JOURNAL OF ECONOMIC THEORY 59, 257–274 (1993)

Common Belief and the Theory of Games with Perfect Information*

PHILIP J. RENY

Department of Economics, University of Western Ontario, London, Ontario N6A 5C2, Canada

Received March 9, 1990; revised February 20, 1992

The statement "it is common belief that player i is Bayesian rational" (where Bayesian rational means an expected utility maximizer) is defined for two-person games with perfect information. It is shown that in most such games it is not possible for a theory to postulate the Bayesian rationality of all players and to be common belief. This bears directly upon the salience of standard solution concepts such as subgame perfect, and sequential equilibria as well as upon the extensive form rationalizability theories of D. Bernheim (*Econometrica* **52**, 1984, 1007–1028) and D. Pearce (*Econometrica* **52**, 1984, 1029–1050) which rely heavily on the common belief of Bayesian rationality. *Journal of Economic Literature* Classification Number: C72.

1. INTRODUCTION

What is a theory of games? From the classical point of view at least, it is a description of "rational" behavior. Equipped with a book entitled "Theory of Games," any individual, in any strategic situation, need only consult the book in order to make a rational decision. Of course, any description of rational behavior which claims to be complete must be immune to defections from it. That is, it must never be to one's advantage to behave in a manner that the theory deems irrational. But in order to check this, one must able to evaluate the effect of *not* conforming to the theory. Thus, the theory must provide a description of the (rational) behavior of the others upon the one's defection. Indeed, because no player

* This paper contains the essential elements of the first chapter of my Ph.D. thesis. I thank my advisor, Hugo Sonnenschein, and the members of my committee, Joe Stiglitz and Barry Nalebuff, for their valuable comments and encouragement. I also thank Robert Aumann, Giacomo Bonanno, Ray Farrow, Faruk Gul, Arthur Robson, Ariel Rubinstein, Matt Spiegel, Tommy Tan, Sergio Werlang, and Robert Wilson for helpful discussions. I am also indebted to an associate editor and especially to an anonymous referee for providing comments which substantially improved the paper. Support from the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

0022-0531/93 $5.00

can ensure that the others conform to it, it is incumbent upon the theory to provide that player with a rational decision *whether or not* the others conform. In the words of von Neumann and Morgenstern,

> [The] description [of rational behavior] must include rules of conduct for all conceivable situations—including those where 'the others' behaved irrationally, in the sense that the theory will set for them. [12, Chap. I, Sect. 4.1.2]

This is the view we shall adopt here. A *theory of games* provides any player with a rational decision whenever a decision must be made, regardless of whether or not the others have made decisions consistent with the theory. Throughout, the phrase "theory of games" will be used in this sense. The purpose of the present paper is to outline some of the constraints that bear upon any theory of games. In particular, it is shown that no theory of finite two-person games with perfect information, unless it applies only to a rather small subset of such games, can contain both of the following assumptions without logical contradiction:

(a) each player believes that his opponent is an expected utility maximizer, so long as this is consistent with the hostory of play,

(b) both players believe (a), both players believe that the other believes (a), etc.

The consequences of this are rather striking.

Suppose, for instance, that the theory does not contain assumption (a). This raises the potential for a player to believe that his opponent is not a maximizer, even though each of the opponent's past choices is consistent with expected utility maximization. On the other hand, if (b) is absent, then although both players believe, when consistent with the history of play, that their opponent is a maximizer, they may not believe that the other believes this (or that the other believes that the other believes this, etc). That is, if one of (a) or (b) is absent from the theory, then there may be situations arising in the course of play that are consistent with each player being an expected utility maximizer, yet expected utility maximization is *not* common knowledge (Aumann [1], Lewis [11]). Why is this important?

Kreps, Milgrom, Roberts, and Wilson [9] provide an extremely elegant model of rational cooperation in the finitely repeated prisoners' dilemma. They show that if there is even the slightest possibility that one of the players is not an expected utility maximizer, then rational play need not consist solely of defecting at each stage. Indeed, if there is only the possibility that one of the players *does not know* that the other is a maximizer (even though he is), again cooperation can result in a rational manner. And generally, so long as expected utility maximization (henceforth EUM) is not common knowledge—and the game is long enough—

rational play need not coincide with backward induction in the finitely repeated prisoners' dilemma or even any game with perfect information and no indifference among terminal nodes.

The consequence of omitting one of (a) or (b) from a theory of games is now clear. Without both (a) and (b), there may be positions in a game consistent with both players being expected utility maximizers in which EUM is not common knowledge. But as shown by Kreps *et al.*, the players, from that point on, may then rationaly choose *not* to play according to backward induction. But this means that playing according to backward induction *before* this position arises need not be the only rational play, since this relies on backward induction play in the future, and future play need not be according to backward induction. Thus, without both (a) and (b)—and it will be shown that both cannot be present—a theory of games might admit non-backward induction outcomes as being consistent with rational play.

Of course, just because both (a) and (b) are absent from a theory does not mean that non-backward induction outcomes are consistent with it. Other assumptions making up the theory may rule out all but backward induction play. The point is that (a) and (b) seem entirely natural and that alternative assumptions geared toward backward induction outcomes are typically ad hoc or unconvincing.

Since (as we shall argue) one of (a) or (b) must be absent from any theory of games, thereby potentially allowing non-backward induction outcomes, the plausibility of standard equilibrium concepts such as subgame perfection (Selten [18]), perfection (Selten [19]), sequential equilibrium (Kreps and Wilson [10]), and others is called into question. Moreover, even the more fundamental ideas put forward independently by Bernheim [5] and Pearce [13] concerning extensive form games are at issue. Both of their notions of extensive form rationalizatility yield backward induction outcomes in games with perfect information and no indifference among terminal nodes. In fact, the extensive form theory developed by Pearce is based almost exclusively on the joint assumptions (a) and (b).

In order to carry out the task at hand, formal meaning shall be given to the statement "during the course of play, it is possible for each player to believe the other is an expected utility maximizer, for each to believe that the other believes this, etc." Referring to expected utility maximizers as *Bayesian rational players*, we thus intend to define the statement "during the course of play, it is possible for Bayesian rationality to be common belief." Note that together (a) and (b) imply that so long as neither player's Bayesian rationality has been contradicted through the course of play, Bayesian rationality is and remains common belief. We shall employ our definition to show that for "most" two-person games having perfect information, this condition on common beliefs simply cannot hold.

It is our feeling that the difficulties others have expressed with backward induction (see for instance Basu [2, 3], Binmore [6], Rosenthal [16]) are intimately connected with the impossibility of Bayesian rationality being common belief. Indeed, every example of a game (having perfect information) we know of used to cast doubt upon the logic of backward induction is such that during the course of some play of the game, Bayesian rationality cannot be common belief according to our definition.

Section 3 introduces the environment and notation, and formally defines those positions in a game that are consistent with the common belief of Bayesian rationality. Section 4 defines the notion of a belief-consistent game as (essentially) one in which (a) and (b) can be jointly assumed without contradiction. Our main theorem shows that few games are belief-consistent.

Section 5 briefly discusses the significance of our result in terms of its implications regarding the backward induction argument, Bernheim [5] and Pearce's [13] notions of extensive form rationalizability, and extensive form game theory in general. A discussion of the finitely repeated prisoner's dilemma is also provided there. We begin with an example.

2. An Example

Consider the following two-person game having perfect information.[1] A referee comes equipped with n dollars and places one in front of players one and two. Player one can take the dollar thereby ending the game, or he can leave it. If he leaves it, the referee places a second dollar in front of the players. Player two now has the opportunity to take the two dollars and end the game or not, in which case the process repeats. In general, at the kth stage of the game, the referee adds one dollar to the pot bringing its total to k dollars. If k is odd (even), player one (two) may take the k dollars and end the game, or leave it. Players' payoffs are assumed strictly increasing in dollars. Finally, should the game continue until the nth stage and the player whose turn it is decides to leave the n dollars, it is then given to the other player. Call this game TOL(n) (Take it Or Leave it).

TOL(n) for n odd is depicted in Fig. 1.

Clearly, the unique subgame-perfect equilibrium involves player one taking the first dollar with probability one regardless of the value of n.

We now move to the problem of common belief. It is enough to consider TOL (3) (see Fig. 2). Consider the rather leading question: Is it possible for Bayesian rationality to be common belief after player one leaves the first dollar (i.e., at two's node)? We proceed informally to conclude that the

[1] This example is similar in spirit to Rosenthal's [16].

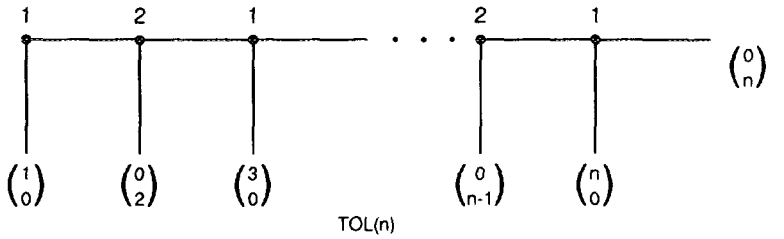


FIGURE 1

answer is no. The argument is really quite straightforward and proceeds by contradiction.

Suppose that Bayesian rationality were common belief at player two's node. Player two, believing that player one is Bayesian rational (i.e., an expected utility maximizer), must believe that at stage 3, player one will take the three dollars, leaving player two with nothing. A rational player two would respond to this by taking the two dollars at the second stage of the game leaving player one with zero. Hence, if at player two's information set it is the case that

(i) player two is Bayesian rational, and

(ii) player two believes that player one is Bayesian rational,

then player two will take the two dollars leaving player one with zero. But since Bayesian rationality is common belief, it must be the case that in particular,

(i') player one believes that player two is Bayesian rational, and

(ii') player one believes that player two believes that player one is Bayesian rational,

i.e., player one believes (i) and (ii) above. Finally, however, this implies that player one believes that player two will take the two dollars leaving

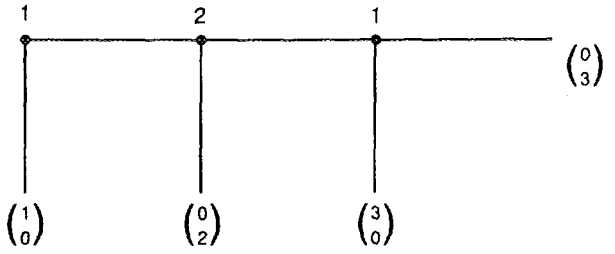


FIGURE 2

player one with zero and rendering player one's choice not to take the dollar in the first round (recall that player two's information set has been reached) an irrational (non-expected utility maximizing) one. Hence, player two must believe that player one is *not* Bayesian rational which contradicts our original assumption and completes the argument.

A similar argument shows that in TOL(n), as soon as player one leaves the first dollar it is not possible for Bayesian rationality to be common belief. That is, the common belief of Bayesian rationality fails precisely at nodes off the backward induction path in TOL(n). This feature of TOL(n) is, however, somewhat misleading. An example in the next section is provided in which common belief of Bayesian rationality is possible at a node which is *off* the backward induction path. More on this is provided in the remark preceding the Proposition.

3. Bayesian Rationality and Common Beliefs

We shall restrict our attention throughout to two-person finite extensive form games with perfect information and assume the reader to be familiar with the formal definition (see Selten [19]). The set of decision nodes, X, is partitioned into X_1 and X_2, where X_i is player i's set of decision nodes, and X is partially ordered by $>$. We shall say that x *weakly follows* y when $x > y$ or $x = y$ and write $x \geqslant y$. M_i denotes player i's set of mixed strategies. For any $x \in X$ and $m_i \in M_i$, we say that m_i reaches x (or x is reachable by m_i) if for some $m_{-i} \in M_{-i}$, (m_i, m_{-i}) assigns a path through x positive probability. Given x and $x' \in X$, x *immediately precedes* x' if $x < x'$ and $\forall y \in X$, $y < x' \Rightarrow y \leqslant x$. In all diagrams, payoffs will be represented by the transpose of the vector (a, b), where a, b denotes 1's, 2's payoff, respectively.

We assume (as in Pearce [13]) that players, when required to make a decision in the face of uncertainty, are able to obtain subjective probabilities over the uncertain events. Decision-making under uncertainty is thus reduced to a Bayesian decision problem as in Savage [17]. Hence, associated with each decision node of player i, $x \in X_i$, is a conjecture $c_i(x)$ about the mixed strategy his opponent, $-i$, is using. Although $c_i(x)$ should be thought of as a probability measure over M_{-i}, Pearce [13] has shown (Appendix A, Lemmas 1 and 2) that for the purpose of expected utility calculation, $c_i(x)$ may be taken to be a member of M_{-i}. In fact, Pearce shows that if player i's beliefs about his opponent's behavior have support $A \subseteq M_{-i}$, then we may take $c_i(x) \in \bar{A}$, where $\bar{A}$ denotes the convex hull of A. We will also need to speak of i's beliefs at $x \in X_{-i}$, about the strategy $-i$ is using. Consequently, $c_i(x) \in M_{-i}$ will denote i's beliefs about his opponent's mixed strategy at $x \in X$ (as opposed to just $x \in X_i$). Let

$c_i \equiv \{c_i(x)\}_{x \in X}$ denote then player i's conjecture profile, and C_i the set of i's conjecture profiles. Given a conjecture profile c_i for player i, $x \in X_i$ and $m_i \in M_i$, we say that m_i *is a best response at* x *given* $c_i(x)$ if for all $m_i' \in M_i$ inducing the same probability distribution over choices at every $x' \in X_i$ such that $x' \not\geq x$, i's expected utility playing m_i given $c_i(x)$ is at least as large as i's expected utility playing m_i'.

DEFINITION 1. $(m_i, c_i) \in M_i \times C_i$ satisfies *Bayesian rationality* if

(i) $\forall x \in X_i$, $c_i(x)$ reaches x

(ii) $\forall x \in X_i$, $\forall x' \in X$ such that x immediately precedes x', and m_i reaches x', $c_i(x) = c_i(x')$

(iii) $\forall x \in X_i$ reachable by m_i, m_i is a best response at x given $c_i(x)$.

If (m_i, c_i) satisfies Bayesian rationality, we shall also say that m_i is a *Bayesian rational* strategy. Condition (i) above simply asks that a player's conjecture be consistent with the information he possesses. Reflected in (ii) is the idea that a player's conjecture about his opponent's behavior, which includes beliefs about what his opponent would do were he to make any one of the choices currently available to him, not change as a result of having actually made a particular choice. This is a particularly weak form of Bayesian updating. Finally, (iii) asks for exected utility maximizing behavior at all decision nodes not precluded by i's strategy.

We wish now to define what it means for Bayesian rationality to be common belief upon reaching a particular node x, or rather, for this simply to be *possible* once x is reached. For this to be possible, there must be a pair of Bayesian rational strategies m_1, m_2 reaching x, and consistent with each of the statements: i believes that $-i$ is Bayesian rational at x, i believes that $-i$ believes that i is Bayesian rational at x, ... for $i = 1, 2$.[2]

Now, suppose that we have determined, for both players, sets of strategies $R_i^x(n) \subseteq M_i$, $i = 1, 2$, such that $R_i^x(n)$ includes *all Bayesian rational strategies* that reach x, and are also consistent with each of the statements: i believes that $-i$ is Bayesian rational at x, ... of length n or less. When will a Bayesian rational strategy, m_i, reaching x, also be consistent with each of the statements: i believes that $-i$ is Bayesian rational at x, ..., of length $n+1$ or less? Precisely when there is a conjecture profile, c_i, so that (m_i, c_i) satisfies Bayesian rationality and i's conjecture at x,

[2] The iterative definition we are preparing for bears a close resemblance to the iterative techniques employed in Bernheim's [5] and Pearce's [13] independent investigations of the implications of the common knowledge of expected utility maximization on behavior in general games. They did not address the question of whether or not this kind of common knowledge was possible in the extensive form, however. Tan and Werlang [20] provide an alternative to Aumman's [1] definition of common knowledge based upon a similar iterative procedure applied to beliefs.

namely $c_i(x)$, reflects i's belief that: $-i$ is Bayesian rational at x and $-i$ believes that i is Bayesian rational at $x, \ldots$, for all such statements up to length n. But this means that i's conjecture, $c_i(x)$, about $-i$ must have support contained in $R^x_{-i}(n)$. Therefore, a Bayesian rational strategy, m_i, reaching x, is also consistent with each of the statements: i believes that $-i$ is Bayesian rational at $x, \ldots$, of length $n+1$ or less if and only if for some conjecture profile c_i, (m_i, c_i) satisfies Bayesian rationality, and $c_i(x) \in \overline{R^x_{-i}(n)}$. But this then defines $R^x_i(n+1)$ in terms of $R^x_{-i}(n)$ as:

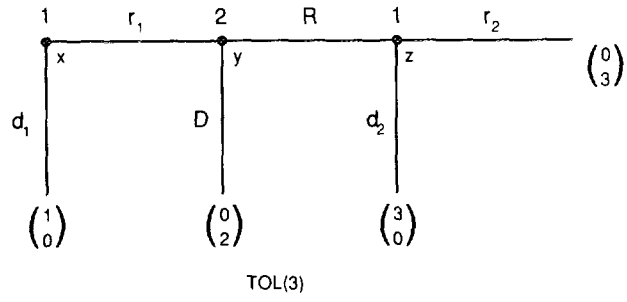
$$R^x_i(n+1) = \{m_i \in M_i \mid \text{for some conjecture profile } c_i,$$
$$\text{(i)} \quad m_i \text{ reaches } x,$$
$$\text{(ii)} \quad (m_i, c_i) \text{ satisfies Bayesian rationality,}$$
$$\text{(iii)} \quad c_i(x) \in \overline{R^x_{-i}(n)}\}. \qquad (*)$$

Finally, note that $R^x_i(1)$ for $i = 1, 2$ is simply the set of Bayesian rational strategies for i that reach x. With $R^x_1(1)$ and $R^x_2(1)$ pinned down, $R^x_i(n)$ is inductively determined by $(*)$ for $i = 1, 2$ and every $n \geq 1$. The definition to follow is now entirely natural:

DEFINITION 2. A node $x \in X$ is *consistent with the common belief of Bayesian rationality* if $\bigcap_{n=1}^{\infty} R^x_i(n) \neq \varnothing$ for $i = 1, 2$.

We now apply Definition 2 to TOL(3) and formally show that y is not consistent with the common belief of Bayesian rationality (henceforth CBR) (see Fig. 3). Since $R^y_1(1)$ is the set of player 1's Bayasian rational strategies reaching y, it consists precisely of those mixed strategies for 1 that induce a behavioral strategy giving positive weight to r_1 at x and all weight to d_2 at z. Since at y, both D and R are best responses for 2 to some strategy in M_1, $R^y_2(1) = M_2$. Using $(*)$ we then have that $R^y_2(2) = \{D\}$, since 2's conjecture at y about 1's future play must be d_2, and D is 2's unique best response at y given this. But this renders $R^y_1(3)$ empty since 1's conjecture at y must have support in $R^y_2(2)$ and so must be D. But reaching y—which every strategy in $R^y_1(3)$ must do—is then inconsistent with 1's Bayesian rationality. Since $R^y_1(3)$ is empty, we conclude that y is not consistent with CBR.

A straightforward application of Definition 2 shows that in TOL(n), only the origin is consistent with CBR. Hence in TOL(n), once player one leaves the first dollar it is never again possible that Bayesian rationality be common belief. This is true despite the fact that every decision node in TOL(n) is consistent with the Bayesian rationality of both players, and only at the last decision node does a player ever have a strictly dominant choice. We conclude that any theory of games that includes a description of rational



TOL(3)

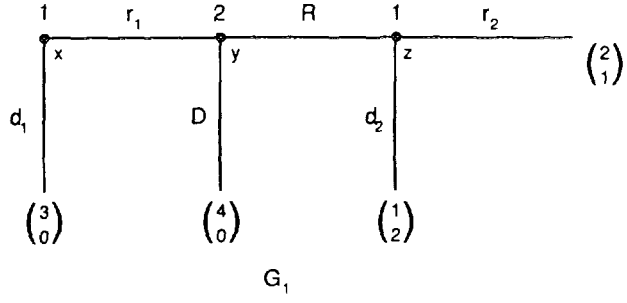
FIGURE 3

play in TOL(n) for $n > 3$, cannot contain both assumptions (a) and (b) from the Introduction, since the necessary common beliefs are precluded by the game's structure.

Although somewhat peripheral to our main line, the following example provides some additional insight into the nature of Definition 2. It might well be thought that in a two-person finite extensive form game with perfect information, if some node is not consistent with CBR then from that point on no node will be. This turns our to be false (at least according to our definition). Consider the game G_1 shown in Fig. 4.

Calculations verify that y is not consistent with CBR, yet z is, even though z succeeds y. The intuition behind this is straightforward. At y, the only way player 2 can believe that player 1 is Bayesian rational, is if 2 believes that 1 believes that 2 will play D at y with high enough weight. (Otherwise 1's best response would involve playing d_1 with probability one.) But if 2 is Bayesian rational, 2 will play D with probability zero. Thus, if at y, 2 believes that 1 is Bayesian rational, then 2 must also believe that 1 believes that 2 is *not* Bayesian rational. Consequently, y is not consistent with CBR.

How then can z be? Since Definition 2 is concerned only with whether



G_1

FIGURE 4

or not it is *possible* for rationality to be common belief at z, we need only provide a single example of beliefs for which this is so. For instance then, suppose 1's beliefs at x about 2 are that 2 plays D with probability 1. Hence, 1 does not believe that 2 is Bayesian rational. Being Bayesian rational himself, player 1 accordingly chooses r_1 so that y is reached. However, 1 is wrong about 2. Player 2 is Bayesian rational and therefore chooses R. This is a surprise to 1 who previously placed probability zero on this event. Accordingly, player 1 is free to revise his beliefs about 2 in any manner consistent with z being reached. In particular, suppose that after updating, player 1 believes that 2 employed the strategy R. In this game, this is equivalent to 1 believing at z that 2 is Bayesian rational. Since each player can explain z having been reached in this manner and each can believe the other believes this, etc., and this explanation is consistent with both players' Bayesian rationality, z is consistent with CBR.

Remark. If (s_1^*, s_2^*) is a subgame perfect equilibrium, then a straightforward inductive argument establishes that for any node, x, along the equilibrium path, $s_i^* \in R_i^x(n)$ for $i = 1, 2$, and all $n \geqslant 1$. Consequently, every node along a subgame perfect equilibrium path is consistent with CBR. The converse, however, fails. As illustrated in game G_1, node z is consistent with CBR, yet off the unique SPE path.

We now introduce an alternative to Definition 2 which, owing to its simplicity, will be useful in the sequel.[3] Given nonempty subsets $R_i^* \subseteq M_i$ for $i = 1, 2$, the pair (R_1^*, R_2^*) will be called a *jointly rational belief system* (JRBS) *for* $x \in X$ if for $i = 1, 2$, m_i is a member of R_i^* precisely when there is a conjecture profile c_i such that: (i) m_i reaches x; (ii) (m_i, c_i) satisfies Bayesian rationality; and (iii) $c_i(x) \in \overline{R^*}_j$.

PROPOSITION. *A node $x \in X$ is consistent with CBR if and only if there is a jointly rational belief system for x.*

Proof. See the Appendix.

4. BELIEF-CONSISTENT GAMES

As exemplified by TOL(n), it is clear that a theory containing assumptions (a) and (b) cannot be applied without contradiction to all games. It is of some interest to know in particular to which two-person games having perfect information such a theory can be applied. Call a two-person game having perfect information and no indifference among terminal nodes, *simple*.

[3] I owe thanks to an anonymous referee, who suggested this alternative definition.

Suppose that a particular theory of games is proposed. Furthermore suppose that it contains (among perhaps others) assumptions (a) and (b) suitably reproduced below:

(a) each player, so long as it is consistent with the history of play, believes that his opponent is Bayesian rational, and

(b) assumption (a) is common belief among the players.

Our aim is to show that such a theory can be applied to only "very few" simple games. To make the point even more forcefully, we shall, in fact, show that an assumption even *less* stringent than (a), together with (b), suitably modified, can be consistently applied only rarely. Thus, consider the following assumptions:

(a′) *if effective*, each player, so long as it is consistent with the history of play, believes that his opponent is Bayesian rational, and

(b′) assumption (a′) is common belief among the players.

When is it ineffective for a player to believe that his opponent is Bayesian rational? When that player has a strategy that is best for him *no matter how his opponent behaves.* In simple games, this is the case precisely when a player has a strictly dominant choice at the current decision node.

We shall call a game *belief-consistent* if (a′) and (b′) can together be applied to it without logical contradiction. In order to formally define a belief-consistent game, we will need some additional notation. At i's decision node x, a choice c (identified by an immediate successor of x) is *strictly dominant* if there is a pure strategy $s \in M_i$ reaching c such that for any pure strategy $s' \in M_i$ reaching x but not c, and any $m_{-i} \in M_{-i}$ reaching x with probability one, (s, m_{-i}) yields i a strictly higher payoff than (s', m_{-i}). We say that i's decision node x is *consistent with* i's *Bayesian rationality* if i has a Bayesian rational strategy reaching x.

Call a decision node, x, *relevant* if it is consistent with both players' Bayesian rationality and no strictly dominant choice is available at x. Let X_r denote the set of all relevant decision nodes. Thus, for a fixed game, X_r contains those decision nodes that could be reached by Bayesian rational players and at which it is effective for the owner to believe that his opponent is Bayesian rational.

Focusing attention merely on the set of relevant decision nodes rather than on all those that are consistent with both players' Bayesian rationality corresponds to our desire to incorporate the weaker assumption (a′) rather than the stronger one (a). Since the set of simple games that (a′) and (b′) can be applied to without contradiction is no smaller than (and is, in fact, strictly larger than) the set that (a) and (b) can be applied to, showing that

the former is in some sense small, shows also that the latter is. We now move toward the definition of a belief-consistent game.

In order for (a′) and (b′) to apply without contradiction to a game, it certainly must be the case that every relevant decision node is consistent with CBR. But more than this must be true. If both x and y are relevant, and say, neither succeeds the other, then if x were reached, not only would (a′) and (b′) imply that Bayesian rationality were common belief at x, but they would also imply that *at* x it is common belief that *had* y *been reached* Bayesian rationality would have been common belief at y.

To see that this has force, consider Fig. 5.

Both x and y are relevant in G_2, and both are consistent with CBR. Yet, we wish to argue (informally) that this game is not belief-consistent. Assumptions (a′) and (b′) together produce a logical contradiction. To see this note that if Bayesian rationality were common belief at x, and player 2 were Bayesian rational, then 2 would play left at x. Since both the origin and x are relevant, (a′) and (b′) together imply that at the origin, player 1 believes that if he chooses left, reaching x, then player 2 will play left at x. Since this eventually gives player 1 his highest possible payoff of 2, he will choose left at the origin if he is Bayesian rational. Therefore, choosing right at the origin, so that y is reached, is jointly inconsistent with (1) player 1 being Bayesian rational; (2) 1 believing that 2 is Bayesian rational; and (3) it being common belief that were x reached, Bayesian rationality would be common belief. So, if y is reached, player 2 must conclude that one of (1)–(3) fails to hold. But since y is relevant, this violates (a′) or (b′). We conclude that (a′) and (b′) cannot be consistently applied to the above game. This is not because one of x or y is inconsistent with CBR, but because *together* they are inconsistent with CBR.

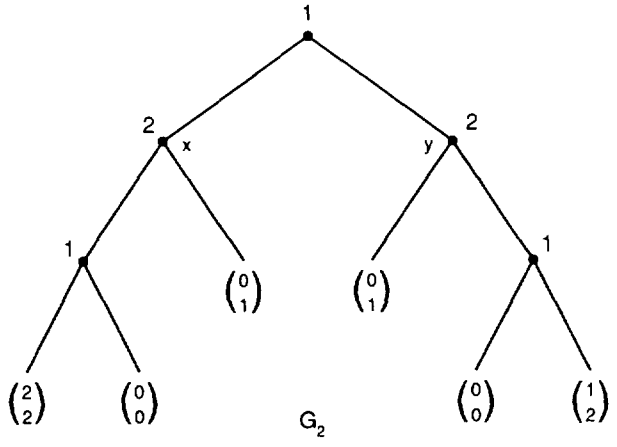


FIGURE 5

In light of this, and drawing on our alternative characterization of Definition 2, consider the following. Fix a nonempty subset of decision nodes $Y \subseteq X$, and a nonempty subset of each player's strategy set $R_i^* \subseteq M_i$, $i = 1, 2$. We will say that (R_1^*, R_2^*) is a *jointly rational belief system* (JRBS) *for* $Y \subseteq X$ if for $i = 1, 2$, m_i is a member of R_i^* precisely when there is a conjecture profile c_i such that (i) m_i reaches some $y' \in Y$, (ii) (m_i, c_i) satisfies Bayesian rationality, and (iii) $c_i(y) \in \overline{R^*}_i$ for every $y \in Y$.

Note that it is condition (iii) that both distinguishes this definition from its predecessor and captures the intuition of the previous example. Condition (iii) ensures not only that at $y' \in Y$ player i's conjecture about $-i$'s strategy is consistent with CBR, it also ensures that at y' it is common belief that *had any other node in* Y *been reached*, i's conjecture about $-i$'s strategy *would have been* consistent with CBR.

DEFINITION 3. A finite two-person game having perfect information is *belief-consistent* if there is a jointly rational belief system for the set of relevant decision nodes.

Remark. As with Definition 2, there is an equivalent inductive manner of defining belief-consistency. We leave the details to the reader.

We now wish to convince the reader that very few two-person finite extensive form games having perfect information and no indifference among terminal nodes (that is, simple games) are belief-consistent. The following theorem is intended to be evidence of this:

THEOREM. *A simple game is belief-consistent if and only if for every decision node* x *not on the unique backward induction path,* x *is irrelevant. That is, either*

(i) *a strictly dominant choice is available at* x, *or*

(ii) x *is inconsistent with at least one player's Bayesian rationality.*

Proof. See the Appendix.

Note for instance then that all simple games in which player 1 moves first, and player 2 moves second, ending the game, are belief-consistent. This may be one reason for the salience of subgame perfect (i.e., backward induction) outcomes over imperfect Nash outcomes in such games. Of the examples presented here, only TOL(n) for $n = 2$ and G_1 are belief-consistent.

5. Concluding Remarks

The analysis of Section 4 shows that assumptions (a) and (b) (indeed, (a′) and (b′)) cannot be jointly applied to any game taken from a large class of two-person finite extensive form games with perfect information. But the lesson of Kreps *et al.* [9] is that if in some subgame Bayesian rationality is not common belief, which the absence of (a) and (b) would allow, Bayesian rational players need not play according to the standards of backward induction. However, if in such subgames backward induction does not apply, then it does not apply at nodes preceding these subgames either—playing according to backward induction at a particular decision node is not necessarily Bayesian rational if the standards of backward induction are not upheld in the part of the tree that follows. Hence, combining our analysis with the message of Kreps *et al.* [9] casts some doubt on the tenet that backward induction is the only "sensible" mode of behavior in two person games with perfect information. Indeed, even assuming that two Bayesian rational players share a common belief of Bayesian rationality at the origin of the game tree does not imply that the backward induction outcome will result (see Reny [14] or Ben-Porath [4]). For instance, in TOL(n) this eliminates only outcomes in which the last mover chooses not to take the money. No other can be ruled out![4]

With respect to the finitely repeated prisoners' dilemma, we feel some additional insights are suggested by our analysis. In particular, the result of Kreps *et al.* [9] supporting rational cooperation rests on an exogenous lack of common belief of Bayesian rationality (in fact there is a positive probability that one of the players is TIT-FOR-TAT). From the analysis of Kreps *et al.*, as we have mentioned, it is clear that cooperation may result among Bayesian rational players whenever Bayesian rationality is not common belief. But if this is the case, then there need not be an exogenously imposed lack of common belief of Bayesian rationality at the beginning of the game for cooperation to result. The reason is that although the finitely repeated prisoners' dilemma is a game with imperfect information, the imperfections are simple enough so that a straightforward modification of Definition 3 can be applied to show that this game is not belief-consistent. Hence, even if Bayesian rationality is common belief at the beginning of the game, by making appropriate choices the players can reach a position in which Bayesian rationality is not (in fact, cannot be) common belief. And by so doing, both may be better off by rationally cooperating thereafter.

[4] Recently, Ben-Porath [4] has characterized those outcomes consistent with the assumptions of Bayesian rationality and CBR at the origin of the tree in simple games. They are precisely those that result upon one round of deleting weakly dominated strategies and then iteratively eliminating strictly dominated strategies.

This can be formally demonstrated. Of course, in order for this to be the case, it must be possible that some of the choices made are *unexpected.* Otherwise the initial common belief of Bayesian rationality would persist. Thus, the type of behavior described here is necessarily non-equilibrium behavior.

It is perhaps worth mentioning that none of the difficulties expressed here, in paricular the impossibility of Bayesian rationality being common belief, arises in simultaneous-move games. This is because players never observe any choice made by their opponents in such games. Hence one is never forced (due to an observation) to believe an opponent is not Bayesian rational or that he believes that you are not Bayesian rational, etc. So, from the point of view of beliefs and beliefs about beliefs... about Bayesian rationality, extensive form games and normal form games (simultaneous-move extensive form games) are markedly different. Whether or not this warrants a difference regarding strategic considerations may be worth keeping in mind when evaluating the notion of invariance of solution concepts (Kohlberg and Mertens [8]) which presumes that the extensive and normal form representations are strategically equivalent.

Since it is not in general possible for Bayesian rationality to be common belief at every information set of every extensive form game (in particular, games with perfect information) we cannot hope to formulate a consistent *theory of games* which is itself common belief among the players and which contains the postulate "all players are Bayesian rational." In particular, since the recent work of Berheim [5] and Pearce [13] on extensive form rationalizability involves heavy use of the common belief of Bayesian rationality within a model which is itself presumed common belief, at the very least a reinterpretation of their (extensive form) analysis is called for. The same goes as well for most refinements of Nash equilibrium.

A *theory of games* therefore, must explicitly allow (and take into account the effects of) the absence of the common belief of Bayesian rationality at some points in some games. In an attempt to deal with this as well as with other fundamental concerns associated with extensive form reasoning, Gul [7] introduces the notion of a τ-theory in which it need never be the case that Bayesian rationality is common belief. Another approach is found in Reny [14, 15], where a refinement of Nash equilibrium that does not insist upon common belief of Bayesian rationality at information sets off the equilibrium path is provided.

APPENDIX

Proof of the Proposition. Suppose that x is consistent with CBR. Then, by definition, $\bigcap_{n=1}^{\infty} R_i^x(n) \neq \varnothing$ for $i = 1, 2$. Let $R_i^* = \bigcap_{n=1}^{\infty} R_i^x(n)$. Owing to

the finite number of pure strategies, it can be shown that for N large enough, $R_i^x(n) = R_i^x(N)$ for all $n \geq N$. So, because $R_i^x(n+1) \subseteq R_i^x(n)$ this implies that $R_i^* = R_i^x(N)$. Putting $n = N$ in $(*)$ then yields:

$$R_i^* = \{m_i \in M \mid \text{for some conjecture profile } c_i,$$

(i) m_i reaches x,

(ii) (m_i, c_i) satisfies Bayesian rationality,

(iii) $c_i(x) \in \overline{R^*}_i\}$.

Hence, (R_1^*, R_2^*) is a JRBS for x.

Conversely, suppose that there is JRBS (R_1^*, R_2^*) for x. By definition of R_i^*, it is the case that $R_i^* \subseteq R_i^x(1)$. We shall establish, by induction, that $R_i^* \subseteq R_i^x(n)$ for every n.

So, suppose for $i = 1, 2$ that $R_i^* \subseteq R_i^x(k)$ for every $k = 1, 2, \ldots, n$. Choose $m_i \in R_i^*$. We wish to show that $m_i \in R_i^x(n+1)$. By the definition of a JRBS, there is a conjecture profile c_i such that (i) m_i reaches x; (ii) (m_i, c_i) satisfies Bayesian rationality; and (iii) $c_i(x) \in \overline{R^*}_i$. But by the induction hypothesis, $R^*{}_i \subseteq R^x{}_i(n)$. Therefore $c_i(x) \in \overline{R^x{}_i(n)}$. But this means then that $m_i \in R_i^x(n+1)$ as desired.

Hence, $R_i^* \subseteq R_i^x(n)$ for every $n \geq 1$ so that $R_i^* \subseteq \bigcap_{n=1}^{\infty} R_i^x(n)$. Since by the definition of a JRBS, R_1^* and R_2^* are nonempty, we conclude that x is consistent with CBR. Q.E.D.

Proof of the Theorem. Suppose that every node off the backward induction path is irrelevant. Let $s_i^* \in M_i$ be i's backward induction (pure) strategy. For $i = 1, 2$, setting R_i^* equal to the collection of strategies realization equivalent to s_i^* yields a JRBS for X_Γ.

Conversely, suppose that Γ is belief-consistent. Let then (R_1^*, R_2^*) be a JRBS for X_Γ. We claim that if $x \in X_\Gamma \cap X_i$, and $m_i \in R_i^*$ reaches x, then m_i induces the unique backward induction choice at x. The argument is inductive.

For $x \in X$, let $l(x)$ denote the maximum number of decision nodes weakly following x among all paths through x. If $x \in X_\Gamma$ is a penultimate node (i.e., $l(x) = 1$), the claim is clearly true. So, assume that it is true for all $x \in X_\Gamma$ with $l(x) \leq k$, and suppose that $z \in X_\Gamma \cap X_i$, $l(z) = k+1$, and $m_i \in R_i^*$ reaches z. Hence, there is a conjecture profile c_i such that (m_i, c_i) satisfies Bayesian rationality, and $c_i(y) \in \overline{R^*}_i$ for every $y \in X_\Gamma$. Therefore $c_i(z)$ is a convex combination of $m^1{}_j, \ldots, m^p{}_j \in R^*{}_j$.

Now, consider any successor of z, $y \in X_j$. (If no such y exists then m_i, being Bayesian rational, gives probability one to the unique backward induction choice at z and we are done.) If $y \in X_\Gamma$ and for some $j = 1, \ldots, p$ $m^j{}_j$ reaches y, then because $m^j{}_j \in R^*{}_j$, the induction hypothesis implies (since $l(y) \leq k$) that $m^j{}_j$ induces the backward induction choice at y. Since

j was arbitrary, $c_i(z)$ therefore induces the backward induction choice at every $y \in X_r \cap X_{-i}$ it reaches following z.

Now suppose that $c_i(z)$ reaches $y \in X_{-i} \setminus X_r$. Since $c_i(z) \in \overline{R^*_{-i}}$, y must be consisent with $-i$'s Bayesian rationality. Hence, since y is irrelevant, either y is inconsistent with i's Bayesian rationality, or a strictly dominant choice is available at y. If the latter holds, then since $c_i(z) \in \overline{R^*_{-i}}$, $c_i(z)$ is a convex combination of Bayesian rational strategies and therefore it places probability one on the strictly dominant choice (the backward induction choice) at y.

Hence, if $c_i(z)$ reaches $y \geq z$, then either $c_i(z)$ induces the unique backward induction choice there or y is inconsistent with i's Bayesian rationality.

Since z is consistent with i's Bayesian rationality, $-i$'s choice at any node following z that is inconsistent with i's Bayesian rationality has no effect upon i's best response at z. So, in order to find i's best response at z we may, without loss of generality, suppose that i's conjecture at z, $c_i(z)$, induces the unique backward induction choice at each of $-i$'s decision nodes that is inconsistent with i's Bayesian rationality. But then $c_i(z)$ induces the unique backward induction choice at each of $-i$'s decision nodes that it reaches following z. Hence, i's unique best response at z is the backward induction choice there. Since, by the induction hypothesis, (m_i, c_i) satisfies Bayesian rationality and m_i reaches z, m_i induces the unique backward induction choice at z. This completes the proof of the claim.

Suppose now that $x \in X$ is relevant and not on the backward induction path. We wish to argue to a contradiction. Let y be any node on the backward induction path preceding x such that the backward induction choice at y precludes x from being reached. Suppose that $y \in X_i$. By the claim proven above, any strategy in R_i^* reaching y induces the backward induction choice there and so does not reach x. Since y precedes x, we conclude that no strategy in R_i^* reaches x.

However, choose now any strategy $m_{-i} \in R^*_{-i}$. By definition of a JRBS for X_r, there is an associated conjecture profile c_{-i} such that $c_{-i}(w) \in \overline{R_i^*}$ for every $w \in X_r$. Furthermore, since (m_{-i}, c_{-i}) must satisfy Bayesian rationality, $c_{-i}(w)$ must reach w for every $w \in X_r$. In particular then, $c_{-i}(x)$ both reaches x and is a member of $\overline{R_i^*}$. But this can happen only if some member of R_i^* reaches x, a contradiction. Q.E.D.

REFERENCES

1. R. AUMANN, Agreeing to disagree, *Ann. Statist.* **4** (1976), 1236–1239.
2. K. BASU, Strategic irrationality in extensive games, mimeo, Institute for Advanced Studies, Princeton, 1985.

3. K. BASU, On the non-existence of a rationality definition for extensive games, mimeo, Delhi School of Economics, Delhi, 1988.
4. E. BEN-PORATH, Common belief of rationality in perfect information games, in preparation, Tel Aviv University, Tel Aviv.
5. D. BERNHEIM, Rationalizable strategic behavior, *Econometrica* **52** (1984), 1007–1028.
6. K. G. BINMORE, Modelling rational players, mimeo, London School of Economics and University of Pennsylvania, 1985.
7. F. GUL, Rational strategic behavior and the notion of equilibrium, mimeo, Stanford GSB, 1989.
8. E. KOHLBERG AND J. F. MERTENS, On the strategic stability of equilibria, *Econometrica* **54** (1986), 1003–1037.
9. D. KREPS, P. MILGROM, J. ROBERTS, AND R. WILSON, Rational cooperation in the finitely repeated prisoner's Dilemma, *J. Econ. Theory* **27** (1982), 245–252.
10. D. KREPS AND R. WILSON, Sequential equilibria, *Econometrica* **50** (1982), 863–894.
11. D. LEWIS, "Convention: A Philosophical Study," Harvard Univ. Press, Cambridge, 1969.
12. J. VON NEUMANN AND O. MORGENSTERN, "Theory of Gmes and Economic Behavior," Princeton Univ. Press, Princeton, NJ, 1944.
13. D. PEARCE, Rationalizable strategic behaviour and the problem of perfection, *Econometrica* **52** (1984), 1029–1050.
14. P. J. RENY, "Rationality, Common Knowledge and the Theory of Games," Ph.D. Dissertation, Chap. 1, Princeton University, 1988.
15. P. J. RENY, Backward induction, normal form perfection, and explicable equilibria, *Econometrica* **60** (1992), 627–649.
16. R. W. ROSENTHAL, Games of perfect information, predatory pricing and the chain-store paradox, *J. Econ. Theory* **25** (1981), 92–100.
17. L. SAVAGE, "The Foundations of Statistics," Dover, New York, 1954.
18. R. SELTEN, Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit, *Z. Ges. Staatiswiss.* **121** (1965), 301–324.
19. R. SELTEN, Reexamination of the perfectness concept for equilibrium points in extensive games, *Int. J. Game Theory* **4** (1975), 25–55.
20. T. TAN AND S. WERLANG, "On Aumann's Notion of Common Knowledge—An Alternative Approach," CARESS Working Paper No. 88-09, University of Pennsylvania, 1988.