$ and denote it by $s_2(s_3)$. Next, choose a best reply of player 1 to $(s_2(s_3), s_3)$ in G and denote it by $s_1(s_3)$. Finally, choose $s_3^* \in S_3$ satisfying that

$$U_3(s_1(s_3^*), s_2(s_3^*), s_3^*) \geq U_3(s_1(s_3), s_2(s_3), s_3), \text{ for all } s_3 \in S_3.$$

Let us check that $s^* := (s_1(s_3^*), s_2(s_3^*), s_3^*)$ is a mental equilibrium of G. Consider the mental game u defined below for every $s \in S$:

$$u_1(s) = \begin{cases} 1 & \text{if } s = (s_1(s_3), s_2(s_3), s_3), \\ 1 & \text{if } s = (\hat{s}_1, \hat{s}_2, s_3) \text{ with } \hat{s}_2 \neq s_2(s_3), \text{ and } U_2(\hat{s}_1, \hat{s}_2, s_3) \leq m_2^{s_3}, \\ 0 & \text{otherwise.} \end{cases}$$

$$u_2(s) = \begin{cases} 1 & \text{if } s_2 = s_2(s_3), \\ 0 & \text{otherwise.} \end{cases}$$

$$u_3(s) = \begin{cases} 1 & \text{if } s_3 = s_3^*, \\ 0 & \text{otherwise.} \end{cases}$$

Obviously, $s^* \in NE(N, S, u)$.

- Player 1 cannot gain by choosing a different u_1' because players 2 and 3 are going to play $(s_2(s_3^*), s_3^*)$ in (N, S, u_1', u_2, u_3) and then player 1 will also play $s_1(s_3^*)$.
- Player 2 cannot gain by choosing a different u_2' because in (N, S, u_1, u_2', u_3) player 3 will play s_3^* and player 1 will play $\hat{s}_1$ with $U_2(\hat{s}_1, \hat{s}_2, s_3^*) \leq m_2^{s_3^*}$ for any $\hat{s}_2 \neq s_2(s_3^*)$. Notice that $m_2^{s_3^*} \leq U_2(s^*)$.
- Player 3 cannot gain by choosing a different u_3' because in (N, S, u_1, u_2, u_3'), for any election s_3 of player 3, players 1 and 2 will play $(s_1(s_3), s_2(s_3))$ and, by definition of s_3^*, player 3 will not gain more than $U_3(s^*)$.

Hence, s^* is a mental equilibrium of G. Moreover, it is easy to check that $U(s^*) \geq m$, where m denotes the maxmin vector in G. In fact, it is clear that:

- $m_1 = \max_{s_1} \min_{s_2 s_3} U_1(s_1, s_2, s_3) \leq \max_{s_1} U_1(s_1, s_2(s_3^*), s_3^*) = U_1(s^*)$,
- $m_2 = \max_{s_2} \min_{s_1 s_3} U_2(s_1, s_2, s_3) \leq \max_{s_2} \min_{s_1} U_2(s_1, s_2, s_3^*) \leq U_2(s^*)$,
- $m_3 = \max_{s_3} \min_{s_1 s_2} U_3(s_1, s_2, s_3) \leq \max_{s_3} U_3(s_1(s_3), s_2(s_3), s_3) = U_3(s^*)$.

Then, in view of Proposition 6, s^* is r-mental. □

References

[1] Aumann, R.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, R.J. (2008). "Rule Rationality vs. Act Rationality," Discussion Paper 497, Dec. 2008. The Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem.

[3] Bachi, B., Ghosh S. and Neeman Z. (2013) "Communication and Deception in Two-person Games," mimeo.

[4] Berg, J., Dickhaut, J., and McCabe, K. (1995). "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, 10, 122–142.

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

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

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

[8] Cason, T., Lau S. and Mui V. (2013) "Learning, Teaching, and Turn Taking in the Repeated Assignment Game," *Economic Theory*, 54, 335-357.

[9] Dekel, E., Ely, J.C., and Yilankaya, O. (2007). "Evolution of Preferences," *Review of Economic Studies*, 74(3), 685–704.

[10] 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.

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

[12] Fershtman, C., Judd, K.L., and Kalai, E. (1991). "Observable Contracts: Strategic Delegation and Cooperation," *International Economic Review*, 32(3), 551–559.

[13] Fershtman, C., and Kalai, E., (1997). "Unobserved Delegation," *International Economic Review*, 38(4), 763–774.

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

[15] Fehr, E., and Schmidt, K. (1999). "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 1, 817–868.

[16] Fehr, E., and Falk, A. (2002). "Psychological Foundations of Incentives," *European Economic Review*, 46, 687–724.

[17] Fischbacher, U., Gaechter, S., and Fehr, E. (2001). "Are People Conditionally Cooperative? Evidence from a Public Good Experiment," *Economic Letters*, 71, 397–404.

[18] Gneezy, U. and Imas A. (2013) "Materazzi Effect and the Strategic Use of Anger in Competitive Interactions" PANS July, 2013.

[19] Gueth, W., Schmittbereger, R., and Schwarze, B. (1982). "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, 3, 367–388.

[20] Gueth, W., and Yaari, M. (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.

[21] Gueth, W., and Kliemt, H. (1998). "The Indirect Evolutionary Approach: Bridging between Rationality and Adaptation," *Rationality and Society*, 10, 377–399.

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

[23] Ichino, A., and Maggi, G. (2000). "Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm," *The Quarterly Journal of Economics*, 115(3), 1057–1090.

[24] Issac, R.M., Walker, J.M., and Williams, A.W. (1994). "Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Very Large Groups," *Journal of Public Economics*, 54, 1–36.

[25] Israel, S., Hart E., and Winter E. (2013). "Oxytocin Decreases Accuracy in the Perception of Social Deception," *Psychological Science*, 23, 1–3.

[26] Kalay, A., Kalay, A., and Kalay, A. (2003). "Friends or Foes? Empirical Tes of a Simple One-Period Nash Equilibrium," mimeo.

[27] Kaplan, T., and Ruffle, B. (2011). "Which Way to Cooperate", *Economic Journal* 122, 1042–1068.

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

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

[30] Meshulam, M., Winter, E., Shahar, G.B., and Aharon, Y. (2012). "Rational Emotions in the Lab" *Social Neuroscience*, 7(1), 11–17.

[31] 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.

[32] Olschewski, G. and Swiatczak, L. (2009). "Existence of Mental Equilibria in 2x2 Games," Handelshochschule Leipzig, mimeo.

[33] Ochsner, K.N., and Gross, J.J. (2005). "The Cognitive Control of Emotion." *Trends in Cognitive Science* 9(5), 242–249.

[34] Phan, K.L., Wager, T., Taylor, S.F., and Liberzon I. (2002). "Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMRI" *NeuroImage,* 16, 331–348.

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

[36] Rapoport, A., Guyer, M.J., and Gordon, D.G. (1976). *The 2 X 2 Game*, Ann Arbor, MI: The University of Michigan Press.

[37] Rustichini, A., and Villeval, M.C. (2014) "Moral Hypocrisy, Power and Social Preferences," *Journal of Economic Behavior & Organization*, 107, 10–24.

[38] Sibly, H., Tisdell, J., and Evans, S. (2014) "Turn-Taking in Finitely Repeated Symmetric Games: Experimental Evidence," unpublished manuscript.

[39] 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.

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

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

[42] Winter, E. (2009). "Incentive Reversal," *American Economic Journal: Microeconomics*, 1(2), 133–147.