International Journal of Game Theory (1990) 19: 33–44

On the Non-Existence of a Rationality Definition for Extensive Games

By K. Basu[1]

Abstract: Standard solution concepts, like subgame perfection, implicitly require that players will continue to assume everybody is rational even if this has been revealed to be false by virtue of having reached a node that could not have been reached had all players behaved rationally. Several attempts have been made in the literature to solve this problem. The present paper shows that the problem is insoluble.

Acknowledgements: I am grateful to T.C.A. Anant, Ken Binmore, Bhaskar Dutta, Vijay Krishna, Debraj Ray, Phil Reny, Ariel Rubinstein, Arunava Sen and an anonymous referee for discussions and comments at various stages of the development of this paper. I also benefited from seminars at the Indian Statistical Institute, Delhi, and the London School of Economics.

1 Introduction

In defining rationality in extensive games unreached nodes cause some well-known problems.[2] In games of imperfect information one problem is that of finding suitable probability numbers for the nodes in the initial information set of subgames which are never actually reached. A more general problem which applies to games of both imperfect and perfect information is that standard solution concepts, like subgame perfection, implicitly require that players turn a blind eye to another player's 'irrationality' even if this has been revealed by virtue of having reached a node that could not have been reached had this player behaved rationally.

Attempts to solve this problem seem to run invariably into difficulties. The aim of the present paper is to prove that the problem is, in fact, insoluble. There is a hint of this in some works (see e.g., Binmore, 1987). The aim of this paper is to lend clarity to this debate by precipitating a formal impossibility theorem.[3] The theorem shows that a definition of rational behaviour which is applicable to all extensive

[1] Professor Kaushik Basu, Delhi School of Economics, Delhi – 110007, India, and Department of Economics, Princeton University, U.S.A.

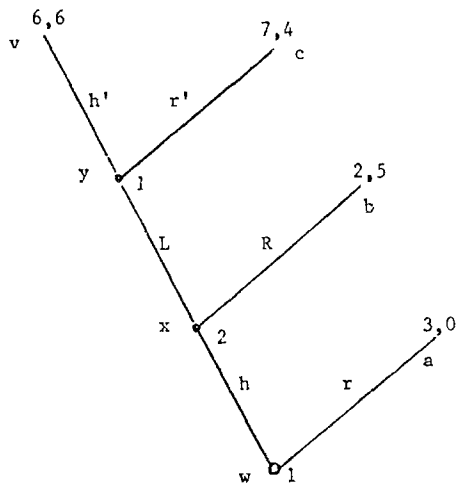
[2] To cite a few references from a large literature: Rosenthal (1981), Kreps and Wilson (1982), Binmore (1987), Reny (1986), Cho (1987), Kreps and Ramey (1987), Dekel and Fudenberg (1987) and Basu (1988).

[3] Abreu and Pearce (1984) have also established some impossibility results for solution concepts but their axiom structure and motivation are quite different from the ones used here.

0020–7276/90/1/33–44 $2.50

games and which does not suffer from the problem of unreached nodes discussed above does not exist. This is because such a definition would run into difficulty with a class of repeated games which include the Prisoner's Dilemma and the games described by Rosenthal (1981) and Reny (1986).

To motivate the problem consider the game Γ, described below, which is a truncated version of Rosenthal's (1981) game. The number at each node denotes the player who has to move there. The game is based on the following principle. At any node, if the relevant player moves left both players get 2 units each, if he moves right he gets 3 and the other gets zero. The pay-off at a terminal node is calculated by counting the number of left moves and right moves that the path entails.



By the backward-induction argument it is clear that at each node, each player will want to move right, and hence the final outcome would be node a with a pay-off of (3,0). It is obvious that all standard solution concepts – Nash, subgame perfection, rationalizability – imply that the final outcome will be node a. But this seems intuitively unacceptable because introspection leads us to believe that that is not the way we would play the game. This comes out more sharply if we consider a longer version of Γ in which there are, say, 100 non-terminal nodes. In such a game, if both players move left throughout they will get a payoff of (200,200). It then seems very unlikely that players will opt to move right and end up with the pay-off (3,0). This is the same problem which arises in the finitely-repeated Prisoner's Dilemma game. For the remaining discussion in this section it may be useful to think of the 100-move game instead of the 3-move one as in Γ.

This conflict between game-theoretic solution and intuition has provoked many attempts to explain formally why players would "cooperate" (in the above example

this entails moving left) in a game. An element of bounded rationality can explain why players may cooperate (Neyman, 1985)[4]. The same may be achieved by allowing for some uncertainty about the 'length' of the game (Basu, 1987). A third and well-known avenue is to allow one player to entertain some doubt about the rationality of the other (Kreps, Milgrom, Roberts and Wilson, 1982).

There is however a more fundamental difficulty with all these escape routes. Suppose two experts in game theory – Messrs. von Neumann and Morgenstern for example – were made to play the Prisoner's Dilemma one hundred times. Most of us would maintain that even here there will not be defection in all hundred games.[5] Here both players are rational and the possibility of one player feeling that the other is irrational is too remote to be treated as an explanation of why cooperation may occur.

In an earlier paper (Basu, 1988), I tried to develop a solution concept based on the argument that if in the first move player 1 cooperates, player 2 will see that a node has been reached which could not have been reached if 1 was 'rational'. Once 2 doubts 1's rationality it may be reasonable for 2 to play cooperatively (at least for some time). But surely 1 can see this and so 1 may actually play cooperatively in the first game. In other words, he plays an 'irrational' move in order to confuse 2. His irrationality is a strategic one.

An important objection can be raised against this. Since 1 manages to do well by feigning 'irrationality', it can be argued that this ought not to be described as 'irrationality' at all.[6] And since others can see this, they will not be misled into believing that 1 is irrational.

It seems to me, however, that the onus of averting this problem lies on the rationality definition being used. This is the line taken in the present paper. The *rationality* definition must be such as to render strategic irrationality ineffective. Once this requirement is coupled with a weak backward-induction axiom we find that it is not just subgame perfection and rationalizability which run into difficulty but so do all existing and potential solution concepts. This is formally established in section 2, and sections 3 and 4 discuss and interpret the axioms and the theorem.

2 The Impossibility Theorem

The term *game* will be used here to refer to any perfect information, no nature-move, two-player, extensive-form game in the sense of Selten (1975) or Kreps and Wilson (1982). A brief resume of the notation follows; for elaboration the reader is referred to Kreps and Wilson.

---

[4] It is worth noting, though, that Rubinstein's (1986) model, using a similar concept of bounded rationality, rules out cooperation in the Prisoner's Dilemma.

[5] An earlier version of this paper figured two *contemporary* game-theorists, but I had to change the names after Ken Binmore assured me, to my dismay, that one of them would defect each time.

[6] Amartya Sen took such a line in his Yrjo Jahnsson lectures in Helsinki in 1987.

Given a node, x, $\alpha(x)$ will be used to denote the *last action* taken to reach x and $p(x)$ to denote the node *immediately preceding* x. In the game described above, $\alpha(y) = L$, $\alpha(b) = R$, and $p(y) = p(b) = x$. Player i's utility function (defined on the terminal nodes) is denoted by u^i. Given a game, Γ, the subgame which consists of node x and all its successors is denoted by Γ_x.

A *strategy* of player i is a mapping s^i which specifies an action at every node where i is supposed to play. S^i is the set of all strategies of i.

A *solution concept*, R, is any mapping defined on the domain of all games such that, given any game Γ, it specifies a non-empty collection of strategy-pairs (one for each player) in the game. Hence, $\phi \neq R(\Gamma) \subset S_1 \times S_2$ where S_i is the set of strategies of i in the game Γ. We refer to $R(\Gamma)$ as the *solution* of Γ (under R).

In the context of individual-rationality discussions it is convenient to impose a technical assumption, which I shall call the *factorability* axiom.

Axiom F: The solution concept, R, must be such that for all Γ, if $(s_1, s_2) \in R(\Gamma)$ and $(\hat{s}_1, \hat{s}_2) \in R(\Gamma)$, then $(s_1, \hat{s}_2) \in R(\Gamma)$.

Axiom F suggests that whether i's strategy s_i is rational or not is independent of what others are using. There are good reasons for defending axioms F (see Bernheim, 1984, and Pearce, 1984), but there is no need to enter this controversy here because in the present context it is easily seen to be irrelevant by taking any of the two following approaches. We could in this paper restrict attention to games in which for each player i the terminal nodes are strictly ordered. That is, $x \neq y$ implies $u_i(x) \neq u_i(y)$. Since our aim is to prove impossibility, nothing is lost by such domain restriction. Secondly, we could interpret standard solution concepts in a way that is consonant with definition 1. If $\hat{R}$ is the Nash solution, we could think of R as a counterpart of this if R is such that, for all Γ, $R_i(\Gamma)$ is the set of all $s_i \in S_i$ such that there exists a strategy pair $(s_1^i, s_2^i) \in \hat{R}(\Gamma)$ where $s_i^i = s_i$.

A solution concept embodies our notion of rationality. If in a game a node is reached that cannot be reached if all players employ strategies in the solution set, it means some player has behaved 'irrationally'. Hence, in any game, Γ, given a solution concept, R, we can at each node, x, specify the set of players, $\Omega^{R\Gamma}(x)$, who have been revealed irrational. The method of doing this is outlined in definition 1. (We shall assume throughout that the solution concept, R, satisfies axiom F. This allows us to write $R(\Gamma) = (R_1(\Gamma), R_2(\Gamma))$.)

Definition 1: Given a solution concept, R, and a game, Γ, we define an irrationality map

$$\Omega^{R\Gamma} : T \to \{\phi, \{1\}, \{2\}, \{1,2\}\}$$

(where T is the set of all nodes in Γ) by induction as follows:

$$\Omega^{R\Gamma}(w) = \phi$$

(where w is the initial node) and for all non-initial node x, we have $\Omega^{R\Gamma}(x) = \{i \mid i \in \Omega^{R\Gamma}(p(x))$ or at $p(x)$ it was i's move and there does not exist $s \in R_i(\Gamma)$ such that $s(p(x)) = \alpha(x)\}$.

Definition 1 tells us that, as in conventional game theory, all players are assumed to be rational to start with (i.e. the set of irrational players, $\Omega^{R\Gamma}(w)$, is empty). At any other node x a player is irrational (i.e. he belongs to the set $\Omega^{R\Gamma}(x)$) if either he is already revealed irrational or at the previous node it was his move and there is no rational strategy of his which could have made him choose the alternative that brings us to node x.

As soon as a solution concept is specified, an implicit rationality definition gets specified as well. It is worth stressing that $R_i(\Gamma)$ is not a definition of rationality but it *embodies* a definition of rationality. The best way to think of a rationality definition is as a statement of moves that a rational player may make at each node. Thus if x is a node in Γ where i has to move and $i \notin \Omega^{R\Gamma}(x)$ then all moves m such that there exists $s \in R_i(\Gamma)$ for which $s(x) = m$ are rational moves. If $i \in \Omega^{R\Gamma}(x)$ then i's move at x of course does not tell us what a rational player would do.

I shall first develop a very weak form of the backward-induction axiom. For this we need some new notation. For any strategy n-tuple s in the game Γ, let $s : \Gamma_x$ represent the restriction of s on the subgame Γ_x. We shall use $\theta(R, \Gamma, x)$ to represent the set of terminal nodes which can be reached in Γ_x by $s{:}\Gamma_x$ for some $s \in R(\Gamma)$.

If Z_1 and Z_2 are subsets of the set of terminal nodes Z in some game Γ, we write $Z_1 >_i Z_2$ if for all $x \in Z_1$ and for all $y \in Z_2$, $u_i(x) > u_i(y)$.

Our backward-induction axiom asserts the following. Suppose in some game y and v are immediate successors of x and it is i's move at x. If all the terminal nodes that can be reached from v by playing strategies in the solution set dominate from i's point of view all the terminal nodes that can be reached from y by playing strategies in the solution set, then if i moves so as to get to y then i is revealed irrational.

Axiom B: The solution concept, R, must be such that for all Γ and for all nodes x, y, v, where y and v are immediate successors of x, if it is i's move at x and $\theta(R, \Gamma, v) >_i \theta(R, \Gamma, y)$, then $i \in \Omega^{R\Gamma}(y)$.

In Basu (1988) cooperation in the Prisoner's Dilemma is explained by allowing a player to make a move that shows him up as irrational and this influences the play of others in a way that could be beneficial to the player. A reasonable objection to this is that other players can surely see through such *strategic* irrationality and hence would not view the original player as irrational at all. To state this formally, I shall abuse the $\theta(.)$ notation developed above a little and use $\theta(s_j, \Gamma, x)$ to denote the set of terminal nodes which can be reached in Γ_x by $s{:}\Gamma_x$ where s is any strategy pair whose j-th component is s_j.

The next axiom states that if in some game, y and v are immediate successors of x and it is i's move at x and for some strategy of the other player belonging to the solution set, all the terminal nodes reachable from y dominate (from i's point of view) all the terminal nodes reachable from v, then if i is described as rational at v, i must described as rational at y.

Axiom S: The solution concept, R, must be such that for all Γ and for all nodes x, y, v where y and v are immediate successors of x, if there exists $s_j \in R_j(\Gamma)$ such that $\theta(s_j, \Gamma, y) >_i \theta(s_j, \Gamma, v)$ and $i \notin \Omega^{R\Gamma}(v)$ then $i \notin \Omega^{R\Gamma}(y)$.

Before stating the next axiom, note that in the approach taken in this paper (in contrast to the traditional, extensive game model) $R_i(\Gamma_x)$ need not be equal to $R_i(\Gamma) : \Gamma_x$. That is, the set of i's possible strategies in the game Γ_x need not coincide with the restriction of i's possible strategies in Γ to the subgame Γ_x. This is because Γ_x considered as a game in itself implies that no one is irrational and i knows this. But in Γ when node x is reached, some players may have been revealed irrational. Even though i himself may not have been revealed irrational, his play in the subgame Γ_x may be influenced by his awareness that there are players who are known to be irrational. It is in this sense that my approach may be described as history-sensitive. How a player plays in a subgame depends on the history of play at the initial node of the subgame.

Since my argument turns on the definition of rationality, it is important to specify what irrationality implies. A simple assumption is that an irrational player is unpredictable.

In terms of the game, Γ, described above, if 1 is known to be irrational at node y, then 1 (it is expected) may play h' or r'.

Axiom U: The solution concept, R, must be such that for all Γ and for any node x if $i \in \Omega^{R\Gamma}(x)$ and it is i's move at x, then $R_i(\Gamma) : \Gamma_x = S_i : \Gamma_x$, where S_i is player i's set of strategies in the game Γ.

The axiom simply states that once a player has been revealed irrational, from there onwards he is treated as unpredictable.

Axioms F, B, S and U are together incompatible.

Theorem 1: There does not exist any solution concept satisfying axioms F, B, S and U.

Proof: This theorem is weaker than theorem 2 proved below. ||

All these axioms can be relaxed without losing the impossibility theorem. This is discussed in a later section but one particular axiom weakening is worth discussing here. It may seem to some that an irrational player should not be treated as totally unpredictable. In that case we may wish to weaken axiom U to state the following. An irrational player is less predictable than a rational player. That is, he may play any strategy that a rational player may play and (wherever possible) there are other strategies which he may play. We no longer require that he may play *any* strategy in his strategy set as in axiom U.

Axiom U:* The solution concept, R, must be such that for all Γ and for any node x, if it is i's move at x, $i \in \Omega^{R\Gamma}(x)$ and $R_i(\Gamma_x)$ is a proper subset of $S_i : \Gamma_x$, then $R_i(\Gamma_x)$ is a proper subset of $R_i(\Gamma) : \Gamma_x$.

The axiom says that if at a node x, i is considered irrational, then the set of strategies that he may be expected to employ thereon (i.e. $R_i(\Gamma) : \Gamma_x$) is a proper super set of the strategies that a rational player may be expected to employ in a similar situation (i.e. $R_i(\Gamma_x)$), assuming, of course that $R_i(\Gamma_x)$ is not already as large as is feasible (i.e. $R_i(\Gamma_x)$ is not equal to $S_i : \Gamma_x$).

To illustrate this axiom suppose Γ, shown above, is actually a subgame of a larger game. In this subgame player 1 has 4 strategies: rr', rh', hr', hh'. Suppose that if 1 is rational, our solution predicts 1 will play rr'. What axiom U* says is that if, at node w, 1 is known to be irrational, the solution must predict a larger set of possible strategies that 1 may employ in this subgame. That is, the predicted set of strategies must include rr' and one or more strategies from the three remaining available.

Actually we could think of an intuitively more correct version of U*. Note that the subgame Γ_x considered as a game in itself treats all players (not just i) as rational. Hence the difference between $R_i(\Gamma) : \Gamma_x$ and $R_i(\Gamma_x)$ in axiom U* is not just that the former takes into account i's irrationality and the latter does not. Instead, the latter treats *everyone* as rational. If we want the only distinction to be i's rationality, then we would have to write axiom U* more elaborately as follows.

The solution concept, R, must satisfy the following condition: for all Γ and for all $x \in T$, if it is i's move at x, $i \in \Omega^{R\Gamma}(x)$ and there exists a game $\hat{\Gamma}$ such that $\hat{\Gamma}_y = \Gamma_x$ and $\Omega^{R\Gamma}(x) - \Omega^{R\hat{\Gamma}}(y) = \{i\}$ and $R_i(\hat{\Gamma}) : \hat{\Gamma}_y$ is a proper subset of $S_i : \hat{\Gamma}_y$ then $R_i(\hat{\Gamma}) : \hat{\Gamma}_y$ is a proper subset of $R_i(\Gamma) : \Gamma_x$.

It will be clear from the proof of theorem 2 that it does not matter formally whether we use axiom U* or its more intuitively appealing variant just described above. Hence, I use the technically simpler axiom U*.

Theorem 2: There does not exist any solution concept satisfying axioms F, B, S and U*.

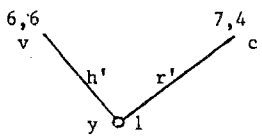
Proof: Let R be a solution concept that satisfies axioms F, B, S and U*. Axiom F allows us to speak of $R_1(\Gamma)$ and $R_2(\Gamma)$ independently. Consider the game, Γ, described in Section I. It follows from definition 2 that either of the following must be true: $\Omega^{R\Gamma}(x) = \phi$ or $\Omega^{R\Gamma}(x) = \{1\}$.

Suppose $\Omega^{R\Gamma}(x) = \phi$. Then $1 \notin \Omega^{R\Gamma}(y)$. By axiom B we know $1 \in \Omega^{R\Gamma}(v)$. Let $s \in R_1(\Gamma)$. Since $1 \notin \Omega^{R\Gamma}(y)$ and $1 \in \Omega^{R\Gamma}(v)$, definition 1 implies $s(y) \neq h'$. Hence, $s(y) = r'$.

This implies $\theta(R, \Gamma, b) >_2 \theta(R, \Gamma, y)$. Hence, by axiom B, $2 \in \Omega^{R\Gamma}(y)$. Since $2 \notin \Omega^{R\Gamma}(x)$, definition 1 implies that if $s' \in R_2(\Gamma)$, then $s'(x) = R$.

By repeating the same argument at w, it can be shown that if $s \in R_1(\Gamma)$, then $s(w) \neq h$. Hence $1 \in \Omega^{R\Gamma}(x)$, which is a contradiction.

Suppose $\Omega^{R\Gamma}(x) = \{1\}$. It follows from definition 1 that $1 \in \Omega^{R\Gamma}(y)$. Consider now the game Γ_y described below:



By axiom B we know that if $s \in R_1(\Gamma_y)$ then $s(y) = r'$. Hence $R_1(\Gamma_y)$ is a proper subset of $S_1 : \Gamma_y$. Since player 1 has to play at y and $1 \in \Omega^{R\Gamma}(y)$, we know by axiom U* that $R_1(\Gamma_y)$ must be a proper subset of $R_1(\Gamma) : \Gamma_y$. Hence the strategy s_1 of player 1 such that $s_1(y) = h'$ must be an element of $R_1(\Gamma) : \Gamma_y$. Hence there exists $s_1 \in R_1(\Gamma)$ such that $v \in \theta(s_1, \Gamma, y)$. Hence $\theta(s_1, \Gamma, y) >_2 \theta(s_1, \Gamma, b)$. Axiom S implies that if $2 \notin \Omega^{R\Gamma}(b)$, then $2 \notin \Omega^{R\Gamma}(y)$. Hence if $2 \in \Omega^{R\Gamma}(y)$, then $2 \in \Omega^{R\Gamma}(b)$. But this is impossible by definition 1, since $2 \notin \Omega^{R\Gamma}(x)$. Hence $2 \notin \Omega^{R\Gamma}(y)$. By definition 1, there exists $s_2 \in R_2(\Gamma)$ such that $s_2(x) = L$. Hence, $\theta(s_2, \Gamma, x) >_1 \theta(s_2, \Gamma, a)$. By axiom S, $1 \notin \Omega^{R\Gamma}(x)$, which is a contradiction. ||

3 Discussion

How does subgame perfection relate to my axioms? It will be shown here that none of my axioms, B, S and U, is *individually* incompatible with subgame perfection (though, of course, taken together, they are). While subgame perfection directly satisfies B and S, its compatibility with U is more indirect and in a sense that needs to be made precise. This is done in the next theorem, but before stating the theorem we need one more definition.

In order to refrain from the technical, but conceptually inconsequential, difficulties that could arise from the fact that a solution concept does not automatically satisfy axiom F, I shall here restrict attention to games in which each player has a strict ordering of the terminal nodes. Such games will be referred to as *strongly ordered games.*

Abusing the $\theta(.)$ notation once more, let us use $\theta(R, \Gamma)$ to denote the subset of the terminal nodes of Γ which can be reached by some strategy combination in $R(\Gamma)$. The solution concepts R and R' will be described as *equivalent* if for all strongly ordered game, Γ, $\theta(R, \Gamma) = \theta(R', \Gamma)$. Now we can state formally our observation about subgame perfection being directly compatible with S and indirectly so with U. It is worth remembering that in stating Theorem 3 we are considering only two-player games with perfect information.

In this section we use $\hat{R}$ to represent the solution concept of subgame perfection. That is, for all Γ, $s^* \in \hat{R}(\Gamma)$ if and only if for all node x, if it is i's move at x, then the terminal node reached in Γ_x by $s^* : \Gamma_x$ is, from i's point of view, as good as any terminal node that may be reached through a unilateral deviation by i from s^*.

Theorem 3: Within the domain of well-ordered games,

(i) $\hat{R}$ satisfies axioms B and S, and
(ii) there exists a solution concept, R, which is equivalent to $\hat{R}$ and which satisfies axioms B and U.

Proof: (i) Consider a strongly-ordered game Γ and nodes x, y and v, where it is i's move at x, y and v are immediate successors of x and $\theta(\hat{R}, \Gamma, y) >_i \theta(\hat{R}, \Gamma, v)$.

Suppose $i \notin \Omega^{\hat{R}\Gamma}(v)$. Hence, there exists $s \in \hat{R}_1(\Gamma)$ such that $s(x) = \alpha(v)$. But it is easy to see that there exists $s' \in S_i$ which would in the subgame Γ_x take i to a superior terminal node than s. For this s' must simply be such that $s'(x) = \alpha(y)$. Hence $s \notin \hat{R}_1(\Gamma)$.

This contradiction establishes that $i \in \Omega^{\hat{R}\Gamma}(v)$. Hence $\hat{R}$ satisfies axiom B. A similar proof can be constructed to show that $\hat{R}$ satisfies S.

(ii) Let us construct a solution concept, R, as follows. For all strongly ordered game, Γ, and for all player, i, in Γ, $R_i(\Gamma)$ is defined in the following way: $s \in R_i(\Gamma)$ if and only if there exists $s' \in \hat{R}_i(\Gamma)$ such that for all x where i has to move and $i \notin \Omega^{\hat{R}\Gamma}(x)$, $s(x) = s'(x)$. It is now easy to check that R is equivalent to $\hat{R}$ and it satisfies axioms B and U. ||

4 Interpretation

The discussion in the previous section suggests that an attempt to break out of the impossibility established in section 2 should be focussed on axiom U or U*. Before going into a discussion of the intuition for accepting or rejecting this axiom, I want to demonstrate that the above theorem can be strengthened by weakening U or U* in an interesting way. For linguistic simplicity, I shall conduct the discussion here in terms of U, though much the same could have been said using U*. Note that what axiom U asserts is that a player who has once been observed behaving irrationally must be, then onwards, treated as completely unpredictable. I feel this is a better assumption than the traditional one which would ignore the revealed irrationality and continue to treat the player as completely rational. It may legitimately be argued however that these two polar assumptions are not the only possible ones; and there may be an escape route inbetween. The aim of this section is to show that these 'in-between' routes are all blocked because theorems 1 and 2 can be generalised to more powerful impossibility theorems.

Let us consider the following weakening of U. Suppose R is the solution concept, and let Γ be a game in which at node x it becomes clear that i is not playing any strategy in $R_i(\Gamma)$. Axiom U would abandon attempts to predict i's play here onwards. However, we could think of an intermediate approach where we have a predefined (first) fall-back rationality definition, R_i^1, such that $R_i^1(\Gamma)$ is a proper subset[7] of S_i, and a player i who violates the rationality definition R_i is expected to play some strategy from the set $R_i^1(\Gamma)$. In other words, $R_i^1(\Gamma)$ is like a second hypothesis about what player i may do, once the first hypothesis (embodied in $R_i(\Gamma)$) is proved false.

At first sight, it seems that modifying axiom U in this manner is a way out. But it is not. This is because it is always possible to construct a game tree, little more elaborate than the one used in the proof of theorem 2, in which there is a node which has the following history. It requires that i violates rationality definition, R_i, and then violates the (first) fall-back rationality definition, R_i^1. What should we assume about i's play here onwards? If we now abandon attempts to predict i's play and treat him as completely unpredictable then an impossibility result can be proved as before. But, of course, there is a way out. This is, to assume that we have a second fall-back rationality definition, R^2, for a player who first violates R and then violates R^1.

But by now it should be evident to the reader that as long as we have a *finite* number of fall-back rationality definitions and concede that a player who has violated all these must be totally unpredicable, we can combine this weaker version of axiom U, with axioms F, B and S to once again establish an impossibility theorem.

Despite the reliance on one principle game in the proofs, it should be obvious that the paradox highlighted in this paper arises in a class of games, including Selten's (1978) chain-store game, the Prisoner's Dilemma, Reny's (1986) game, combinations of these, etc. It should, in principle, be possible to isolate the class of games in which the rationality paradox arises. I did not attempt this in the present paper but it is worth noting that, in a broad sense, the paradox arises in any game which has the following possibility: There is a move such that after a player (call him A) makes the move then (i) if others treat A as irrational, it influences their future play in a way which makes the move a rational one and (ii) if others treat A as rational, then it influences their future play in a way which makes the move irrational for A.

It should be immediately clear that if a certain move is the last one A makes in a game, then that move can never be the basis of the above conflict. It follows that in repeated games with *changing* partners (as in some biological games) the kind of paradox precipitated in this paper does not arises. Subgame perfection may therefore be a solution concept more suitable for evolutionary games.

---

[7] If $R_i^1(\Gamma) = S_i$, then we are back to the *complete* unpredictability assumption of section 2.

Let us now turn to some interpretational issues. As already mentioned, the axiom that is likely to be the most contentious is axiom U or one of its variants. Hence, in the light of the above impossibility theorems we are forced into taking one of the two broad positions. (1) We could treat U or U* as reasonable, and reject the view that every game must have a solution, or that rationality is always definable even in strategic environments. (2) We could reject U and U* and maintain that reasonable solutions can be defined for all games.

(2) seems to direct us towards a model with 'mistakes'. The usual approach is to consider the limit of games with mistakes (as the mistakes vanish). An alternative is a model in which deviations from rationality are treated as possible but of zero-probability, in a measure-theoretic sense. This makes it possible to ignore past deviations. This is an interesting route which has not been adequately explored in the literature[8].

However, as must be evident from the many arguments above, I am at this point inclined to go along with (1). Hence, the present paper considered only a 'no-mistakes' framework. This is not unreasonable since in games in which the *rules* of play are simple (for example, the Prisoner's Dilemma in contrast to Chess) the scope for mistakes is indeed quite limited. In the light of these comments one way of interpreting the theorems in this paper is to treat them as suggesting that rationality cannot be defined unless allowance is made for mistakes. It is not that this paper says that if such an allowance is made, we can define rationality; but it says that the other route is certainly closed.

References

Abreu D, Pearce DG (1984) On the Inconsistency of Certain Axioms on Solution Concepts for Non-Cooperative Games, Journal of Economic Theory, vol. 34, 169-74

Basu K (1987) Modeling Finitely-Repeated Games with Uncertain Termination, Economics Letters, vol. 23, 147-51

Basu K (1988) Strategic Irrationality in Extensive Games, Mathematical Social Sciences, vol. 15, 247-60

Bernheim D (1984) Rationalizable Strategic Behaviour, Econometrica, vol. 52, 1007-28

Binmore K (1987) Modelling Rational Players, Economics and Philosophy, vol. 3, 179-214

Cho I (1987) A Refinement of Sequential Equilibrium, Econometrica, vol. 55, 1367-89

Dekel E, Fudenberg D (1987) Rational Behaviour with Payoff Uncertainty, Working Paper, M.I.T., Cambridge, Massachusetts

Kreps DM, Milgrom P, Roberts J, and Wilson R (1982) Rational Cooperation in the Finitely-repeated Prisoner's Dilemma, Journal of Economic Theory, vol. 27, 245-52

Kreps DM, Ramey G (1987) Structural Consistency, Consistency, and Sequential Rationality, Econometrica, vol. 55, 1331-48

Kreps DM, Wilson R (1982) Sequential Equilibria, Econometrica, vol. 50, 863-94

[8] I owe this suggestion to an anonymous referee.

Neyman A (1985) Bounded Complexity Justifies Cooperation in the Finitely Repeated Prisoner's Dilemma, Economics Letters, vol. 19, 227-29
Pearce DG (1984) Rationalizable Strategic Behaviour and the Problem of Perfection, Econometrica, vol. 52, 1029-50
Reny P (1986) Rationality, Common Knowledge and the Theory of Games, Ph. D. dissertation, Princeton University
Rosenthal RW (1981) Games of Perfect Information, Predatory Pricing and the Chain Store Paradox, Journal of Economic Theory, vol. 25, 92-100
Rubinstein A (1986) Finite Automata Play the Repeated Prisoner's Dilemma, Journal of Economic Theory, vol. 39, 83-96
Selten R (1975) Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games, International Journal of Game Theory, vol. 4, 25-55
Selten R (1978) The Chain-store Paradox, Theory and Decision, vol. 9, 127-59

Received October 1988
Revised version May 1989
Final version September 1989