When all is said and done, how should you play and what should you expect ?

R.J. Aumann and J. Dreze

Discussion Paper 2005-21

Département des Sciences Économiques
de l'Université catholique de Louvain

UCL

When All is Said and Done, How Should You Play and What Should You Expect?

R. J. Aumann*and J. H. Dreze†

Abstract

Modern game theory was born in 1928, when John von Neumann published his Minimax Theorem. This theorem ascribes to all two-person zero-sum games a value—what rational players may expect—and optimal strategies—how they should play to achieve that expectation. Seventy-seven years later, strategic game theory has not gotten beyond that initial point, insofar as the basic questions of value and optimal strategies are concerned. Equilibrium theories do not tell players how to play and what to expect; even when there is a unique Nash equilibrium, it it is not at all clear that the players "should" play this equilibrium, nor that they should expect its payoff. Here, we return to square one: abandon all ideas of equilibrium and simply ask, how should rational players play, and what should they expect. We provide answers to both questions, for all n-person games in strategic form.

*Center for Rationality and Interactive Decision Theory, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

†Center for Operations Research and Econometrics, Université Catholique de Louvain, 34 Voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium

1. Introduction

Modern game theory was born in 1928, when John von Neumann published his Minimax Theorem. This theorem ascribes to all two-person zero-sum games a *value*—what rational players may expect—and *optimal strategies*—how they should play to achieve that expectation.

Seventy-seven years later, strategic game theory has not gotten beyond that initial point, insofar as the basic questions of value and optimal strategies are concerned. To be sure, we do have equilibrium theories: the initial concept of Nash [1951], its various refinements[1] and coarsenings[2], and the selection theory of Harsanyi and Selten [1987]. But when the game is not two-person zero-sum, none of these theories actually tell the players what to expect and how to play[3]. Even when there is just one Nash equilibrium, it is not at all clear that the players "should" play this equilibrium, nor that they should expect its payoff[4].

Here, we return to square one: abandon all ideas of equilibrium and simply ask, how should rational players play, and what should they expect. We provide answers to both questions, for *all* n-person games in strategic[5] form. The answers do turn out related to the idea of equilibrium; specifically, correlated equilibrium. But the relationship is not straightforward.

A good starting point is Kadane and Larkey [1982], who wrote that each player in an n-person game should use ordinary one-person Bayesian decision theory to choose her strategy. Specifically, she should ascribe a (subjective) probability to each $(n-1)$-tuple of (pure) strategies of the other players. For this, they wrote,

[1]Selten [1975], Kreps and Wilson [1982], Kalai and Samet [1984], Kohlberg and Mertens [1986], and many others. For comprehensive surveys, see Hillas and Kohlberg [2002] and van Damme [1987, 2002].

[2]Correlated equilibrium and subjective equilibrium [Aumann 1974] and rationalizability [Bernheim 1984, Pearce 1984].

[3]The Harsanyi-Selten selection theory does choose a unique equilibrium, composed of a well-defined strategy for each player and having a well-defined expected outcome. But nobody—least of all Harsanyi and Selten themselves—would actually recommend using these strategies. This is for many reasons, including the complexity of the theory, its sophistication, and the more or less arbitrary choices that the theory makes at various points.

[4]As in the repeated prisoner's dilemma, the centipede game [Rosenthal 1982], unprofitable games [Harsanyi 1966, 1977, Morgan and Sefton 2002] (Example 7.1), Shapley's [1964] game (Example 5.3), and many others.

[5]A game in *strategic* (or "normal") form is one described by strategies and payoffs only, without reference to the sequence of moves and the information of the players when they make those moves.

classical game theory is of little or no use; rather, one should apply disciplines such as psychology. Once the subjective probabilities are determined, the player should choose her own strategy to maximize her subjective expected payoff.

On its face, the Kadane-Larkey viewpoint seems straightforward and reasonable. But it ignores a fundamental insight of game theory: that a rational player should take into account that *all the players are rational, and reason about each other*. Let's call this "interactive rationality." To be sure, the implications of interactive rationality are not, on their face, clear; but it does seem clear that it *has* logical—not only psychological—implications, which substantively restrict the possible outcomes. Identifying those restrictions is the object of the current work.

Thus, we answer our title questions—"how should you play and what should you expect?"—exactly as Kadane and Larkey do: Play to maximize your expected payoff given your subjective probabilities about the other players' choices, and expect the resulting expectation. But, unlike Kadane and Larkey, we note that the demands of interactive rationality severely restrict the expectations, and go on to characterize precisely what expectations *can* arise under this restriction.

To start with, one must formulate precisely the notion of interactive rationality—itself a nontrivial task. We do so by means of the notions of *common knowledge of rationality* (CKR) and *common priors* (CP). We then characterize the strategies and expectations that are possible under CKR and CP (Theorem A). In the special case of two-person zero-sum games, it turns out (Theorem B) that these expectations coincide with the minmax value of the game; thus, our results really do extend von Neumann's.

The plan of the paper is as follows: After summarizing our results informally in Section 2, and presenting them formally in Section 3, we devote Section 4 to a careful conceptual discussion of Theorem A. An alternative formulation of Theorem A is provided in Section 7. Section 6 is devoted to "determined" games—those with a unique rational expectation—including in particular two-person zero games, which are carefully discussed in 6.1. Numerical examples are adduced in Sections 5 and 6.2 - 6.6; proofs, in Section 8. Section 9 discusses the historical background and the literature. Section 10 is devoted to general discussion, and Section 11 to listing some open questions and directions for research.

2. Results

In this section we state our results informally, for simplicity confining attention to two-person games. The formal treatment in the ensuing sections covers also the general n-person case.

Let G be a two-person game in strategic (normal) form, represented by a bimatrix with r rows and k columns. It is convenient to analyze the game from the viewpoint of one player, called the *protagonist,* whom w.l.o.g. we take to be the row player. Specifically, we ask, what are her[6] *rational expectations* in G—the amounts that she could expect when common knowledge of rationality and common priors are assumed[7]?

Recall that a *correlated equilibrium* of G [Aumann 1974, 1987] is a probability distribution ρ on pure strategy pairs, such that if a pure strategy pair is chosen in accordance with ρ, and each player is informed only of his component of the chosen pair, then it is optimal for him to play that component, assuming that the other is playing his component. The protagonist's conditional expected payoff, given her information (the chosen row), is called a *conditional payoff* to ρ; there may be as many as r different such payoffs, depending on which row was chosen. A *conditional correlated equilibrium payoff*[8] in G is a conditional payoff to some correlated equilibrium ρ of G. Correlated equilibria are described by a limited number[9] of explicit linear inequalities; so they, and the corresponding conditional payoffs, are explicitly describable in terms of G.

The *doubled* game $2G$ is a game with $2r$ rows and k columns, each row of G appearing twice. Doubling a game affects its correlated equilibria in a non-trivial way; they are *not* necessarily just doubled versions of the correlated equilibria of the original game.

[6]The protagonist is female. The other player is—or players are, in the n-person case—of indeterminate gender; to distinguish them from the protagonist, we use masculine pronouns for them. Similarly, we use masculine pronouns for players in general, who may or may not include the protagonist.

[7]See Section 3 for precise definitions.

[8]We emphasize again that unless otherwise specified, all payoffs, expectations, and so on are to the protagonist. Thus a "rational expectation" is a rational expectation of hers; the "expectation range" is the range of her rational expectations; a "conditional payoff" is a conditional payoff to her; a "conditional correlated equilibrium payoff" is a conditional correlated equilibrium payoff to her; and so on.

[9]There are $r^2 - r + k^2 - k$ inequalities in rk variables, plus the $rk + 1$ inequalities that say that the probabilities are non-negative and sum to 1.

THEOREM A. Rational expectations in the game G coincide with conditional correlated equilibrium payoffs in the doubled game $2G$.

More precisely, α is a rational expectation in G if and only if in the doubled game $2G$, there are a correlated equilibrium ρ, and a row such that α is the conditional payoff to ρ given that row. Note that the correlated equilibria in $2G$ are determined by fewer than $(2r+k)^2$ inequalities in $2rk$ variables. For examples, see Section 5 below.

Next, we turn to two-person zero-sum games.

THEOREM B. Every two-person zero-sum game has a unique rational expectation, namely the value of the game.

Thus the notion of rational expectation indeed provides a generalization, to arbitrary games, of the classical minimax value in two-person zero-sum games. Further discussion of this point is provided in Section 6.1, where we argue that characterizing the minimax value via rational expectations—common knowledge of rationality and common priors—is more compelling than any of the arguments hitherto advanced, including the original arguments of von Neumann and Morgenstern [1944].

Theorems A and B are our main results. The following propositions contain some "practical" remarks, helpful in calculating rational expectations.

PROPOSITION C.
(i) Every conditional correlated equilibrium payoff—in particular, every Nash equilibrium payoff—is a rational expectation.
(ii) The rational expectations are unchanged by iterated deletion of strongly dominated strategies.
(iii) Every rational expectation is at least the protagonist's minimax payoff[10].
(iv) The rational expectations are covariant under multiplication of a player's payoffs by a positive constant, under addition of a constant to the row player's payoffs in a column, and under addition of a constant to the column player's payoffs in a row.

Item (i) says that the notion of rational expectation is weaker than that of Nash or even correlated equilibrium. Nevertheless, it is strong enough to yield the value in two-person zero-sum games (Theorem B). Proposition C follows from

[10] In mixed strategies.

Theorem A; for item (iv), we note that these transformations do not change the correlated equilibria.

To state the next proposition, recall that Myerson [1997] called a game *elementary* if it has a correlated equilibrium that assigns positive probability to each strategy of each player, and all the inequalities associated with this equilibrium are strict[11]. We then have:

PROPOSITION D. Every elementary game has a maximum rational expectation, namely the (protagonist's) highest payoff at any strategy pair.

Finally, we have:

PROPOSITION E. If all correlated equilibrium payoffs are the protagonist's minimax payoff v, then v is the only rational expectation.

3. Formal Treatment

This section describes our framework both more generally and more formally than Section 2. More generally, in that it refers to n players rather than just two; and more formally, in that it defines precisely what is meant by outcomes that can occur "under" CKR and CP.

As stated in the previous section, the viewpoint taken here is that of a single player, the *protagonist.* She is the "you" of our title; it is her whom we advise how to play and what to expect. As noted above, she must take into account that there are other players, each analyzing the situation from his viewpoint. But the bottom line is *her* decision and *her* expectation. We designate her Player 1.

Formally, a (strategic) n-person *game* G consists of n abstract sets $S_1, S_2, ..., S_n$ (the *strategy sets* of the players) and n functions $h_1, h_2, ..., h_n$ from $S = S_1 \times S_2 \times ... \times S_n$ to $\mathbb{R}$ (the *payoff functions).* To define CKR and CP, we need the idea of a player's *belief hierarchy,* which specifies his belief about what the others play, about what they believe he—and the others—play, about what they believe about *that,* and so on ad infinitum. Formally, such a hierarchy is most easily represented by means of a *belief system*[12] B for G, consisting of:

[11] I.e., if a player is informed that the chosen strategy pair calls for him to play a certain strategy, then it is strictly *better* for him to choose that strategy than any other strategy. Myerson showed that in a certain sense, all games may be "reduced" to elementary games.

[12] Originated by Harsanyi (1967-8) and subsequently developed, in various versions and under various names, by many workers.

(i) for each player i, a finite[13] set T_i whose members t_i are called *types*[14] of i; and
(ii) for each type t_i of each player i,

(a) a strategy[15] of i in G, denoted $s_i(t_i)$, and

(b) a probability distribution on $(n-1)$-tuples of types of the other players, called t_i's *theory.*

It may be seen that a player's type uniquely determines the whole hierarchy of his beliefs.

A *common prior* is a probability distribution π on $T_1 \times ... \times T_n$ that assigns positive probability to each type of each player, and such that the theory of each type of each player is the conditional of π given that that player is of that type[16]. Less formally, such that each player's probability for an event is its probability under the common prior, conditioned on his information—i.e., on his being the type he is[17]. A type of a player is *rational* if the strategy it prescribes maximizes his expected payoff given its theory. Rationality is *commonly known* if this is so for all types of all players[18]; note that both the existence of a common prior (CP) and common knowledge of rationality (CKR) are properties of the belief system as a whole. A *rational expectation* in G is an expected payoff of some type of the protagonist in some belief system for G in which CKR and CP obtain. We wish to characterize the set of rational expectations.

A *correlated equilibrium* of G is a probability distribution ρ on the set S of strategy profiles, with the following property: if a strategy profile s is chosen in accordance with ρ, and each player i is informed only of his component s_i of s, then it is optimal for him to play that component, assuming that the others are playing their components. The protagonist's expected payoff if she plays s_1 is called *the conditional payoff to ρ given s_1.*

The *doubled* game $2G$ is the n-person game in which 1's strategy set is $S_1 \times \{1,2\}$, the strategy sets of all other players remain as in G, and for all players i, the payoff $h_i(s_1, ..., s_n)$ is assigned to both the strategy profiles $((s_1, 1), s_2, ..., s_n)$ and $((s_1, 2), s_2, ..., s_n)$. In words, there are two copies of each of the protagonist's strategies, and the payoff does not depend on *which* copy is used.

[13]Finiteness is assumed here for convenience. For a general treatment, see, say, Aumann and Heifetz (2002).

[14]Or *information sets.*

[15]Member of S_i.

[16]In symbols, $\pi_i(t^{-i}; t_i) = \pi(t)/\pi(t_i)$ for each i and each t in $T_1 \times ... \times T_n$, where $\pi_i(\cdot; t_i)$ is t_i's theory and t^{-i} is the $(n-1)$-tuple of types assigned by t to players other than i.

[17]Examples are provided in the next section.

[18]This definition is equivalent to the more familiar definition, in terms of iterated knowledge.

With these definitions, the results stated in Section 2 remain correct as they stand[19].

4. Conceptual Discussion

4.1. Game Situations

The work described here began with the following question: Game Theory tells us what to expect in a two-person zero-sum game—the maxmin value. But how about games G that are *not* two-person zero-sum? What should we expect there, given just the formal description of G?

As stated, the question has no answer; the problem is underspecified. Formally, a game is defined by its strategy sets and payoff functions. But in real life, many other parameters are relevant; there is a lot more going on. Situations that substantively are vastly different may nevertheless correspond to precisely the same strategic game. For example, in a parliamentary democracy with three parties, the winning coalitions are the same whether the parties hold a third of the seats in parliament each, or, say, 49%, 39%, and 12% respectively. But the political situations are quite different. The difference lies in the attitudes of the players, in their expectations about each other, in custom, and in history, though the rules of the game do not distinguish between the two situations. Another example revolves around the ultimatum game [Güth, Schmittberger, and Schwarze 1982], which when played in different cultures, leads to systematically different outcomes [Roth, Prasnikar, Okuno-Fujiwara, and Zamir 1991].

Thus if one is given only the abstract formulation of a game, one cannot reasonably hope for an expectation and optimal strategies. Somehow, the real-life context in which the game is played must be taken into account.

The essential element in the notion of "context" is the mutual expectations of the players about the actions and expectations of the other players. As we saw in the previous section, such mutual expectations may be represented by a "belief system." So we define a *game situation* Γ to consist of a strategic game G as defined by its strategy sets and payoff functions, together with a belief system B for that game, and a particular type t_1 of the protagonist in that belief system.

[19] *Mutatis mutandi.* Specifically, "strategy pair" must be replaced by "strategy profile;" in Proposition C(iii), the protagonist's minimax strategy must take into account that the other players may correlate among each other; and in Proposition C(iv), the results are covariant when to a player's payoff function one adds a function that does not depend on that player's choice.

Of course, with this definition the question in the title becomes trivial, as the expectations are implicit in t_1. What we do is turn the question around: Given only G, what can we say about expectations in game situations Γ that are "based[20] on" G? That is the question addressed here.

Unlike a game, a game situation is a setting with "differential" (or "incomplete") information. Not about objective, exogenous factors like payoffs or utilities, but simply about the actions (or intentions) of the players. Each player has a probability distribution over—i.e., beliefs about—the actions of the others, but *knows* neither their actions nor their beliefs. The *game* is commonly known, but the *game situation* is not.

This state of affairs is inherent in the idea of "game situation." The 39% party may believe that the 49% party prefers a coalition with the 12% party; but it isn't sure, and is even less sure of the other parties' beliefs about its beliefs.

4.2. Ex Ante and Interim Viewpoints

Economists distinguish three stages in differential information environments. *Ex ante,* no one has any information; in the *interim,* each agent has his private information only; *ex post,* all information is revealed to all. In our context, ex ante the protagonist knows only the belief system B—which is commonly known by all players; in the interim, each player knows his type, but not those of the others; ex post, each player knows the types of all players.

Some readers may be curious about the relationship of the current work to Aumann [1987], which appears to make the same assumptions—Common Knowledge of Rationality and Common Priors—but reaches distinctly different conclusions[21]. Namely, the 87 paper arrives at *unconditional* expectations of correlated equilibria (CE's) of the *given* game G; here, we arrive at *conditional* expectations of CE's of the *doubled* game $2G$.

The puzzle is solved by noting that the current work concerns the interim stage, while the 87 paper concerns the ex ante stage. The expectations here are those of a protagonist who knows her type, as is indeed natural in a game situation. In contrast, in the 87 paper, the expectations—and the distribution of strategy profiles—are taken over the entire belief system B. Specifically, the protagonist

[20] I.e., that consist of G together with some belief system B for G and some type t_1 of the protagonist in B.

[21] Indeed, many who have been present at preliminary presentations of this material have expressed puzzlement on this matter.

does not know her own type, and in particular does not know the strategy she will play once she is informed of her type. The correlated equilibrium that emerges in the 87 paper is the protagonist's[22] prior distribution—when she knows only B—of the strategy profiles that *will* be played once the players are informed of their types. It can also be viewed as the distribution of strategy profiles of an outside observer who does not know the players' types.

5. Examples

In the two-person games below, the row and column players are Rowena and Colin respectively. Rowena is the protagonist.

5.1. Rational Expectations may be Mutually Inconsistent

	L	R
T	$6,6$	$2,7$
B	$7,2$	$0,0$

Figure 1a
The game G

	L	R
T	$1/2$	$1/2$
B	$7/8$	$1/8$

Figure 1b
Rowena's beliefs

	L	R
T	$1/2$	$7/8$
B	$1/2$	$1/8$

Figure 1c
Colin's beliefs

	L	R
T	$7/22$	$7/22$
B	$7/22$	$1/22$

Figure 1d
The common prior

The game G in Figure 1a ("Chicken") has three Nash equilibria: two pure, yielding $(2,7)$ and $(7,2)$, and one mixed, yielding $(4\frac{2}{3}, 4\frac{2}{3})$. Consider now a belief system with four states, TL, TR, BL, and BR, with each player's probabilities *for* each state *in* each state as depicted in Figures 1b and 1c. For example, in BL as well as in BR, Rowena's probabilities for BL and BR are $7/8$ and $1/8$ respectively, while for TL and TR they are 0. Rowena has the two types T and B, Colin the two types L and R.

The expectation of Rowena's type B is $6\frac{1}{8}$. She attributes probability $1/8$ to Colin's type being R, in which case his expectation, too, will be $6\frac{1}{8}$. So in that case, the players will each expect $6\frac{1}{8}$. These expectations are mutually inconsistent; $(6\frac{1}{8}, 6\frac{1}{8})$ is infeasible—it is outside the convex hull of the possible payoff vectors. And this in spite of common knowledge of rationality, which the reader may verify, and the existence of a common prior, depicted in Figure 1d.

[22] Also that of any other player, which by the common prior assumption, is the same as the protagonist's.

BR is the conflict outcome in Chicken. We see here that conflict may occur even when the players reason perfectly rationally and attribute rationality to each other; both players know about the inconsistency, and indeed it is commonly known that it may occur. Contrary to common wisdom (or rather, foolishness), the conflict is not due to any irrationality, but simply to differing assessments, which may well ensue when players are provided with different information.

To be sure, the mixed Nash equilibrium may also lead to conflict. But in that case, the players' assessments of the situation are not inconsistent. It is the inconsistency of the assessments that is noteworthy here.

The distribution in Figure 1d is a correlated equilibrium of G, and $6\frac{1}{8}$ is its conditional payoff given B.

	L	R
T	1/2	1/2
B	1	0

Figure 1e
Rowena's beliefs

	L	R
T	1/2	1
B	1/2	0

Figure 1f
Colin's beliefs

	L	R
T	1/3	1/3
B	1/3	0

Figure 1g
The common prior

Another belief system for G is depicted in Figures 1e - 1g. Here it is common knowledge that the conflict outcome BR is impossible. In particular, a type B Rowena expects 7 and knows that Colin is of type L, so expects 4. The payoff pair $(7, 4)$ is, however, infeasible. Thus here again, the expectations of the players are mutually inconsistent, in spite of there being no element of irrationality in the system.

5.2. Different Conditional Correlated Equilibrium Payoffs in 2G and G

	L	C	R
T	$0,0$	$4,5$	$5,4$
M	$5,4$	$0,0$	$4,5$
B	$4,5$	$5,4$	$0,0$

Figure 2a
The game G

The game G of Figure 2a [Shapley 1964] has a single Nash equilibrium, namely, $((1/3, 1/3, 1/3), (1/3, 1/3, 1/3))$, yielding the payoff $(3, 3)$. Consider now a belief system with seven states, $T_1R, T_2C, T_2R, ML, MR, BL$ and BC, with each player's

probabilities set forth in Figures 2b and 2c (as in Example 5.1); Rowena's strategy in rows T_1 and T_2 is T. If Rowena's type is T_1, the game situation has expectation 5. She knows that Colin's type is R, so that his expectation is $4\frac{1}{2}$. Thus Rowena *knows* that the players' expectations are the infeasible[23] pair $(5, 4\frac{1}{2})$—in spite of common knowledge of rationality and a common prior (depicted in Figure 2d). Here, unlike in the previous example, 5 is not a conditional payoff to a correlated equilibrium of G, given any strategy of Rowena[24].

	L	C	R
T_1	0	0	1
T_2	0	2/3	1/3
M	1/2	0	1/2
B	1/2	1/2	0

Figure 2b
Rowena's beliefs

	L	C	R
T_1	0	0	1/4
T_2	0	1/2	1/4
M	1/2	0	1/2
B	1/2	1/2	0

Figure 2c
Colin's beliefs

	L	C	R
T_1	0	0	1/12
T_2	0	1/6	1/12
M	1/6	0	1/6
B	1/6	1/6	0

Figure 2d
The common prior

But it *is* a conditional payoff to a correlated equilibrium of the doubled game $2G$, depicted in Figure 2e; the correlated equilibrium in question is depicted in Figure 2f. Note that if we eliminate the rows in which all the probabilities vanish, Figure 2f becomes Figure 2d. A similar relationship obtained in our first example, but with G instead of $2G$.

0, 0	4, 5	5, 4
0, 0	4, 5	5, 4
5, 4	0, 0	4, 5
5, 4	0, 0	4, 5
4, 5	5, 4	0, 0
4, 5	5, 4	0, 0

Figure 2e
The doubled game $2G$

0	0	1/12
0	1/6	1/12
1/6	0	1/6
0	0	0
1/6	1/6	0
0	0	0

Figure 2f
A correlated equilibrium in $2G$

Alternatively, a correlated equilibrium of this game may be obtained by assigning probability 1/6 to the entries with payoffs $(4, 5)$ or $(5, 4)$. The associated

[23]The payoffs sum to $9\frac{1}{2}$, whereas the maximum sum in the matrix is 9.

[24]By G's symmetry, we may suppose that Rowena's component of the strategy pair is T. The correlated equilibrium cannot then assign positive probability to TC, as Rowena's conditional payoff would then be < 5. So everything is eliminated by a sequence of strict dominations: C by L, then B by M, then L by R, then M by T, and finally R by C, leaving nothing. In fact, the highest conditional payoff to a correlated equilibrium of G, given a strategy of Rowena, is $5 - (1/3126) \approx 4.99968$.

inequalities are all strict, so the game is elementary. So by Proposition D, the maximum rational expectation is 5.

5.3. The Set of Rational Expectations Need not be Convex

	L	R
T	$1,1$	$0,0$
B	$0,0$	$0,0$

Figure 3
The game G

In all the examples adduced up to now, the set of rational expectations is an interval. In the game G of Figure 3, this is not so; here, there are precisely two rational expectations: 1 and 0.

6. Determined Games

Call a game *determined* if it has only one rational expectation. That is, the context doesn't matter after all; the expectation is independent of the context. Theorem B says that all two-person zero-sum games are determined. So in a sense, we are back to our starting point: in two-person zero-sum games, game theory provides unequivocal answers. But in fact, we have come further. As we shall see, there are games that are neither two-person zero-sum, nor easily reducible to such games, but are nevertheless determined.

6.1. Two-Person Zero-Sum Games

Von Neumann and Morgenstern [1944] advance two kinds of argument to support the minmax value of two-person zero-sum games—*equilibrium* arguments and *guaranteed value* arguments. The equilibrium argument says that if game theory is going to recommend something, then that recommendation must be a Nash equilibrium, and all Nash equilibria of two-person zero-sum games yield the value. The guaranteed value argument says that the row player can guarantee getting the value v, and the column player can guarantee not paying more than v, "so" rational players must reach precisely the value.

The equilibrium argument is of a formal, mathematical nature; one *proves* that there can be only one equilibrium payoff. But the guaranteed value argument is

more tenuous. We purposely put the word "so" in quotation marks, because there is a bit of a non-sequitor there. Fully to justify this kind of argument, one needs a formal framework. In fact, though this "argument" appears to depend only on the rationality of the players, it is not true that players who are merely rational must necessarily reach the value; one needs *common knowledge* of rationality[25], and one also needs common priors[26].

One may think of Theorem B as reflecting the guaranteed value argument, but in a rather subtle way. The players do *not* actually guarantee the value. In many two-person zero-sum games[27], it is in fact impossible to do so in pure strategies; and here, we think of the players as using pure strategies only. Rather, the protagonist *expects* the value. Guarantees enter the argument in showing that what she expects cannot be less than the value, because she *could*—by using mixed strategies—attain at least the value in expectation. One needs further arguments, revolving around the common knowledge, the common prior, and the zero-sumness to show that she also cannot expect more than the value.

Indeed, the current perspective shows exactly where the "classical" argument breaks down. It is true that Players 1 and 2 can guarantee v and $-v$ respectively; so, since the sum of payoffs is 0, the only feasible "individually rational[28]" payoff pair is $(v, -v)$. It is also true that any rational expectation (of either player) must be individually rational; that is Proposition C(iii). What is *not* in general true is that the players' expectations must constitute a feasible pair, i.e., be "consistent." Indeed, we saw in Section 5 that inconsistent expectations are the rule rather than the exception; in particular, we saw that it is possible for Rowena to *know* that Colin expects a payoff that is inconsistent with the payoff she *knows* she is getting, even though rationality is commonly known and there is a common prior. On the face of it, there is no reason to suppose that a similar situation could not arise also in two-person zero-sum games.

But in fact, it cannot. Theorem B says that there is something special about two-person zero-sum games that makes it impossible. So this theorem goes considerably beyond the classical "guaranteed value" argument for the value in such games.

As for the "equilibrium argument," this is certainly entirely precise; but it is less compelling, because Nash equilibrium is a much stronger assumption than

[25]See Example 6.6.

[26]See Example 6.5.

[27]Specifically, unless the game has a pure strategy saddle point.

[28]This means that each player gets at least what he can guarantee to himself.

rational expectation. Indeed, we have seen (Proposition C(i)) that every Nash equilibrium payoff is a rational expectation, but the converse is certainly not true. So saying that every rational expectation is the value is saying a good deal more than that every Nash equilibrium payoff is the value.

Moreover, when subjected to close substantive examination, also the equilibrium argument falls apart. It assumes that the putative "recommendation of game theory" must be for each player to play some specified (mixed or pure) strategy, *known to all players.* Game theory need not make that kind of recommendation; its recommendation could be—indeed, *should* be—"respond optimally to your private information." As pointed out in Section 4.1 above, the players are faced with a game *situation,* not just a game. Even though the *game* is commonly known, the *game situation* is not. It is, indeed, replete with private information, which there is no reason for the players to ignore.

6.2. Some Determined Two-Person Non-Zero-Sum Games

By Proposition C(ii), the prisoner's dilemma is determined. The game G of Figure 4a also is. Indeed, the unique correlated equilibrium ρ of G is depicted in Figure 4b. The conditional expected payoff to ρ given either T or B is $1/2$, and the maxmin is also $1/2$. So by Proposition C(iii), every rational expectation is $\geq 1/2$. If in $2G$ there were a correlated equilibrium and a row with a conditional expectation that is $> 1/2$, then there would also have to be a row with conditional expectation $< 1/2$, contradicting our conclusion that every rational expectation is $\geq 1/2$.

	L	R
T	$1,0$	$0,1$
B	$0,2$	$1,0$

Figure 4a
The game G

	L	R
T	$1/3$	$1/3$
B	$1/6$	$1/6$

Figure 4b
The correlated equilibrium ρ

Note that ρ is equivalent to the Nash Equilibrium $(\frac{2}{3}T + \frac{1}{3}B, \frac{1}{2}L + \frac{1}{2}R)$, which yields $1/2$, but does not *guarantee* it. To guarantee $1/2$, Rowena would have to play $\frac{1}{2}T + \frac{1}{2}B$; but this is not part of any Nash equilibrium. Such games—in which no Nash equilibrium yields either player more than his maxmin payoff—are called *unprofitable* [Harsanyi 1966, 1977, Morgan and Sefton 2002].

Our point of view is in any case somewhat different. Rather than guaranteeing something, the players maximize, given their beliefs; and they use pure strategies,

which cannot guarantee anything here.

That G is determined also follows from Theorem B and Proposition C, since G can be transformed into a zero-sum game by "C(iv)" transformations.

The repeated prisoner's dilemma and Rosenthal's [1982] centipede are also determined, but this does not appear to follow easily from the general results in Section 2.

6.3. Some Determined Three-Person Games

6.3.1. Suppose G has three players, and each pair plays "matching pennies." So each player plays in two matches, and so has four strategies; his payoff in G is the sum of his payoffs in his two matches. The inequalities defining correlated equilibria imply that all correlated equilibria of $2G$ yield conditional expected payoffs of 0 for each strategy. Intuitively, each player can guarantee 0 to himself in each of his two matches, so also in the overall game; but as we saw in 5.1, this in itself is not enough—one must go through the actual calculations.

6.3.2. Another game of this kind is where each player displays stone, scissors, or paper, the payoff for any pair is determined as usual, and as before, each player's payoff is the sum of his payoffs in his two matches. Here each player has just three strategies.

6.4. A Non-determined Game with a Unique Correlated Equilibrium

	L	R
W	$1,1,-1$	$-1,-1,1$

	L	R
E	$-1,-1,1$	$1,1,-1$

Figure 5a
The game G

	L	R
W	$1/4$	$1/4$

	L	R
E	$1/4$	$1/4$

Figure 5b
The correlated equilibrium ρ

The game G depicted in Figure 5a has three players. Rowena chooses the only row there is; Colin chooses a column, L or R; and Matt chooses a matrix, W or

E. As between Colin and Matt, this is matching pennies. So there is a unique correlated equilibrium ρ, depicted in Figure 5b; it yields Rowena 0. But the game $2G$, depicted in Figure 5c, has many correlated equilibria. For example, the correlated equilibrium ρ', depicted in Figure 5d; the conditional expected payoff given T is 1, whereas given B it is -1. Thus the set of rational expectations in G is $[-1, 1]$.

	L	R
T	$1, 1, -1$	$-1, -1, 1$
B	$1, 1, -1$	$-1, -1, 1$

W

	L	R
T	$-1, -1, 1$	$1, 1, -1$
B	$-1, -1, 1$	$1, 1, -1$

E

Figure 5c
The game $2G$

	L	R
T	1/4	0
B	0	1/4

W

	L	R
T	0	1/4
B	1/4	0

E

Figure 5d
The correlated equilibrium ρ'

6.5. Failure of Theorem B without Common Priors

	L	R
T	$1, -1$	$-1, 1$
B	$-1, 1$	$1, -1$

Figure 6a
The game G

	L	R
T	.9	.1
B	.1	.9

Figure 6b
Rowena's beliefs

	L	R
T	.1	.9
B	.9	.1

Figure 6c
Colin's beliefs

The game G is "matching pennies." With the depicted belief system—which has no common prior—it is common knowledge that it is optimal for each player to play that strategy with which his type is designated. In particular, Rowena's type T plays T and expects .8, whereas the value of the game is 0.

Careful consideration of the example leads to some discomfort. It is commonly known that Rowena believes that[29] Colin believes that Rowena does the opposite

[29]Short for "ascribes probability .9 to".

of what she really does—and vice versa[30]; i.e., that each player ascribes grave errors to the other. This is typical of situations without common priors. Common knowledge of rationality does obtain in this example.

6.6. Failure of Theorem B without CKR

	L_1	L_2	R_1	R_2	L_3
T_1	.1	.1	0	0	0
B_1	.1	0	.1	0	0
B_2	0	.1	0	.1	0
T_2	0	0	.1	0	.1
T_3	0	0	0	.1	.1

Figure 7
The common prior

Figure 7 depicts a common prior for a belief system in the game "matching pennies" (see Figure 6a). Type L_3 of Colin is irrational, but all other types of both players are rational. It follows that type T_1 of Rowena is rational, knows that Colin is rational, knows that he knows that she is rational, knows that he knows that she knows that he is rational, and knows that he knows that she knows that he knows that she is rational; moreover, T_1's expectation is 1, whereas the value of the game is 0. The example can be extended to an arbitrarily high level of iterated mutual knowledge of rationality; but by Theorem B, not to common knowledge.

7. An Alternative Formulation of Theorem A

Theorem A is stated in terms of the doubled game $2G$. It can also be stated in terms of a set of "augmented" games, in each of which a single strategy of the protagonist is "doubled." Formally, given a strategy r_1 of the protagonist in G, define the *augmented* game G_{2r_1} as the n-person game in which

1's strategy set is $(S_1 \backslash \{r_1\}) \cup (\{r_1\} \times \{1, 2\})$;

the strategy sets of all other players remain as in G;

for all players i, the payoff $h_i(r_1, s_2, ..., s_n)$ is assigned to both the strategy profiles $((r_1, 1), s_2, ..., s_n)$ and $((r_1, 2), s_2, ..., s_n)$; and

[30]That is, with Rowena and Colin interchanged.

the payoff is as in G for all other strategy profiles.

In words, r_1 is replaced by two copies, and the payoff does not depend on which copy is used.

THEOREM A′. Rational expectations in the game G coincide with conditional correlated equilibrium payoffs in the augmented games G_{2s_1}, where s_1 ranges over the protagonist's strategies.

More precisely, α is a rational expectation in G if and only if there is a strategy s_1 of the protagonist such that in the augmented game G_{2s_1}, there are a correlated equilibrium ρ and a strategy such that α is the conditional payoff to ρ given that strategy.

8. Proofs

PROOF OF THEOREM A′. Let B be a belief system for G with CKR and a common prior π. Let t_i^1 and t_i^2 be two types of player i who play the same strategy. Let B' be the belief system obtained from B by *amalgamating* t_i^1 and t_i^2 into a single type u_i^0. Specifically, in B', the type space of each player j other than i is T_j; the type space U_i of i is obtained from T_i by removing t_i^1 and t_i^2 and replacing them by u_i^0, where

(1) $s_i(u_i^0) := s_i(t_i^1) = s_i(t_i^2)$;

and the common prior π' in B' is defined by

(2) $\pi'(u_i, t^{-i}) := \pi(u_i, t^{-i})$ if $u_i \neq u_i^0$, and

(3) $\pi'(u_i^0, t^{-i}) := \pi(t_i^1, t^{-i}) + \pi(t_i^2, t^{-i})$.

LEMMA 4. CKR obtains in B'.

PROOF. We must show that in B', the strategy of each type maximizes that type's expectation. For types of players j other than i, this is immediate, since their (conditional) expectations are the same in B' as in B, whether or not they play the strategies prescribed by their types. The same holds for types u_i of i other than u_i^0. For i's type u_i^0, one must show that i's conditional expectation

$\sum_{t^{-i} \in T^{-i}} \pi'(u_i^0, t^{-i}) h_i(s_i(u_i^0), t^{-i}) / \sum_{t^{-i} \in T^{-i}} \pi'(u_i^0, t^{-i})$

if he plays the strategy $s_i(u_i^0)$ prescribed by u_i^0 is at least as great as his conditional expectation

$\sum_{t^{-i}\in T^{-i}} \pi'(u_i^0, t^{-i}) h_i(r_i, t^{-i}) / \sum_{t^{-i}\in T^{-i}} \pi'(u_i^0, t^{-i})$

if he plays some other strategy r_i; i.e., since the denominators are the same in the two expressions, that

(5) $\sum_{t^{-i}\in T^{-i}} \pi'(u_i^0, t^{-i}) h_i(s_i(u_i^0), t^{-i}) \geqq \sum_{t^{-i}\in T^{-i}} \pi'(u_i^0, t^{-i}) h_i(r_i, t^{-i})$.

But by the same token, the optimality (in B) of $s_i(t_i^1)$ for t_i^1 and of $s_i(t_i^2)$ for t_i^2 yield

(6) $\sum_{t^{-i}\in T^{-i}} \pi(t_i^1, t^{-i}) h_i(s_i(t_i^1), t^{-i}) \geqq \sum_{t^{-i}\in T^{-i}} \pi(t_i^1, t^{-i}) h_i(r_i, t^{-i})$, and

(7) $\sum_{t^{-i}\in T^{-i}} \pi(t_i^2, t^{-i}) h_i(s_i(t_i^2), t^{-i}) \geqq \sum_{t^{-i}\in T^{-i}} \pi(t_i^2, t^{-i}) h_i(r_i, t^{-i})$;

and, by (1) and (3), adding (6) and (7) yields (5). This establishes Lemma 4.

Corollary. Amalgamation does not affect the expectation of any type of any player, except for the types that have been amalgamated.

Let α be a rational expectation in G; we must prove that it is a conditional correlated equilibrium payoff in one of the augmented games. By definition of "rational expectation," there is for G a belief system B with CKR, a common prior π, and a type u_1 of the protagonist whose expectation is α. Let $r_1 := s_1(u_1)$ be the strategy played by type u_1. W.l.o.g., there is another type—different from u_1—who plays r_1. For if not, we may split u_1 into two identical types, with the same strategy and theory as u_1. The new common prior (after the split) is then obtained from the original one by halving the probabilities of all states affected by the split.

By repeatedly amalgamating types and using Lemma 4 and its corollary, we may arrive at a belief system B'' with CKR and a common prior π'', such that (i) for each strategy of each player—other than r_1—at most one type plays that strategy; (ii) the protagonist has a type u_1^1 that is "like" u_1, in that it plays the same strategy r_1, and has the same expectation α; and (iii) the protagonist has exactly one other type, u_1^2, that plays[31] r_1.

The game G_{2r_1} is exactly like G, except that the strategy r_1 is "doubled:" call the duplicates r_1^1 and[32] r_1^2. Each type in B'' corresponds to a strategy in G_{2r_1}; specifically, u_1^1 and u_1^2 correspond to r_1^1 and r_1^2. Hence π'' induces a probability distribution ρ on the strategy profiles in G_{2r_1}, it being understood that strategy profiles without a counterpart in B'' are assigned probability 0.

Claim 8. ρ is a correlated equilibrium in G_{2r_1}.

[31] Obtained by amalgamating all the types of the protagonist who play r_1, other than u_1^1.

[32] Rather than the more cumbersome $(r_1, 1)$ and $(r_1, 2)$.

Proof. CKR in B'' tells us that the expectation of any type is at least as great if it plays the strategy prescribed for that type, as if it plays some other strategy. Rephrasing, any strategy in G_{2r_1} with positive ρ-probability yields to its player a conditional expected payoff under ρ at least as great as any other strategy.

From Claim 8 and the Corollary to Lemma 4, it follows that α is a conditional correlated equilibrium payoff in G_{2r_1}. This completes the proof of Theorem A$'$ in one direction.

For the other direction, let β be a conditional correlated equilibrium payoff in an augmented game G_{2r_1}; specifically, the conditional payoff to the correlated equilibrium ρ, given some strategy s_1 in the augmented game. Define a belief system B for G as follows: the types of i in B are in one-one correspondence with those of his strategies in G_{2r_1} to which ρ assigns positive probability; the theory of a type is the conditional of ρ given the strategy corresponding to that type. Then ρ is a common prior for B, and that CKR holds in B is the same as saying that ρ is a correlated equilibrium is CKR. This completes the proof of Theorem A$'$.

PROOF OF THEOREM A. If α is a rational expectation in G, then by Theorem A$'$, it is a conditional payoff to a correlated equilibrium ρ' in some augmented game G_{2r_1}, given a strategy x' of the protagonist in that game The strategy profiles in $G_{2r_1}$are in one-one correspondence with those strategy profiles in $2G$ whose first component is $(r_1, 1), (r_1, 2)$, or $(s_1, 1)$ for an s_1 other than r_1. Assigning to any such profile the ρ'-probability of the corresponding profile in G_{2r_1}, and 0 to all other profiles (those of the form $((s_1, 2), s_2, ..., s_n)$ for $s_1 \neq r_1$), yields a correlated equilibrium ρ in $2G$. The strategy x' in G_{2r_1} corresponds to some strategy x in $2G$, and then α is the conditional payoff to ρ in $2G$ given x. This completes the proof of Theorem A in one direction.

In the other direction, let β be a conditional payoff to a correlated equilibrium ρ in $2G$, given a strategy x of the protagonist; w.l.o.g., x has the form $(r_1, 1)$. "Amalgamating" $(s_1, 1)$ and $(s_1, 2)$ for all s_1 other than r_1 yields a correlated equilibrium ρ' in G_{2r_1}, and turns $(r_1, 1)$ into a strategy in G_{2r_1}. Then β is a conditional payoff to ρ' in G_{2r_1}, given $(r_1, 1)$. So by Theorem A$'$, β is a rational expectation in G. This completes the proof of Theorem A.

PROOF OF PROPOSITION C.

(i) The first part follows from Theorem A, as every correlated equilibrium in G can also be viewed as a correlated equilibrium in $2G$, since one can simply assign

probability 0 to one of the two duplicates of each of the protagonist's strategies. The part about Nash equilibrium follows from the fact that at a Nash equilibrium ν, all "active" strategies of the protagonist (indeed, of any player)—i.e., those with positive probability at ν—must get the same payoff, so that the conditional expected payoffs coincide with the total expected payoff.

(ii) Follows from the fact that a strictly dominated strategy can never appear with positive probability in a correlated equilibrium, as it is always worthwhile to switch to the dominating strategy.

(iii) Let α be a rational expectation, t_1^* a type of the protagonist with expectation α in a belief system with CKR and a common prior, $s_1^* := s_1^*(t_1^*)$ the strategy played by type t_1^*, and p the probability distribution over $S_{-1} := S_2 \times ... \times S_n$ that t_1^*'s theory induces. Thus $\alpha = \sum_{s_{-1} \in S_{-1}} p_{s_{-1}} h_1(s_1^*, s_{-1})$. By CKR, s_1^* maximizes t_1^*'s expectation given its theory, so

$$\alpha = \max_{s_1 \in S_1} \sum_{s_{-1} \in S_{-1}} p_{s_{-1}} h_1(s_1, s_{-1}) \geq \min_q \max_{s_1 \in S_1} \sum_{s_{-1} \in S_{-1}} q_{s_{-1}} h_1(s_1, s_{-1}),$$

where q ranges over all probability distributions over S_{-1}.

(iv) Follows from Theorem A, since correlated equilibria are covariant in the required manner.

PROOF OF PROPOSITION E. Proposition C(iii) says that every rational expectation is $\geq v$. Suppose α is a rational expectation that is $> v$. By Theorem A, α is a conditional payoff of a correlated equilibrium ρ in the doubled game $2G$. By Proposition C(iii) and Theorem A, all other conditional payoffs to ρ in $2G$ are $\geq v$. Since the unconditional payoff to ρ is the expectation of the conditional payoffs, and α appears in this expectation with positive probability, it follows that the unconditional payoff to ρ is $> v$. But an unconditional payoff to the correlated equilibrium ρ in $2G$ is also an unconditional payoff to a correlated equilibrium in G, obtained by amalgamating duplicated strategies. So we get a correlated equilibrium payoff in G that is $> v$, contrary to hypothesis.

PROOF OF THEOREM B. Follows from Proposition E, since in two-person zero-sum games, the (unconditional) expected payoff to every correlated equilibrium is the value[33].

PROOF OF PROPOSITION D. By definition, the given game G has a correlated equilibrium μ that assigns positive probability to each strategy of each player, and in which the associated inequalities are strict. Let S be the set of strategy

[33] Aumann [1974], last paragraph of Section 2.

profiles in G. If ι assigns equal probabilities to all strategy profiles, and $\varepsilon > 0$ is sufficiently small, then $\lambda := (1-\varepsilon)\mu + \varepsilon\iota$ assigns positive probability to each strategy profile, and the associated inequalities are still strict. Let w be a strategy profile in G that yields the protagonist her highest payoff in G. For each strategy profile s in G, let s^1 and s^2 be the two copies[34] of s in $2G$. Let $0 < \delta < \lambda_w$. Define a probability distribution ρ on the set $2S$ of strategy profiles in $2G$ by $\rho_{s^1} := \lambda_s,\ \rho_{s^2} := 0$ for $s \neq w$, and $\rho_{w^1} := \lambda_w - \delta,\ \rho_{w^2} := \delta$. We will show that ρ is a correlated equilibrium of $2G$ when δ is sufficiently small.

Indeed, the inequalities associated with ρ in $2G$ are the same as those associated with λ in G, except for those that correspond to w_1^1 and w_1^2 in $2G$. Since δ is small, and the inequalities corresponding to w_1 in G are strict, those corresponding to w_1^1 in $2G$ still hold. As for w_1^2: If the protagonist is informed of w_1^2, she knows for sure that she will get the highest possible payoff in the whole game if she indeed plays w_1^2, so it certainly is not worthwhile for her to switch. Therefore ρ is indeed a correlated equilibrium of $2G$.

It then follows from Theorem A that the conditional payoff corresponding to w_1^2 is a rational expectation. This conditional payoff is the protagonist's payoff in G at w, which is her highest payoff at any strategy pair.

9. Background and Literature

While the theory of interactive rationality presented here is new, to a large extent it flows naturally from previous developments in game theory. First was von Neumann's (1928) minimax for two-person zero-sum games; this led to Nash's (1951) strategic equilibrium; this, in turn, to correlated equilibrium (Aumann 1974, 1987); and this, to interactive rationality. The really new, crucial, element here is looking at game *situations* rather than games—viewing games from "inside," without common knowledge of the situation—and it is Harsanyi's (1967-8) theory of types that enables us to define this precisely.

Theories of games may be roughly classified by "strength:" the fewer outcomes allowed by the theory, the "stronger"—more specific—it is. In a sense, interactive rationality is strongest, because in a game situation Γ, it tells you exactly what to do: choose the strategy that maximizes your expected payoff given your information. But one may also ask—as we do here—about *all* the rational expectations that can arise from a given game G, which is, of course, a much larger set.

[34] In the notation of Section 3, $s^m = ((s_1, m), s_2, ..., s_n)$ for $m = 1, 2$.

Viewed thus, the Harsanyi-Selten (1987) selection theory, which specifies a single outcome for each game, is the strongest. Next come refinements of Nash equilibrium, like Kohlberg-Mertens (1986); next, Nash equilibrium itself; next, correlated equilibrium; and then, interactive rationality. Weaker is rationalizability (Bernheim 1984, Pearce 1984), and weaker still, the Kadane-Larkey (1982) "theory,"[35] which excludes only strongly dominated strategies.

Two-person zero-sum games constitute an important watershed in this classification; indeed, this research began with the idea of generalizing the value from two-person zero-sum to general games. Up to and including rational expectations, all the above theories yield precisely the value—no more and no less—in two-person zero-sum games. Beyond that, they do not; two-person zero-sum games may have rationalizable outcomes—and a fortiori, undominated strategies—that do not yield the value.

We end this section by briefly discusssing Mariotti (1995). Kadane and Larkey attack game theory because it does not depend exclusively on Bayesian decision theory. Mariotti takes the precisely opposite stand: that there is a "fundamental incompatibility between Bayesian decision theory and game theory" (p.1108, IV).

T	G'
B	$2,2$

Figure 8a
The game G

	L'	R'
T'	$1,7$	$0,0$
B'	$0,0$	$3,3$

Figure 8b
The game G'

	L'	R'
TT'	$1,7$	$0,0$
TB'	$0,0$	$3,3$
B	$2,2$	$2,2$

Figure 8c
The game GG'

Figure 8a represents an extensive game: if Rowena chooses T, then G' is played; otherwise, both players get 2. Mariotti argues[36] that in G, a prudent Rowena might well play B, which assures her 2, whereas if she plays T, she might get only 1—her payoff at a reasonable outcome of G' (the Pareto undominated strict Nash equilibrium (T', L')). Then she would also play B in GG', which is simply the strategic form of G. But in GG', we may first eliminate TT', by strong domination; then L', by weak domination;[37] and then B, as $3 > 2$.

The perspective of game situations resolves the difficulty. In the abstract, (T', L') indeed cannot be ruled out in G'. But if Rowena chose T in G, it's unlikely

[35]The quotation marks are because Kadane-Larkey do not really propose a theory; rather, they attack the existing equilibrium theories.

[36]This is a simplified and more transparent version of the example in Mariotti's Figure 2.

[37]Mariotti uses a slightly different argument for this, but it comes to the same thing.

that she would choose T' in G'. The G' in Figure 8a is a game *situation,* not a game; unlike the abstract G' in Figure 8b, it has a context. When Rowena plays T in G, she is not merely deciding to play G'; she is deciding to play G' *in a situation where she could have gotten* 2 *for sure.* That's an altogether different kettle of fish.

10. Discussion

10.1. Kadane-Larkey, Nash, and Interactive Rationality

Basically, our viewpoint is that of Kadane and Larkey (1982): A rational player should do the best she can, given how she thinks the others will play. In technical terminology, she should maximize her utility, given her subjective probabilities for the other players' choices.

When one thinks about it, it seems hard to disagree with this[38]. The difficulty is not in the position itself, but in what Kadane and Larkey do—and don't do—with it. What they do is to conclude that formal game theory—in particular, Nash equilibrium—is irrelevant in the practical analysis of games; that it should be replaced by disciplines like cognitive psychology, which might help players in estimating probabilities for other players' strategy choices. What they don't do is to bear in mind that *in estimating how the others will play,* a rational player must take into account that the others are—or should be—estimating how *she* will play. The interactive element is of crucial importance in games; by ignoring it, Kadane and Larkey miss the whole point of game theory.

On its face, the apparently circular nature of interactive reasoning, in which each player thinks about how the others think, leads to Nash equilibrium. It is precisely this that motivated von Neumann and Morgenstern to formulate the minimax (or equilibrium) solution of two-person zero-sum games, which led to Nash's concept of equilibrium in general strategic games.

Nash equilibrium is indeed "circular:" In a two-person game, Rowena plays a because it is optimal against b, and Colin plays b because it is optimal[39] against a. But interactive reasoning is *not* circular. Rowena has beliefs about Colin's actions,

[38]Though in fact, many do disagree; for example, Mariotti (1995).

[39]In the interpretation of Aumann and Brandenburger (1995) (henceforth A-B), a mixed strategy of Rowena represents Colin's beliefs about her action, and vice-versa. So in a mixed equilibrium (α, β), Rowena believes β about Colin's actions, and Colin believes α about Rowena's actions, where the actions in β's support are optimal against α, and those in α's support are optimal against β.

about his beliefs about her actions, about his beliefs about her beliefs about his actions, and so on; and similarly, Colin has beliefs about Rowena's actions, about her beliefs about his actions, about her beliefs about his beliefs about her actions, and so on. That's interactive, but it *ain't* circular. It's more like an infinite double helix: The strands wind around each other, but never come together. If one wants interactive *rationality*—not just interactive reasoning—then one must add CKR and CP, which are the substance holding the strands of the double helix together.

Thus, Nash equilibrium demands a good deal more than interactive rationality. To make sense of Nash equilibrium, one needs more than "global" assumptions[40] like CKR and CP; one needs also some mutual[41] or common knowledge of the players' *actions* or beliefs[42]. For Rowena to play optimally against b, she must know what b is. From where does this knowledge come?

The answer, as indicated by Nash himelf [1951], is that Nash equilibrium defines a norm of behavior. If in a certain kind of situation, people usually take certain actions, then the actions are known; so under appropriate rationality assumptions, they should constitute a Nash equilibrium.

To conclude: Interactive rationality, and so rational expectations, apply in arbitrary game situations. Nash equilibrium applies to norms, or other situations in which there is some mutual or common knowledge of the actions or beliefs of the players.[43]

Please see also the first and last paragraphs of Section 6.1.

10.2. Common Knowledge of Rationality and Common Priors

We have defined "interactive rationality" as comprising common knowledge of rationality (CKR) and common priors (CP). While these assumptions are undoubtedly strong[44], they sound appropriate—"ring true," so to speak.

Any reasonably intelligent child knows that tic-tac-toe "is" a draw—that that is the "right" outcome of the game. That is not to say that she would always

[40] Assumptions not depending on a specific game or the actions taken in that game.

[41] *Mutual* knowledge of p means that all players know p.

[42] A-B describe precisely what is needed.

[43] Specifically, under the assumptions of A-B, an n-tuple of rational expectations is a Nash equilibrium payoff.

[44] Possible weakenings are discussed in 10.5. We will not rehash here the conceptual discussion of these assumptions, particularly of CP; interested readers are referred to Aumann (1987, Section 5), to the controversy between Gul (1998) and Aumann (1998), to the characterizations of CP due to Morris (1994), Samet (1998a,1998b), and Feinberg (2000), and to the review by Aumann and Heifetz (2002, Section 9).

necessarily play for a draw; she might play for a win, and lose, even while knowing that the game "is" a draw. In arbitrary two-person zero-sum games, the value captures the idea of "right" outcome; and in arbitrary games, it is captured by the notion of rational expectation.

CKR alone—without CP—calls only for the iterated deletion of strictly dominated strategies.[45] Even in two-person zero-sum games, the remaining expected payoffs include far more than the value.

10.3. Conan Doyle, Morgenstern, and Interactive Rationality

In one of the Sherlock Holmes stories (Conan Doyle 1893), Watson and Holmes, pursued by the archcriminal Moriarty, arrive at Victoria station and jump on a train to Dover. As it pulls out, they see Moriarty, who has just missed the train. Watson is delighted, as they have eluded their pursuer. But Holmes says,

"My dear Watson, you evidently did not realize my meaning when I said that this man may be taken as being quite *on the same intellectual plane as myself* (our emphasis). You do not imagine that if I were the pursuer I should allow myself to be baffled by so slight an obstacle. Why, then, should you think so meanly of him?"

"What will he do?"

"What I should do."

"What would you do, then?"

"Engage a special."

The story continues with Holmes and Watson alighting at Canterbury, an intermediate station. Holmes's anticipation turns out correct—Moriarty did indeed enagage a special, and they watch with satisfaction from behind a pile of luggage as it thunders through Canterbury.

Discussing this episode in the dawn of Game Theory, Oskar Morgenstern (1935) wrote, "Holmes recognizes that Moriarty is very clever. But what if Moriarity had been still more clever, had estimated Holmes' mental abilities better and had foreseen his actions accordingly? Then, obviously, he would have travelled to Canterbury. Holmes again would have had to calculate that, and he himself would have decided to go on to Dover. Whereupon, Moriarity would again have 'reacted' differently."

Morgenstern is of course right that Conan Doyle failed to follow through on his own reasoning. But Morgenstern himself also missed an important point: Mo-

[45]This is closely related to the notion of rationalizability [Bernheim 1984, Pearce 1984].

riarty is not merely "very clever," he is *on the* ***same*** *intellectual plane* as Holmes. Since Holmes knows that he and Moriarty are on the same intellectual plane, so does Moriarty.[46] So if Holmes gets out at Canterbury, so should Moriarty; he need not be "still more clever." The rest of Morgenstern's reasoning also follows.

A beautifully succinct informal characterization of interactive rationality in a game G is that it is what Holmes would do if he were playing G opposite Moriarty. To be "on the same intellectual plane" may be taken to involve not only CKR—which is Morgenstern's point—but also that the players proceed from the same basic assumptions about the world: i.e., the common prior assumption.

11. Some Open Questions and Research Projects

1. Characterize the set of all rational expectations in a game G, in terms of G's payoff matrix.

2. Other than two-person zero-sum games and their close relatives, it does not seem easy to find determined games. Are they *exceptional,* in the sense that the complement is generic (open dense) in the set of all games of a given size and number of players? Note that two-person zero-sum games *are* exceptional.

For two-person 2×2 games, the answer is "no," because all such games that do not possess a pure strategy Nash equilibrium are equivalent to zero-sum games under the transformations described in Proposition C(iv) (Section 2). But the question does apply to larger games, say 3×3 or larger.

3. Calculate the rational expectations in games appearing in various applications, such as auctions.

4. Extend the notion of interactive rationality to extensive games and to games of incomplete information.

5. We noted in 9.2 that CKR and CP are rather strong assumptions. To what extent—if at all—can they be weakened, while still getting results in the spirit of those presented here?

[46]Strictly speaking, Holmes's knowing that he and Moriarty are on the same intellectual plane does not necessarily imply that Moriarty knows it—because of possible differences in information. But it is in the spirit of the story that Moriarty respected Holmes as much as Holmes respected him.

In the case of CKR, the most natural weakening to explore would appear to be the common p-beliefs of Monderer and Samet (1989), about which a good deal is already known. In the case of CP, a possible approach might be to stipulate that all the priors have mutual Radon-Nikodym derivatives that differ from 1 by less than ε; in the finite case, this is like saying that the ratios of the priors for all atoms differ from 1 by less than ε. But unlike the common p-beliefs, this notion has not been previously explored at all, in any context.

12. References

Aumann, R. J. (1974), "Subjectivity and Correlation in Randomized Strategies," *J. Math. Econ.* 1, 67-96.

—— (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica* 55, 1-18.

—— (1998), "Common Priors: A Reply to Gul," *Econometrica* 66, 929-938.

—— and A. Brandenburger (1995), "Epistemic Conditions for Nash Equilibrium," *Econometrica* 63, 1161-1180 (**A-B** in the text).

—— and A. Heifetz (2002), "Incomplete Information," in *Handbook of Game Theory with economic applications, Vol. 3,* R. J. Aumann and S. Hart (eds.), Amsterdam: Elsevier, 1665-1686.

Bernheim, B. D. (1984), "Rationalizable Strategic Behavior," *Econometrica* 52, 1007-1028.

Conan Doyle, Sir Arthur (1893), "The Final Problem," London: *The Strand Magazine.*

van Damme, E. (1987), *Stability and Perfection of Nash Equilibria,* Berlin: Springer.

—— (2002), "Strategic Equilibrium," in *Handbook of Game Theory with economic applications, Vol. 3,* R.J. Aumann and S. Hart (eds.), Amsterdam: Elsevier, 1521-1596.

Feinberg, Y. (2000), "Characterizing Common Priors in the form of Posteriors," *J. Econ. Th.* 91, 127-179.

Gul, F. (1998), "A Comment on Aumann's Bayesian View," *Econometrica* 66, 923-927.

Güth, W., R. Schmittberger, and B. Schwarze (1982), "An Experimental Analysis of Ultimatum Bargaining," *J. Econ. Behav. Org.* 3, 367-388.

Harsanyi, J. C. (1966), "A General Theory of Rational Behavior in Game Situations," *Econometrica* 34, 613-634.

—— (1967-8), "Games of Incomplete Information Played by Bayesian Players I, II, III," *Manag. Sci.* 14, 159-182, 320-334, 486-502.

—— (1977), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations,* Cambridge: Cambridge University Press.

—— and R. Selten (1987), *A General Theory of Equilibrium Selection in Games,* Cambridge, Mass.: MIT Press.

Hillas, J., and E. Kohlberg (2002), "Foundations of Strategic Equilibrium," in *Handbook of Game Theory with economic applications, Vol. 3,* R. J. Aumann and S. Hart (eds.), Amsterdam: Elsevier, 1597-1663.

Kadane, J. B., and P. D. Larkey (1982), "Subjective Probability and the Theory of Games," *Manag. Sci.* 28, 113-120.

Kalai, E., and D. Samet (1984), "Persistent Equilibria in Strategic Games," *Int. J. Game Theory* 13, 129-144.

Kohlberg, E., and J.-F. Mertens (1986), "On the Strategic Stability of Equilibria," *Econometrica* 54, 1003-1037.

Kreps, D., and R. Wilson (1982), "Sequential Equilibria," *Econometrica* 50, 863-894.

Mariotti, M. (1995), "Is Bayesian Rationality Compatible with Strategic Rationality?" *Econ. J.* 105, 1099-1109.

Monderer, D., and Samet, D. (1989), "Approximating Common Knowledge with Common Beliefs," *Games Econ. Behav.* 1, 170-190.

Morgan, J., and M. Sefton (2002), "An Experimental Investigation of Unprofitable Games," *Games Econ. Behav.* 40, 123-146.

Morgenstern, O. (1935), "Perfect Foresight and Economic Equilibrium," in A. Schotter (ed.): *Selected Economic Writings of Oskar Morgenstern,* New York: New York University Press, 1976.

Morris, S. (1994), "Trade with Heterogeneous Prior Beliefs and Asymmetric Information," *Econometrica* 62, 1327-1347.

Myerson, R. B. (1997), "Dual Reduction and Elementary Games," *Games Econ. Behav.* 21, 183-202.

Nash, J. F. (1951), "Non-cooperative Games," *Ann. Math.* 54, 286-295.

von Neumann, J., (1928), "Zur Theorie der Gesellschaftsspiele," *Math. Annalen* 100, 295-320.

—— and O. Morgenstern (1944), *Theory of Games and Economic Behavior,* Princeton: Princeton University Press.

Pearce, D. G. (1984), "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* 52, 1029-1050.

Rosenthal, R. (1982), "Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox," *J. Econ. Th.* 25, 92-100.

Roth, A., M. Okuno-Fujiwara, V. Prasnikar, and S. Zamir (1991), "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *Amer. Econ. Rev.* 81, 1068-1095.

Samet, D. (1998a), "Iterated Expectations and Common Priors," *Games Econ. Behav.* 24, 131-141.

—— (1998b), "Common Priors and Separation of Convex Sets," *Games Econ. Behav.* 24, 173-175.

Selten, R. (1975), "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *Int. J. Game Th.* 4, 25-55.

Shapley, L. S. (1964) "Some Topics in Two-Person Games," in *Advances in Game Theory,* Ann. of Math. Studies 52, M. Dresher, L. S. Shapley, and A. W. Tucker (eds.), Princeton: Princeton University Press, 1-28.

Département des Sciences Économiques
de l'Université catholique de Louvain
Institut de Recherches Économiques et Sociales

Place Montesquieu, 3
1348 Louvain-la-Neuve, Belgique

ISSN 1379-244X D/2005/3082/021