*Economics and Philosophy*, 7 (1991), 37–65. Printed in the United States of America.

# THE LOGIC OF RATIONAL PLAY IN GAMES OF PERFECT INFORMATION

GIACOMO BONANNO
*University of California, Davis*

For the past 20 years or so the literature on noncooperative games has been centered on the search for an equilibrium concept that expresses the notion of rational behavior in interactive situations. A basic tenet in this literature is that *if a "rational solution" exists, it must be a Nash equilibrium.*[1] The consensus view, however, is that not all Nash equilibria can be accepted as rational solutions. Consider, for example, the game of Figure 1.

Both $(B, Y)$ and $(A, X)$ are Nash equilibria. The strategy pair $(B, Y)$, however, is usually rejected as a rational solution on the grounds that it is sustained by player II's threat to play $Y$, which is not a credible threat, since a rational player II would not choose $Y$ if her decision node were actually reached.

It is examples like this one that have stimulated the search for a

This is a substantial revision of a Working Paper with the same title which appeared in 1987. I am grateful to Michael Bacharach for many detailed and illuminating comments on the first draft of this article and to John Roemer and Roberto Lucchetti for their encouragement and very helpful suggestions. I have also benefited greatly from the comments and criticisms of Jerry Cohen, Robin Cubitt, Daniel Hausman, Aanund Hylland, David Kreps, Paul Milgrom, Philip Reny, Ariel Rubinstein, and Hyun Shin.

1. A number of arguments have been suggested in support of this claim: the concept of transparency of reason (cf. Bacharach, 1987; Bjerring, 1978), the notion of self-enforcing agreement (cf. Kreps, 1987), the notion of an "authoritative game theory book" consulted by all the players (cf. Binmore, 1990, pp. 60–61).

   There are also exceptions to this claim. Important ones are Aumann (1987a), Bernheim (1984), Brandenburger and Dekel (1987), Pearce (1984), and Tan and Werlang (1984).

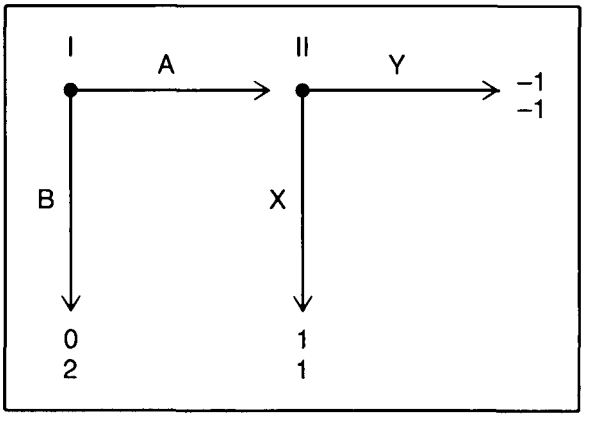


FIGURE 1.

suitable refinement of the Nash equilibrium concept.[2] The central idea of the refinement program is that if a player's strategy is to be part of a rational solution, it must prescribe a rational choice of action *in all circumstances*, including those that are ruled out by the candidate equilibrium. Thus, the central question of the refinement program is what constitutes rational behavior at information sets that are not reached by the equilibrium path.

What kind of reasoning lies behind the attempt to determine "rational" choices of action at information sets that are not reached by the equilibrium path? In discussing the game of Figure 1, Harsanyi and Selten write:

> In modern logic the problem . . . can be restated as follows: The assumption that player II will use strategy Y is equivalent to the conditional statement S, "If player I were to make move A, then player II would make move Y." If this conditional statement is interpreted as *Material Implication*, it will automatically become *vacuously true* whenever the stated condition (player I's making move A) does not arise. But if statement S is interpreted as a *Subjunctive Conditional* . . . it will be simply *false*. If player I does make move A, then player II (assuming that he is a rational individual who tries to maximize his payoff) would most certainly *not* make move Y.

2. The following is a (possibly incomplete) list of published contributions, each proposing a different refinement of the Nash equilibrium concept: Banks and Sobel (1987), Cho (1987), Cho and Kreps (1987), Grossman and Perry (1986), Harsanyi and Selten (1988), Kalai and Samet (1984), Kohlberg and Mertens (1986), Kreps and Wilson (1982a), McLennan (1985), Myerson (1978), Okada (1981), Selten (1965, 1975), and Wu Wen-Tsun and Jiang Jia-He (1962). There are also a number of as yet unpublished articles where more solution concepts are put forward. For an overview of the literature see Van Damme (1987).

> The strategy pair (*B*, *Y*) is formally an equilibrium point. For this to be the case, all we need is that statement S should be true when it is interpreted as a Material Implication. . . . Nevertheless, our game-theoretical intuition judges (*B*, *Y*) to be an *irrational* equilibrium point because this intuition would accept the truth of statement S only if it remained true even when interpreted as a Subjunctive Conditional, which is obviously not the case. (1988, pp. 18–19)

This quotation suggests that: (1) strategies ought to be thought of as hypothetical statements;[3] however, if strategies are treated as instances of material implication, then all one can prove is that *all* the Nash equilibria and *only* the Nash equilibria of a particular game constitute the rational solutions of that game; and (2) in order to obtain sharper results, that is, in order to refine the notion of Nash equilibrium, it is necessary to construe strategies as subjunctive conditionals or, as some authors put it (cf. Bicchieri, 1988a; Selten and Leopold, 1982), as *counterfactuals*.

Unfortunately, an explicit definition of rationality and a clear illustration of the logical reasoning that leads from the notion of rationality to the hypothetical statements that make up the "rational strategies" cannot be found in standard game theory.[4] This is unfortunate, since the very notion of rationality seems to refer to a process of logical deduction from given premises, and, indeed, the desirability of an axiomatic approach has been recognized by some of the leading contributors to the refinement program: "An ideal way to discuss which equilibria are stable . . . would be to proceed axiomatically" (Kohlberg and Mertens, 1986, p. 1005).

The purpose of this article is to attempt a logical analysis of extensive games (in particular, games of perfect information) that does not take Nash equilibrium as a starting point. We start from an explicit and intuitively plausible axiom of individual rationality and define a strategy profile to be a rational solution of a given game if it can be deduced from the axiom of rationality and the description of the game by using the language of propositional logic.

The approach suggested in this article is unconventional in other respects, too. First, of all, we treat strategies as instances of *material implication* rather than as counterfactuals. Second, we use propositional logic rather than epistemic logic, so that we do not model players' knowl-

3. A strategy for a player in an extensive-form game is defined as a function that associates with every information set of that player a choice of action at that information set.
4. Bacharach (1987) is an exception, although he is not concerned with extensive-form games or with counterfactuals. Other exceptions, although at a less formal level, are Bicchieri (1988a) and Cubitt (1989). For more recent contributions, see Kaneko and Nagashima (1990a, 1990b) and Shin (1989). See section on related literature for a more detailed discussion of some of these contributions.

edge. In view of the commonly held opinion about the correct interpretation of strategies (cf. the earlier quotation from Harsanyi and Selten, 1988) and the general agreement that common knowledge of the game and of players' rationality is a necessary ingredient of a meaningful analysis of games (cf., e.g., Aumann, 1987b), one might expect that no interesting results could possibly be obtained within our approach. Instead, we show that all games of perfect information that have the property that no player moves more than once along any given play (we call such games "nonrecursive") have a rational solution, and all the rational solutions give rise to the same play, namely, the play associated with the unique subgame-perfect equilibrium (we consider only generic games where no player is indifferent between any two terminal nodes). *Thus, this article can be seen as a counterexample to the commonly held view that it is necessary to define strategies as counterfactuals in order to eliminate "bad" Nash equilibria.* Furthermore, since we model strategies as material implication, it is perhaps not surprising that *there are (nonrecursive) games where there is a rational solution that is not a Nash equilibrium.*

With the axiom of individual rationality suggested in this article, only nonrecursive games can be shown to have a rational solution. It is conceivable that a stronger axiom (or the addition of further axioms) would suffice to solve all games of perfect information and maybe also games of imperfect information. We do not know if such an axiom exists.

It is worth emphasizing that we are not proposing a new theory of games or even suggesting that our approach is the right one. We are merely showing that, by taking a point of view that is different from the conventional one (strategies as material implication rather than subjunctive conditionals or counterfactuals, and propositional logic rather than epistemic logic), one can go surprisingly far. Whether or not it is possible to go even further remains an open question. If it is *not* possible, then the contribution of this article will have been to show the precise sense in which it is necessary to construe strategies as counterfactuals: not because otherwise all Nash equilibria turn out to be rational solutions, but because otherwise only a small class of games can be solved in a satisfactory way. If it *is* possible, then the contribution of this article will have been to raise the question: what do we gain by thinking of strategies as counterfactuals and by modeling players' knowledge?

The article is organized as follows. The next section contains the definition of rational solution, while in the section following that, we introduce an axiom of individual rationality and show that, when applied to one-person games, the solution concept proposed yields results that are consistent with standard decision theory. The analysis of $n$-person games is contained in another section. The last two sections contain a brief discussion of related papers in the literature and some concluding remarks.

## DEFINITION OF RATIONAL SOLUTION

Given an extensive game $G$, we shall denote by $\Gamma$ a formula (a conjunction of propositions) that gives a complete description of the game-tree. An example will be useful. Consider again the game of Figure 1. Let the following symbols have the following interpretation:

*A:* "player I takes action A";
*X:* "player II takes action X";
*B:* "player I takes action B";
*Y:* "player II takes action Y";
$(\pi_i = t)$: "player $i$'s payoff is $t$" $(i = \text{I, II};\ t \in R)$.

The description of the game of Figure 1, denoted by $\Gamma^1$, is given by the conjunction of the following propositions.[5]

$$(\Gamma_1^1) \quad A \vee B$$

$$(\Gamma_2^1) \quad \neg(A \wedge B)$$

$$(\Gamma_3^1) \quad (X \vee Y) \Leftrightarrow A$$

$$(\Gamma_4^1) \quad \neg(X \wedge Y)$$

$$(\Gamma_5^1) \quad B \Rightarrow ((\pi_{\text{I}} = 0) \wedge (\pi_{\text{II}} = 2))$$

$$(\Gamma_6^1) \quad X \Rightarrow ((\pi_{\text{I}} = 1) \wedge (\pi_{\text{II}} = 1))$$

$$(\Gamma_7^1) \quad Y \Rightarrow ((\pi_{\text{I}} = -1) \wedge (\pi_{\text{II}} = -1))$$

$(\Gamma_1^1)$ and $(\Gamma_2^1)$ say that player I must take one and only one of the two actions $A$ and $B$. $(\Gamma_3^1)$ says that player II must take one of the two actions $X$ and $Y$ if and only if player I takes action $A$. $(\Gamma_4^1)$ says that player II can take only one of those two actions. $(\Gamma_5^1)$–$(\Gamma_7^1)$ describe the payoffs.

5. The symbols "$\neg$," "$\wedge$," "$\vee$," "$\Rightarrow$," and "$\Leftrightarrow$" read: "not," "and," "or," "implies," "if and only if," respectively. The symbol "$\Rightarrow$" denotes material implication; thus $(P \Rightarrow Q)$ is equivalent to $(\neg P \vee Q)$. Throughout this article the superscript of $\Gamma$ will denote the number of the figure that illustrates the game. Thus, for example, $\Gamma^{2a}$ is the formula that describes the game illustrated in Figure 2a.

($\Gamma_1^1$)–($\Gamma_7^1$) imply that the only possible assignments of truth-values to the atomic sentences $A$, $B$, $X$, and $Y$ are the following:

| $A$ | $B$ | $X$ | $Y$ |
|---|---|---|---|
| F | **T** | F | F |
| **T** | F | **T** | F |
| **T** | F | F | **T** |

These three assignments are in one-to-one correspondence with the only three plays of the game.

As explained in the previous section, strategies can be thought of as hypothetical statements, and we shall formalize these statements as instances of material implication. Thus, for example, in the game of Figure 1 a possible strategy for player II is the statement: "If player II finds herself in the situation of having to choose between action $X$ and action $Y$, then she will choose action $Y$." Formally, the strategies of player I in the game of Figure 1 are the formulas:

$$(A \vee B) \Rightarrow A \quad \text{and} \quad (A \vee B) \Rightarrow B$$

while the strategies of player II in the game of Figure 1 are the formulas:

$$(X \vee Y) \Rightarrow X \quad \text{and} \quad (X \vee Y) \Rightarrow Y$$

In general, a strategy for player $i$ will be a formula $S_i$ that is a conjunction of hypothetical statements. A *strategy-profile* is a formula $S$ that is the conjunction $(S_{\text{I}} \wedge S_{\text{II}} \wedge \ldots \wedge S_n)$, where $S_i$ is a strategy of player $i$ ($i = \text{I}, \ldots, n$). For example, a possible strategy-profile for the game of Figure 1 is the formula: $[((A \vee B) \Rightarrow A) \wedge ((X \vee Y) \Rightarrow X)]$.

Let $R_i$ ($i = \text{I}, \text{II}, \ldots, n$) be the proposition "player $i$ is rational." The content of the proposition "player $i$ is rational" will be specified by some axioms or rules of inference that express the notion of individual rationality. In the next section we shall propose a very natural rule of inference.

DEFINITION OF RATIONAL SOLUTION. *The strategy profile* S *is a rational solution of the game described by* $\Gamma$ *if and only if* S *is deducible from the conjunction* $(\Gamma \wedge R_{\text{I}} \wedge R_{\text{II}} \wedge \ldots \wedge R_n)$, *that is, if and only if the following is a theorem* (*in the sense of propositional logic*):[6]

$$\Gamma \wedge R_{\text{I}} \wedge R_{\text{II}} \wedge \ldots \wedge R_n \Rightarrow S.$$

6. See Appendix 1 for an exact definition in terms of the formal language of propositional logic.

In other words, it is a logical consequence of the definition of individual rationality used that, if the description of the game is correct, and it is true that all players are rational, then the propositions $S_{\mathrm{I}}$, $S_{\mathrm{II}}$, . . . , $S_n$ are true.

Every definition must be judged on the basis of the results it produces. In the next section we shall show that the rational solutions of one-person games are consistent with standard decision theory, while in the following section we shall show that for a general class of $n$-person games, there is a well-defined correspondence between rational solutions and subgame-perfect equilibria. However, the above definition is appealing also on intuitive grounds: if all players share a common method of reasoning, have a common understanding of the meaning of the word "rational," and start from the same set of hypotheses (namely, that every player is rational and that the game they are playing is as described by $\Gamma$), then they must all reach the same conclusion. This common conclusion is what we call a rational solution.

## AXIOM OF INDIVIDUAL RATIONALITY

There are two basic properties that any definition of rational solution ought to satisfy: (i) it ought to be clear in what sense it is an expression of rationality; and (ii) when applied to one-person games, it ought to yield results that are consistent with standard decision theory.

It is clear that the solution concept proposed here satisfies property (i): given one or more axioms of individual rationality, a strategy-profile $S$ is defined to be a rational solution if and only if it is a logical implication of those axioms. In this section we shall propose a natural axiom of individual rationality and show that, with this axiom, property (ii) is also satisfied. In a later section we shall turn to the analysis of $n$-person games.

The weakest definition of individual rationality seems to be the following: "given two alternatives, a player will always choose the one he prefers, i.e., the one with the larger utility" (Luce and Raiffa, 1957, p. 55). This definition can be translated into the following sentence:

> ⟨⟨If it is true that player $i$ either takes action $A_1$ or takes action $A_2$ or . . . or takes action $A_m$, and action $A_j$ leads to a payoff of at most $\alpha$ and action $A_k$ leads to a payoff of at least $\beta$, and $\alpha < \beta$, then the following is true: if player $i$ takes action $A_j$ he is irrational (or, equivalently, if he is rational he will not take action $A_j$).⟩⟩

Formally, let

$A_{ih}$ = "player $i$ takes action $A_h$" ($h = 1, \ldots, m$),
$(\pi_i \geq (\leq)\, t)$ = "player $i$'s payoff is $\geq (\leq)\, t$" (where $t$ is a real number),
$R_i$ = "player $i$ is rational."

Then we could express this notion either in the form of an axiom scheme or in the form of a rule of inference, as follows. Let $P$ denote the formula

$$(A_{i1} \vee A_{i2} \vee \ldots \vee A_{im}) \wedge (A_{ij} \Rightarrow \pi_i \leq \alpha) \wedge (A_{ik} \Rightarrow \pi_i \geq \beta) \wedge (\alpha < \beta) \quad \textbf{(1)}$$

and $Q$ denote the formula[7]

$$(A_{ij} \Rightarrow \neg R_i) \quad \textbf{(2)}$$

*Axiom scheme:*

$$P \Rightarrow Q \quad \textbf{(3)}$$

*Rule of inference:* if

$$\Gamma \wedge R_{\mathrm{I}} \wedge R_{\mathrm{II}} \wedge \ldots \wedge R_n \Rightarrow P \quad \textbf{(4)}$$

is a theorem, then also

$$\Gamma \wedge R_{\mathrm{I}} \wedge R_{\mathrm{II}} \wedge \ldots \wedge R_n \Rightarrow Q \quad \textbf{(5)}$$

is a theorem.[8]

Unfortunately, both the axiom scheme and the rule of inference given above are inconsistent.[9] We say that an axiom scheme or rule of inference is *inconsistent* if for every game that has a rational solution, it is a logical consequence of the axiom scheme or rule of inference that the proposition $(\Gamma \wedge R_{\mathrm{I}} \wedge R_{\mathrm{II}} \wedge \ldots \wedge R_n)$ is false (and, hence, by the "paradox of material implication," *every* strategy combination is a rational solution!). In other words, by adding the axiom scheme or rule of inference to the system of propositional logic, the proposition $\neg(\Gamma \wedge R_{\mathrm{I}} \wedge R_{\mathrm{II}} \wedge \ldots \wedge R_n)$ becomes a theorem. The inconsistency is due to the fact that circular arguments are possible: the axiom scheme (or the rule of inference) allows one to deduce **(1)** from a formula containing the proposition $R_i$ and then conclude that $\neg R_i$. While a general proof that the above axiom scheme and rule of inference are inconsistent is given in Appendix 2, we shall give an illustration here with reference to the

7. Note that **(2)** is equivalent to $(R_i \Rightarrow \neg A_{ij})$.
8. As Chellas (1984, p. 15) notes, "a rule of inference is properly understood as meaning that its conclusion is a theorem if each of its hypotheses is."
9. I am grateful to Aanund Hylland and Philip Reny for clarifying my thoughts on this.

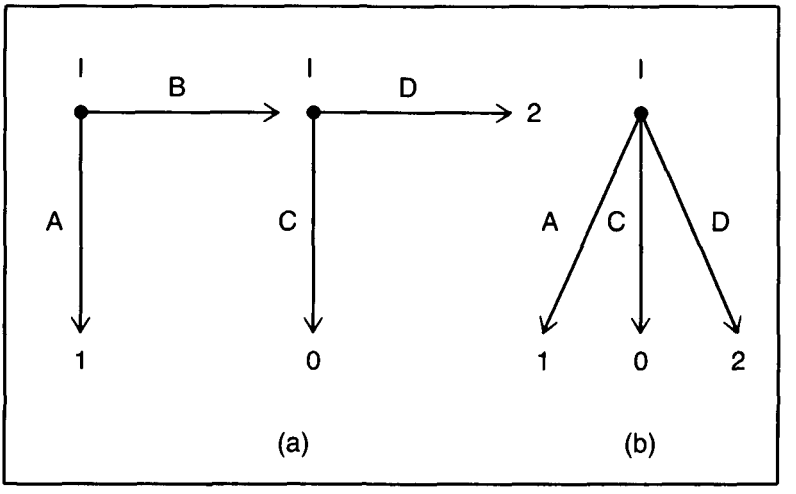


FIGURE 2.

one-person game (decision problem) illustrated in Figure 2b. The logical description, $\Gamma^{2b}$, of this game is as follows:

$$(A \vee C \vee D) \wedge \neg(A \wedge C) \wedge \neg(A \wedge D) \wedge \neg(C \wedge D)$$
$$\wedge (A \Rightarrow \pi_I = 1) \wedge (C \Rightarrow \pi_I = 0) \wedge (D \Rightarrow \pi_I = 2) \qquad (6)$$

Since the following is a tautology:

$$\Gamma^{2b} \Rightarrow (A \vee C \vee D) \wedge (A \Rightarrow \pi_I = 1) \wedge (C \Rightarrow \pi_I = 0)$$
$$\wedge (D \Rightarrow \pi_I = 2) \wedge (0 < 2) \wedge (1 < 2),$$

applying the above axiom scheme or rule of inference we obtain that the following are theorems:

$$\Gamma^{2b} \Rightarrow (A \Rightarrow \neg R_I) \quad \text{and} \quad \Gamma^{2b} \Rightarrow (C \Rightarrow \neg R_I)$$

which are equivalent to:

$$\Gamma^{2b} \wedge R_I \Rightarrow \neg A \quad \text{and} \quad \Gamma^{2b} \wedge R_I \Rightarrow \neg C.$$

Thus, since both $[\Gamma^{2b} \wedge R_I \Rightarrow \Gamma^{2b}]$ and $[\Gamma^{2b} \Rightarrow (A \vee C \vee D)]$ are tautologies, we have:

$$\Gamma^{2b} \wedge R_I \Rightarrow D.$$

Since the following formula (known as the "paradox of material implication") is a tautology: $\neg C \Rightarrow (C \Rightarrow \pi_I = 15)$, we can conclude (using the fact that $\Gamma^{2b} \wedge R_I \Rightarrow \neg C$ is a theorem) that the following is a theorem:

$$\Gamma^{2b} \wedge R_I \Rightarrow (A \vee C \vee D) \wedge (D \Rightarrow \pi_I = 2) \wedge (C \Rightarrow \pi_I = 15) \wedge (2 < 15)$$

and applying the axiom scheme or rule of inference we obtain: $[\Gamma^{2b} \wedge R_I \Rightarrow (D \Rightarrow \neg R_I)]$, which is equivalent to

$$\Gamma^{2b} \wedge R_I \Rightarrow \neg D.$$

But the conjunction of $(\Gamma^{2b} \wedge R_I \Rightarrow D)$ and $(\Gamma^{2b} \wedge R_I \Rightarrow \neg D)$ is equivalent to $\neg(\Gamma^{2b} \wedge R_I)$.

We shall therefore use a weaker version of this axiom, which rules out circular arguments.

DEFINITION OF PLAYER-*i*-ADMISSIBLE HYPOTHESIS. *A formula of the form*

$$\Gamma \wedge (A_{i1} \vee A_{i2} \vee \ldots \vee A_{im}), \quad \text{or} \tag{7a}$$

$$\Gamma \wedge R_k \wedge (A_{i1} \vee A_{i2} \vee \ldots \vee A_{im}) \quad (\textit{for some or all } k \in \{I, \ldots, n\} \setminus \{i\}) \tag{7b}$$

*where* $\Gamma$ *is the description of the game-tree,* $R_k$ *is the proposition "player* k *is rational" (with* $k \neq i$*), and* $A_{ih}$ *has the usual meaning ("player* i *takes action* $A_h$*,"* $h = 1, \ldots, m$; $m \geq 1$).

Thus, a player-*i*-admissible hypothesis, $H_i$, is a proposition of the form: the game-tree is as described by $\Gamma$, player $i$ has to choose among actions $A_{i1}, \ldots, A_{im}$ and (possibly) the *other* players are rational. Since we want to be able to say what choices would be *irrational* for $i$ in the situation expressed by the hypothesis $H_i$, we should not include in this hypothesis the proposition "player $i$ is *rational*."

*Rule of inference of individual rationality* (NERD):[10] If

$$H_i \Rightarrow [(A_{i1} \vee A_{i2} \vee \ldots \vee A_{im}) \wedge (A_{ij} \Rightarrow \pi_i \leq \alpha) \wedge (A_{ik} \Rightarrow \pi_i \geq \beta) \wedge (\alpha < \beta)] \tag{8}$$

is a theorem, then the following is a theorem

$$H_i \Rightarrow [A_{ij} \Rightarrow \neg R_i] \tag{9}$$

10. NERD stands for Necessary Element of a Rationality Definition: we believe that the above rule of inference is very weak, and it is hard to think of a definition of individual rationality that would not contain or imply it.

provided that

1. $H_i$ is a player-$i$-admissible hypothesis, and
2. there is a proof of **(8)** such that each formula in the proof does not contain the atomic sentence $R_i$ (for a definition of proof, see Appendix 1).

We shall now investigate the NERD-rational solutions of one-person games. It may seem that in one-person games a direct application of NERD ought to automatically select the payoff-maximizing action. This is not so, since even in a one-person game choices may have to be made sequentially, and the proviso of NERD does not allow sequential application of the axiom to the same player. Consider the following example, illustrated in Figure 2a.

The logical description of the game, denoted by $\Gamma^{2a}$, is the following formula:

$$(A \vee B) \wedge \neg(A \wedge B) \wedge (B \Longleftrightarrow C \vee D) \wedge \neg(C \wedge D)$$

$$\wedge (A \Rightarrow \pi_1 = 1) \wedge (C \Rightarrow \pi_1 = 0) \wedge (D \Rightarrow \pi_1 = 2) \qquad \textbf{(10)}$$

Any reasonable solution concept ought to select $(B, D)$ as the unique solution. This is indeed the case with our solution concept. It may be instructive, however, to show first that axiom NERD does not allow a proof based on backward induction. In fact, backward induction is represented by the following argument: (T) next to a formula signifies that the formula is a tautology, (NERD) means that the corresponding formula is obtained from the preceding one by applying the rule of inference of individual rationality, and, finally, (IM) means that the corresponding formula is implied by the preceding one:

$$\Gamma^{2a} \wedge (C \vee D) \Rightarrow [(C \vee D) \wedge (C \Rightarrow \pi_1 = 0) \wedge (D \Rightarrow \pi_1 = 2) \wedge (0 < 2)] \qquad \text{(T)}$$

$$\Gamma^{2a} \wedge (C \vee D) \Rightarrow (R_1 \Rightarrow \neg C) \qquad \text{(NERD)}$$

$$\Gamma^{2a} \wedge R_1 \Rightarrow (B \Rightarrow \pi_1 = 2) \qquad \text{(IM)}$$

$$\Gamma^{2a} \wedge R_1 \Rightarrow [(A \vee B) \wedge (A \Rightarrow \pi_1 = 1) \wedge (B \Rightarrow \pi_1 = 2) \wedge (1 < 2)] \qquad \text{(IM)}$$

$$\Gamma^{2a} \wedge R_1 \Rightarrow (R_1 \Rightarrow \neg A) \qquad (*)$$

$$\Gamma^{2a} \wedge R_1 \Rightarrow B \wedge D. \qquad \text{(IM)}$$

This argument, however, is not valid, since to obtain formula (*), one would need to invoke NERD, but the proviso of NERD does not allow this.

We now show that the conclusion ($\Gamma^{2a} \wedge R_I \Rightarrow B \wedge D$) can be reached *without* recurring to backward induction and, therefore, without violating the proviso of NERD. It is easy to check that the following is a tautology:

$$\Gamma^{2a} \Rightarrow \Gamma^{2b} \tag{11}$$

where $\Gamma^{2b}$ is given by **(6)** and $\Gamma^{2a}$ by **(10)**.[11] Furthermore, a direct application of NERD to $\Gamma^{2b}$ yields the following theorem:

$$\Gamma^{2b} \wedge R_I \Rightarrow D \tag{12}$$

Thus, by **(11)** and **(12)** we obtain that ($\Gamma^{2a} \wedge R_I \Rightarrow D$) is a theorem. Finally, since the following is a tautology: $\Gamma^{2a} \Rightarrow (D \Rightarrow B)$, we can conclude that ($\Gamma^{2a} \wedge R_I \Rightarrow B \wedge D$) is a theorem.

In fact, the following general result can be proved:[12]

*Proposition 1:* Consider a finite, one-person game of perfect information with a unique maximum payoff (and without chance moves). Let $p$ be the unique play that leads to the maximum payoff and $S$ be a strategy that gives rise to $p$. Then $S$ is a rational solution of the game. Conversely, if $S$ is a rational solution of the game, then $S$ gives rise to $p$.

We can now turn to the analysis of games with more than one player.

## APPLICATION TO *N*-PERSON GAMES

We shall first highlight some properties of our solution concept by applying it to a few examples and then prove some general results.

*Claim 1:* In the game of Figure 1, $(A, X)$ is the unique rational solution (thus the Nash equilibrium $(B, Y)$ is not a rational solution).

*Proof:* The logical description of the game-tree of Figure 1, denoted by $\Gamma^1$, was given in a previous section. The proof is as follows:

$$\Gamma^1 \wedge (X \vee Y) \Rightarrow (X \vee Y) \wedge (X \Rightarrow \pi_{II} = 1) \wedge (Y \Rightarrow \pi_{II} = -1) \wedge (-1 < 1) \tag{T}$$

11. In fact, $[(A \vee B) \wedge (B \Leftrightarrow C \vee D)]$ implies $(A \vee C \vee D)$; $\neg(A \wedge B)$ is equivalent to $(B \Rightarrow \neg A)$ and, since $(C \Rightarrow C \vee D)$ is a tautology, it follows that $(C \Rightarrow \neg A)$, which is equivalent to $\neg(A \wedge C)$; and so forth.
12. The proof is similar to the one given for the game of Figure 2a and consists in showing that the description of any one-person extensive game of perfect information with sequential moves implies a game without sequential moves (to which NERD can be applied without violating the proviso).

$$\Gamma^1 \wedge (X \vee Y) \Rightarrow (Y \Rightarrow \neg R_{II}) \quad \text{(NERD)}$$

$$\Gamma^1 \wedge R_{II} \Rightarrow (A \Rightarrow X) \quad \text{(IM)}$$

$$\Gamma^1 \wedge R_{II} \Rightarrow (A \vee B) \wedge (A \Rightarrow \pi_I = 1) \wedge (B \Rightarrow \pi_I = 0) \wedge (0 < 1) \quad \text{(IM)}$$

$$\Gamma^1 \wedge R_{II} \Rightarrow (B \Rightarrow \neg R_I) \quad \text{(NERD)}$$

$$\Gamma^1 \wedge R_I \wedge R_{II} \Rightarrow A \wedge X \quad \text{(IM)}$$

$$\Gamma^1 \wedge R_I \wedge R_{II} \Rightarrow [(A \vee B) \Rightarrow A] \wedge [(X \vee Y) \Rightarrow X]. \quad \text{(IM)}$$

This shows that $(A, X)$ is a rational solution. Uniqueness is a consequence of proposition 2. ∎

*Thus, this example shows that, despite the fact that we defined strategies as instances of material implication, not all Nash equilibria are rational solutions.*

Consider now the game illustrated in Figure 3. Apart from the payoffs, the logical description of this game is the same as that of the game of Figure 1.

*Claim 2:* The rational solutions of the game of Figure 3 are $(B, X)$, which is a Nash equilibrium, and $(B, Y)$, which is not a Nash equilibrium.

*Proof:*

$$\Gamma^3 \wedge (X \vee Y) \Rightarrow (X \vee Y) \wedge (X \Rightarrow \pi_{II} = 1) \wedge (Y \Rightarrow \pi_{II} = -1) \wedge (-1 < 1) \quad \text{(T)}$$

$$\Gamma^3 \wedge (X \vee Y) \Rightarrow (Y \Rightarrow \neg R_{II}) \quad \text{(NERD)}$$

$$\Gamma^3 \wedge R_{II} \Rightarrow (A \Rightarrow X) \quad \text{(IM)}$$

$$\Gamma^3 \wedge R_{II} \Rightarrow (A \vee B) \wedge (A \Rightarrow \pi_I = -1) \wedge (B \Rightarrow \pi_I = 0) \wedge (-1 < 0) \quad \text{(IM)}$$

$$\Gamma^3 \wedge R_{II} \Rightarrow (A \Rightarrow \neg R_I) \quad \text{(NERD)}$$

$$\Gamma^3 \wedge R_I \wedge R_{II} \Rightarrow B \quad \text{(IM)}$$

$$\Gamma^3 \Rightarrow [B \Rightarrow \neg(X \vee Y)] \quad \text{(T)}$$

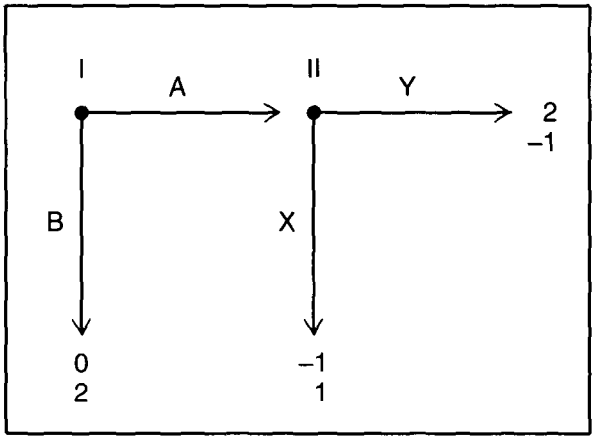


FIGURE 3.

$$\neg(X \vee Y) \Rightarrow [(X \vee Y) \Rightarrow X] \wedge [(X \vee Y) \Rightarrow Y] \qquad \text{(T)}$$

$$\Gamma^3 \wedge R_{\text{I}} \wedge R_{\text{II}} \Rightarrow [(A \vee B) \Rightarrow B] \wedge [(X \vee Y) \Rightarrow X] \qquad \text{(IM)}$$

$$\Gamma^3 \wedge R_{\text{I}} \wedge R_{\text{II}} \Rightarrow [(A \vee B) \Rightarrow B] \wedge [(X \vee Y) \Rightarrow Y]. \qquad \text{(IM)}$$

This shows that $(B, X)$ and $(B, Y)$ are both rational solutions. That they are the only rational solutions follows from proposition 2. ∎

*Thus, this example shows that rational solutions need not be Nash equilibria.* The two rational solutions of this game, however, are equivalent in the sense that they both give rise to the same outcome, namely, the outcome associated with action $B$. It will be shown later (proposition 2) that this is always true: whenever there are multiple solutions, they are all outcome-equivalent. Note also that in the games of both Figure 1 and Figure 3, the rational outcome (or play) identified by the rational solution(s) coincides with the subgame-perfect equilibrium outcome. We shall see later that this is true for a general class of games.

The reason why the non-Nash equilibrium $(B, Y)$ is a rational solution of the game of Figure 3 is that since from $(\Gamma^3 \wedge R_{\text{I}} \wedge R_{\text{II}})$ we can deduce that it is not the case that $(X \vee Y)$, then any hypothetical statement with $(X \vee Y)$ as antecedent, that is, any strategy of player II, is necessarily a true statement. This is, of course, a consequence of the fact that we defined strategies as instances of material implication. One could suggest that, since for the game of Figure 3 the following is a theorem,

$$\Gamma^3 \wedge R_{\text{II}} \Rightarrow [(X \vee Y) \Rightarrow X],$$

this theorem could be used to construct the "counterfactual" statement

$(X \vee Y) \rightarrow X$ (where "$\rightarrow$" denotes counterfactual implication)

that would then be considered to be the unique rational strategy of player II. What kind of reasoning lies behind this suggestion? We know that the following is a theorem: $[\Gamma^3 \wedge R_I \wedge R_{II} \Rightarrow \neg(X \vee Y)]$. But this is equivalent to

$$X \vee Y \Rightarrow \neg(\Gamma^3 \wedge R_I \wedge R_{II}).$$

Thus, we cannot consistently ask the question "suppose the description of the game-tree, $\Gamma^3$, is correct and *both* players are rational and player II has to make a choice between $X$ and $Y$; what choice would be rational for her?" The supposition is inconsistent, and from an inconsistent premise we can deduce anything. The suggestion we are considering, therefore, amounts to the following: if player II's information set happens to be reached, abandon the hypothesis that player I is rational and maintain the other hypotheses, namely, that $\Gamma^3$ and $R_{II}$ are true. There are a number of objections to this suggestion.

First of all, this suggestion cannot be made by those who insist on the assumption that the structure of the game and the rationality of the players be common knowledge. Any reasonable definition of knowledge requires that only true propositions can be known and that, if one knows $P$, and $P$ implies $Q$, then one knows $Q$, too. Thus we would have (where $CK(P)$ stands for "$P$ is common knowledge")

$CK(\Gamma^3 \wedge R_I \wedge R_{II}) \Rightarrow (\Gamma^3 \wedge R_I \wedge R_{II})$ (axiom of knowledge)

$(\Gamma^3 \wedge R_I \wedge R_{II}) \Rightarrow \neg(X \vee Y)$ (consequence of NERD)

$\therefore CK[\neg(X \vee Y)]$ (axiom of knowledge)

Thus, if player II's information set happens to be reached, then she would know that $(X \vee Y)$ and she would also know that $\neg(X \vee Y)$, which contradicts the notion of knowledge.[13]

13. Binmore (1984, p. 55) makes a similar observation: "If players insist that 'perfect rationality' is to be taken as given and then ask what conjecture is reasonable at an information set which is unreachable given this hypothesis, then the only viable conjecture would seem to be that the model they are using is refuted." From this observation Binmore draws the conclusion that the notion of "perfect rationality" is not well defined and proceeds to develop a theory of limited rationality (Binmore, 1987a, 1987b).

However, in this article we are not making the assumption of common knowledge.[14] Thus, we can consistently talk about starting from an initial set of hypotheses and then revising them when the evidence (e.g., the fact that player II's information set is actually reached) shows that the set is inconsistent. The question is, however: why should the revision $(\Gamma^3 \wedge \neg R_I \wedge R_{II})$ be considered more natural than, or preferable to, say, the revision $(\neg\Gamma^3 \wedge R_I \wedge R_{II})$?[15] Consider the following example. Two players are asked to play the following game: player I will enter the room first and can either take the \$100 bill that is on the table, in which case the game ends, or leave the room without taking it. In the latter case, player II can either enter the room or not, and if she does, then she can take the \$100 bill or leave the room without it. Suppose that player II is strongly convinced of player I's rationality, and yet she observes that player I comes out of the room without the \$100 bill. Is it obvious that we ought to label player II as irrational if she decides not to enter the room? An explanation of the following type seems to be entirely rational: "I know that player I would never forgo the opportunity of getting \$100 for free, so I have concluded that the description of the game was not correct. I don't know what the true situation is, but I can conjecture that maybe there wasn't a \$100 bill in the room, or that it was a counterfeit, or. . . ." Some game theorists would probably argue that the description of the game ought to be the last hypothesis to be abandoned. It is hard to see why, and, indeed, some recent contributions (Dekel and Fudenberg, 1987; Fudenberg, Kreps, and Levine, 1988) have investigated the consequences of allowing players to react to unexpected moves by questioning the correctness of the description of the game, in particular the description of the payoffs of the other players.

The last objection to the suggestion that in the game of Figure 3, the theorem $[\Gamma \wedge R_{II} \Rightarrow ((X \vee Y) \Rightarrow X)]$ could be used to define the counterfactual statement that a rational player II would choose $X$ if her

14. In her analysis of strategies as counterfactuals, Bicchieri (1988a) attributes to players not knowledge but beliefs. Bicchieri introduces into game theory the notion, due to Gaerdenfors (1978), of belief revisions that involve a minimum loss of information. Bicchieri's article is very interesting but is not in the spirit of this article, since she assumes (p. 157) that the initial set of beliefs of the players include, among other things, the sentence "the players only consider Nash equilibria." Thus, that rationality implies Nash equilibrium is a datum in her analysis and not a theorem. Furthermore, she does not start from an explicit (axiomatic) definition of individual rationality.

15. Bicchieri's (1988a) suggestion would probably be that the belief revision $(\Gamma^3 \wedge \neg R_I \wedge R_{II})$ involves a smaller loss of information than the revision $(\neg\Gamma^3 \wedge R_I \wedge R_{II})$. This is because she assumes (p. 157) that the initial belief set of the players, which is common knowledge, includes the propositions: "player I chooses to play $B$" and "players always play what they choose." She suggests that abandoning the latter hypothesis, that is, explaining unexpected events as involuntary deviations involves a smaller belief revision than the one required by the hypothesis that the deviation was voluntary. This suggestion is related to Selten's (1975) notion of deviations as mistakes, which will be discussed in the next section.

decision node were reached is that this suggestion does not have a counterpart in recursive situations, as the following examples show.

*Example 1.*[16] We are told that an individual is faced with the decision whether or not to join an expedition to the North Pole and that he is aware that, if he does, he will find himself in the situation of having to choose between eating canned meat or starving (since, we are told, canned meat is the only kind of food that can be taken on an expedition). We are also told that the individual is a vegetarian and eating meat is the second worst thing that could happen to him (the worst being dying). We are asked to make a prediction on the assumption that he is a rational individual. The prediction is, of course, that he will not join the expedition. We are then told that, contrary to our prediction, he did join the expedition and we are again asked to make a prediction of whether he will eat meat or choose to die. Wouldn't we react by saying: "the evidence you are giving me (his joining the expedition) contradicts the hypotheses you gave me at the beginning. I must now conclude that those hypotheses are wrong: either he is not rational, or he is not a vegetarian, or he didn't really have a choice between joining and not joining the expedition, or it is not true that meat is the only food available, or. . . . If you ask me to retain the hypothesis that the original description of the situation is correct, then I have to conclude that this individual is not rational and it is hard to predict the behavior of irrational people."[17]

*Example 2.* Consider the game illustrated in Figure 4a. In this game, player I gets the largest possible payoff if he plays $D_1$. Hence, it is an immediate consequence of the rule of inference NERD that $(\Gamma^{4a} \wedge R_I \Rightarrow D_1)$ and, therefore, that (by the "paradox of material implication") all the strategy combinations with $D_1$ as first component are rational solutions. Moreover, neither $[\Gamma^{4a} \wedge R_{II} \Rightarrow ((A_2 \vee D_2) \Rightarrow A_2)]$ nor $[\Gamma^{4a} \wedge R_{II} \Rightarrow ((A_2 \vee D_2) \Rightarrow D_2)]$ can be proved using NERD. Furthermore, concerning player I's last decision node we have that, by NERD, the following is a theorem: $[\Gamma^{4a} \wedge R_I \Rightarrow ((A_3 \vee D_3) \Rightarrow D_3)]$. However, since $[\Gamma^{4a} \Rightarrow (D_1 \Rightarrow \neg(A_3 \vee D_3))]$ and $(\Gamma^{4a} \wedge R_I \Rightarrow D_1)$ are both theorems, it follows that also $[\Gamma^{4a} \wedge R_I \Rightarrow ((A_3 \vee D_3) \Rightarrow A_3)]$ is a theorem. Thus, following this suggestion, we would obtain two contradictory counterfactual statements for player I at his last decision node and we would not be able to obtain a counterfactual statement for player II.

We can now turn to some general results that can be proved for our solution concept. These results are very general, in the sense that *they do not require acceptance of the rule of inference NERD,* even though they

16. This example was suggested to me by Jerry Cohen.

17. In this example, most game theorists would probably suggest an explanation in terms of mistakes: the original description of the decision problem *is* correct and the individual *is* rational and he decided *not* to join the expedition, but he mistakenly boarded the ship. We shall return to this type of explanation in the next section.

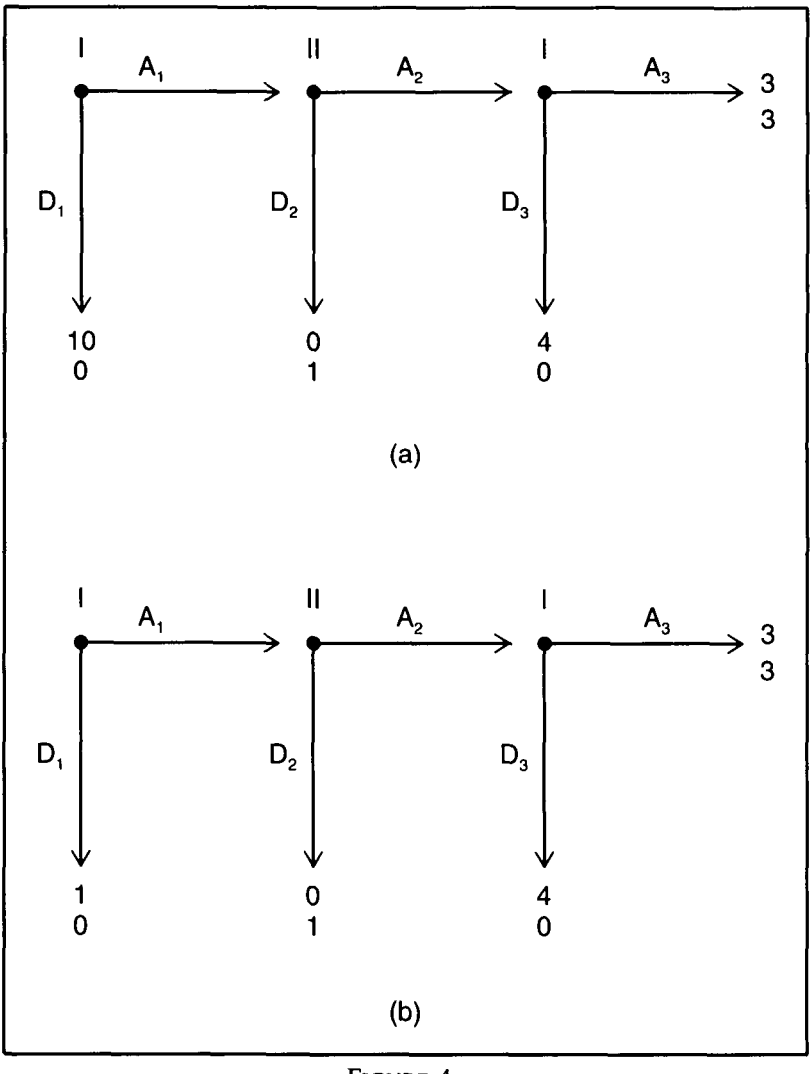


FIGURE 4.

require acceptance of the definition of rational solution and of the definition of strategies as instances of material implication.

*Proposition 2:* If the hypothesis that the description of the game is correct and that all players are rational is consistent,[18] then all the rational solutions of a game are equivalent in the sense that they give rise to the *same play* and, hence, to the *same outcome.*

Thus, proposition 2 says that even though there may be many rational solutions, the prediction in terms of (sequence of actions

18. That is, if it is not the case that the proposition $\neg(\Gamma \wedge R_{\mathrm{I}} \wedge R_{\mathrm{II}} \wedge \ldots \wedge R_{n})$ is a theorem.

and) outcome is unique. The proof is quite simple: if there were two rational solutions that gave rise to two different plays, then there would be an information set of a player, say player $i$, that is reached by both solutions, and two *different* actions of player $i$, call them $A_{ij}$ and $A_{ik}$, one prescribed by one solution and the other prescribed by the other solution. Then – given the definition of rational solution and assuming that $A_{i1}, A_{i2}, \ldots, A_{im}$ are the actions available to player $i$ at that information set – we would have that the following proposition is a theorem

$$\Gamma \wedge R_1 \wedge \ldots \wedge R_n \Rightarrow (A_{i1} \vee \ldots \vee A_{im}) \wedge \neg(A_{i1} \wedge A_{i2})$$

$$\wedge \ldots \wedge \neg(A_{im-1} \wedge A_{im}) \wedge [(A_{i1} \vee \ldots \vee A_{im}) \Rightarrow A_{ij}]$$

$$\wedge [(A_{i1} \vee \ldots \vee A_{im}) \Rightarrow A_{ik}]$$

which implies that $\neg(\Gamma \wedge R_1 \wedge \ldots \wedge R_n)$ is a theorem.

*Proposition 3:* If a game has a unique rational solution (and each player has at least two choices at every information set), then the corresponding play reaches all the information sets.

The proof is clear: if an information set is not reached, then any strategy concerning that information set is a hypothetical statement with a false antecedent, hence, true. Thus, there would be at least two true hypothetical statements (strategies) concerning that information set. It is worth noting that the set of perfect-information games where there is a play that crosses all the information sets is a trivial subset of the set of games of perfect information. Hence, proposition 3 implies that most games do *not* have a *unique* rational solution.

If the notion of individual rationality is expressed by the rule of inference NERD alone, then not every game has a solution. An obvious example is a one-person game with only two actions, both of which give the same payoff. The following proposition identifies those games that do have a NERD-rational solution.

DEFINITION. *A game in extensive form is called* nonrecursive *if along every play of the game each player moves at most once.*

*Proposition 4:* If a game of perfect information (without chance moves) is nonrecursive and has a unique[19] subgame-perfect equilibrium, then the subgame-perfect equilibrium is a NERD-rational solution of the game, but, in general, not the only one. However, if there are many NERD-rational solutions, they all give rise to the subgame-perfect equilibrium play.

The proof of proposition 4 relies on the fact that, in a nonrecursive game, a backward induction argument is allowed by NERD, since it is never the case that the proviso is violated: backward induction can be

19. Uniqueness holds generically.

done in exactly the same way as it was done in the proof of claims 1 and 2. Furthermore, by proposition 3, we know that if the subgame-perfect equilibrium play does not cross all the information sets, there cannot be a unique rational solution. Finally, by proposition 2, we know that all the rational solutions give rise to the same play.

For games that have a recursive structure, a general result similar to that of proposition 4 cannot be proved because of the proviso of NERD. Consider, for example, the game of Figure 4b. The unique subgame-perfect equilibrium is still $(D_1, D_2, D_3)$. The proviso of NERD, however, prevents us from claiming that $(D_1, D_2, D_3)$ is a rational solution. In fact, using NERD we can prove that $[\Gamma^{4b} \wedge R_I \Rightarrow ((A_3 \vee D_3) \Rightarrow D_3)]$, which in turn yields the following theorem: $[\Gamma^{4b} \wedge R_I \wedge R_{II} \Rightarrow ((A_2 \vee D_2) \Rightarrow D_2)]$. At this stage the proviso of NERD does not allow us to proceed any further. *There are good logical reasons why this is so*: Without the proviso we would be using the following circular argument: suppose that player I is rational and his second decision node is reached, then he will choose $D_3$; use this result to conclude that if player I is rational he will play $D_1$. *But this conclusion invalidates the hypothesis "player I is rational and his second decision node is reached" used to obtain the first conclusion.*

Of course, one could "solve" this logical problem by considering the *agent form* of the game, that is, by treating the same player at two different information sets as two different players with the same payoff. By doing so, one can transform every recursive game into a nonrecursive one and then apply proposition 4. However, in doing so, one "solves" the problem by ignoring it.

The discussion so far can be used to raise doubts about commonly accepted views, such as the one according to which the finitely repeated prisoners' dilemma has a unique rational solution where each player acts noncooperatively at each stage. Even though, so far, we have only considered games of perfect information, there is nothing in what we said that limits the applicability of our approach to such games. In particular, our definition of rational solution applies also to games of imperfect information (in Appendix 3 we are given an example of a game of imperfect information that has two pure-strategy Nash equilibria, only one of which is sequential; we show that that game has a unique NERD-rational solution that coincides with the sequential equilibrium). Given the recursive structure of the finitely repeated prisoners' dilemma game, the proviso of NERD makes it impossible to prove that defection at every stage is indeed a rational solution. However, we can imagine that a different axiom (or additional axioms) of individual rationality could do the job. In such a case, proposition 3 tells us that "defection at every stage" cannot be the *unique* rational solution (note that the proof of proposition 3 is equally valid for games of imperfect information). In other words, the prediction would be that rational players would defect

at every stage, but it would not be justified to claim that, once one player has played cooperatively, the unique rational response of his opponent is defection: if there is defection at some stage, then the original set of hypotheses, namely, $(\Gamma \wedge R_1 \wedge \ldots \wedge R_n)$, must be abandoned.

## RELATED LITERATURE

In this article we used the language of propositional logic to analyze extensive games. Bacharach (1987) was the first to introduce logic into the analysis of games. However, he considered only *normal-form, simultaneous* games and used first-order epistemic logic. One of the results he proves is that only Nash equilibria can be solutions. It is worth noting, however, that the proof relies on the fact that a *defining property* of a solution concept – according to Bacharach – is that it select a *unique* strategy profile.

We defined strategies as instances of material implication and observed that this approach is at variance with the prevailing view that strategies ought to be construed as counterfactual statements. Bicchieri (1988a) seems to have been the first to study the issue of counterfactual reasoning in extensive games. Counterfactuals seem to require moving away from common knowledge of players' rationality and attributing to players a commonly known hierarchy of beliefs or a commonly known procedure of belief revision.[20] We pointed out in notes 14 and 15 some important differences between the objectives of this article and those of Bicchieri's (1988a) article.

Foundational issues in game theory have been raised in a number of recent articles.[21] Space limitations prevent us from discussing all these contributions.

Binmore (1984, 1987a) argues as follows: (i) the attempt to determine rational behavior at unreachable information sets involves counterfactual reasoning; (ii) counterfactual reasoning, however, is unavoidable, because it is off-the-equilibrium-path behavior that determines equilibrium behavior; (iii) in order to avoid logical inconsistencies, deviations must be explained as mistakes; (iv) mistakes can take place at the level of players' reasoning, because there is no such thing as "perfect rationality." Binmore (1987a, p. 179) himself summarizes his contribution as follows: "The essential point is that the traditional or axiomatic approach needs to be abandoned in favor of a constructive or algorithmic approach."

20. Recent interesting contributions along these lines are Battigalli (1989), Bicchieri (1988b, 1988c), Pettit and Sugden (1989), Shin (1989), and Sugden (1988).
21. See Basu (1988, 1990), Battigalli (1989), Bicchieri (1988a, 1988b, 1988c), Binmore (1984, 1987a, 1987b), Cubitt (1988, 1989), Pettit and Sugden (1989), Reny (1985, 1988), Samuelson (1989), Shin (1988), Sugden (1988), and Tan and Werlang (1984).

Reny (1988) suggests a definition of common belief of Bayesian rationality at a node of a game of perfect information with two players, and shows that common belief of rationality is possible *at every node* if and only if the subgame-perfect equilibrium play reaches all the decision nodes. Thus, there is a point in common between Reny's conclusion and our result that if a game has a unique rational solution, the resulting play must reach all the information sets.

Some points that are in common as well as some important differences between our approach and Cubitt's (1989) were pointed out by Cubitt himself (pp. 120–21).

One of the implications of our approach is that if a game has a rational solution and the corresponding play does not reach a certain information set, then the hypothesis ($\Gamma \wedge R_{\mathrm{I}} \wedge R_{\mathrm{II}} \wedge \ldots \wedge R_n$) – that is, the hypothesis that the description of the game is correct and that all players are rational – cannot yield any nontrivial predictions as to what the relevant player would do at that information set. Selten (1975) suggests a notion of individual rationality according to which a deviation from a "rational solution" is interpreted as an involuntary – and *ex ante* very unlikely – mistake on the part of one of the players. This notion is the rationale for most of the refinements proposed in the literature (perfect, proper, sequential equilibrium, and so forth). According to this approach, it is never the case that from the description of the game and the hypothesis that all players are rational, one can deduce that a certain information set will not be reached. Furthermore, if *common knowledge of rationality* is assumed, then deviations from the equilibrium are always necessarily interpreted as involuntary mistakes. Rosenthal (1981) and Binmore (1987a) have argued forcefully that this approach yields unacceptable results in games with a recursive structure. Another common objection to this approach is that it is not clear why we should consider it desirable to base a definition of individual rationality on the possibility of mistakes. After all, Selten (1975, p. 25) himself recognizes that "there cannot be any mistakes if the players are absolutely rational." If the purpose of introducing mistakes is to eliminate "irrational" Nash equilibria, such as Nash equilibria that involve incredible threats, then the approach suggested in this article provides an alternative: it eliminates incredible threats without resorting to mistakes (cf. the game of Figure 1, where the only rational solution is $(A, X)$).

## CONCLUSION

The literature of the past 10 years shows that considerable effort has been spent in the attempt to find a suitable refinement of the Nash equilibrium concept that would provide the "rational solution" of any extensive-form game. The basic tenets of this approach are: (i) a "rational solution" must be a Nash equilibrium; and (ii) the Nash equilibrium

concept needs to be refined to deal with the problem of "rational" behavior at information sets that are not reached by the equilibrium path. Many equilibrium concepts have been suggested within this approach, but no one has achieved the goal of the research program.

In this article we suggested a different approach based on an explicit definition of rationality. The need to start from an explicit definition is based not only on the need to avoid ambiguity, but also on the fact that it may be misleading to try to express the notion of rationality by means of an equilibrium concept. In fact, the concept of rationality seems to refer to a process of logical deduction, while the concept of equilibrium (or at least of stable equilibrium) refers to a deviation-correcting process.

As emphasized earlier, the purpose of this article was not to propose a new theory of games or to suggest that our approach is the correct one. We merely showed that by taking a point of view that is different from the conventional one (strategies as material implication rather than subjunctive conditionals or counterfactuals, and propositional logic rather than epistemic logic) one can go surprisingly far. In fact, we were able to show that the solution concept put forward in this article is consistent with decision theory, yields solutions that are equivalent (in the sense that they all give rise to the same outcome), and gives logical foundation to the selection of outcome implied by the notion of subgame-perfect equilibrium in nonrecursive games (where no player moves more than once along any given play). It is conceivable that with a stronger axiom of individual rationality (or with the addition of further axioms) one could go even further. If such a stronger axiom *does not exist,* then the contribution of this article will have been to show the precise sense in which it is necessary to construe strategies as counterfactuals: not because otherwise all Nash equilibria turn out to be rational solutions, but because otherwise only a small class of games can be solved in a satisfactory way. If such a stronger axiom *does* exist, then the contribution of this article will have been to raise the question: what do we gain by thinking of strategies as counterfactuals?

**REFERENCES**

Aumann, R. 1987a. "Correlated Equilibrium as an Expression of Bayesian Rationality." *Econometrica* 55:1–18.

———. 1987b. "Game Theory." In *The New Palgrave: A Dictionary of Economics,* edited by J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

Bacharach, M. 1987. "A Theory of Rational Decision in Games." *Erkenntnis* 27:17–55.

Banks, J., and J. Sobel. 1987. "Equilibrium Selection in Signalling Games." *Econometrica* 55:647–61.

Basu, K. 1988. "Strategic Irrationality in Extensive Games." *Mathematical Social Sciences* 15:247–60.

———. 1990. "On the Non-Existence of a Rationality Definition for Extensive Games." *International Journal of Game Theory* 19:33–44.

Battigalli, P. 1989. "On Rationalizability in Extensive Games." Mimeo. Bocconi University, Milan.

Bernheim, D. 1984. "Rationalizable Strategic Behaviour." *Econometrica* 52:1007–28.

Bicchieri, C. 1988a. "Strategic Behavior and Counterfactuals." *Synthese* 76:135–69.

———. 1988b. "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge." *Erkenntnis* 29:69–85.

———. 1988c. "Common Knowledge and Backward Induction: A Solution to the Paradox." In *Theoretical Aspects of Reasoning about Knowledge*, edited by M. Vardi. Los Altos: Morgan Kaufmann.

Binmore, K. 1984. "Equilibria in Extensive Games." *Economic Journal* 95:51–59.

———. 1987a. "Modeling Rational Players: Part I." *Economics and Philosophy* 3:179–214.

———. 1987b. "Modeling Rational Players: Part II." *Economics and Philosophy* 4:9–55.

———. 1990. *Essays on the Foundations of Game Theory*. Oxford: Basil Blackwell.

Bjerring, A. K. 1978. "The Tracing Procedure." In *Foundations and Applications of Decision Theory*, edited by C. A. Hooker, J. J. Leach, and E. F. McClennen. Dordrecht: Reidel.

Brandenburger, A., and E. Dekel. 1987. "Rationalizability and Correlated Equilibria." *Econometrica* 55:1391–1402.

Chellas, B. F. 1984. *Modal Logic: An Introduction*. Cambridge: Cambridge University Press.

Cho, I. 1987. "A Refinement of Nash Equilibrium." *Econometrica* 55:1867–90.

Cho, I., and D. Kreps. 1987. "Signalling Games and Stable Equilibria." *Quarterly Journal of Economics* 102:179–221.

Cubitt, R. 1988. "Dominance and Rationality in Noncooperative Games." Mimeo. Oxford: The Queen's College.

———. 1989. "Refinements of Nash Equilibrium: A Critique." *Theory and Decision* 26:107–31.

Dekel, E., and D. Fudenberg. 1987. "Rational Behavior with Payoff Uncertainty." Mimeo. Berkeley: University of California.

Fudenberg, D., D. Kreps, and D. Levine. 1988. "On the Robustness of Equilibrium Refinements." *Journal of Economic Theory* 44:354–80.

Gaerdenfors, P. 1978. "Conditionals and Changes of Belief." *Acta Philosophica Fennica* XXX:381–404.

Grossman, S., and M. Perry. 1986. "Perfect Sequential Equilibrium." *Journal of Economic Theory* 39:97–119.

Harsanyi, J. C., and R. Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Kalai, E., and D. Samet. 1984. "Persistent Equilibria." *International Journal of Game Theory* 13:129–44.

Kaneko, M., and T. Nagashima. 1990a. "Game Logic I: Deductions and the Common Knowledge of Deductive Abilities." Working Paper No. E90-03-1. Virginia Polytechnic Institute and State University.

———. 1990b. "Final Decisions: The Nash Equilibrium Concept and Solvability in Non-Cooperative Games with Common Knowledge of Logical Abilities." Working Paper No. E89-12-01. Virginia Polytechnic Institute and State University.

Kohlberg, E., and J. F. Mertens. 1986. "On the Strategic Stability of Equilibria." *Econometrica* 54:1003–37.

Kreps, D. 1987. "Nash Equilibrium." In *The New Palgrave: A Dictionary of Economics*, edited by J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

Kreps, D., and R. Wilson. 1982a. "Sequential Equilibria." *Econometrica* 50:863–94.

Luce, R. D., and H. Raiffa. 1957. *Games and Decisions*. New York: Wiley.

McLennan, A. 1985. "Justifiable Beliefs in Sequential Equilibrium." *Econometrica* 53:889–904.

Myerson, R. B. 1978. "Refinement of the Nash Equilibrium Concept." *International Journal of Game Theory* 7:73–80.

Okada, A. 1981. "On Stability of Perfect Equilibrium Points." *International Journal of Game Theory* 10:67–73.

Pearce, D. 1984. "Rationalizable Strategic Behaviour and the Problem of Perfection." *Econometrica* 52:1029–50.

Pettit, P., and R. Sugden. 1989. "The Backward Induction Paradox." *The Journal of Philosophy* 86:169–82.

Reny, P. 1985. "Rationality, Common Knowledge, and the Theory of Games." Ph.D. thesis. Princeton University.

———. 1988. "Backward Induction and Common Knowledge in Games of Perfect Information." Mimeo. Department of Economics, University of Western Ontario.

Rosenthal, R. 1981. "Games of Perfect Information, Predatory Pricing, and the Chain-Store Paradox." *Journal of Economic Theory* 25:92–100.

Samuelson, L. 1989. "Dominated Strategies and Common Knowledge." Department of Economics Working Paper. Pennsylvania State University.

Selten, R. 1965. "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetraegheit." *Zeitschrift fuer die gesmate Staatswissenschaft* 12:301–24.

———. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4:25–55.

Selten, R., and U. Leopold. 1982. "Subjunctive Conditionals in Decision and Game Theory." In *Philosophy of Economics*, edited by W. Stegmueller, W. Balzer, and W. Spohn. Berlin: Springer-Verlag.

Shin, H. S. 1988. "Correlated Equilibrium as a Consequence of Evidential Rationality with Lessons for 'No Trade' Equilibria." Mimeo. Oxford: Magdalene College.

———. 1989. "Counterfactuals and a Theory of Equilibrium in Games." Discussion Paper No. 42. Oxford: Nuffield College.

Sugden, R. 1988. "Game Theory without Backward Induction." Mimeo. School of Economic and Social Studies, University of East Anglia, U.K.

Tan, T., and S. Werlang. 1984. "The Bayesian Foundations of Rationalizable Strategic Behavior and Nash Equilibrium Behavior." Mimeo. Princeton University.

Van Damme, E. 1987. *Stability and Perfection of Nash Equilibria*. Berlin: Springer-Verlag.

Wu Wen-Tsuen and Jiang Jia-He. 1962. "Essential Equilibrium Points of *n*-Person Non-Cooperative Games." *Science Sinica* 11:1307–22.

# APPENDIX 1

In this appendix we give a rigorous definition of rational solution. With every $n$-person extensive game of perfect information we can associate:

(1) the alphabet $V$ consisting of: the atomic formulas used to describe the game-tree, the $n$ propositions $R_{\mathrm{I}}, R_{\mathrm{II}}, \ldots, R_n$ (recall that the interpretation of $R_i$ is "player $i$ is rational"), the propositions $(\pi_i = t)$, $(\pi_i > t)$, $(\pi_i < t)$ $(i = \mathrm{I}, \ldots, n; t \in R$; the interpretation is "player $i$'s payoff is equal to/greater than/less than $t$), and the connectives of propositional logic,[1]

(2) the language $\mathfrak{L}$, which is the set of all well-defined formulas obtained from $V$ using the composition rules of propositional logic, together with the ordering of the real numbers.

Thus, $\Gamma$, the description of the game-tree, is an element of $\mathfrak{L}$, and the set of strategies $\Sigma_i$ of player $i$ is a subset of $\mathfrak{L}$.[2]

Let $\mathfrak{L}^*$ be the language $\mathfrak{L}$ together with the axioms of propositional calculus, the rule of inference modus ponens and the rule(s) of inference (or axiom scheme(s)) that express the notion of individual rationality (e.g., the rule of inference NERD).

DEFINITION. *The strategy combination* $(S_{\mathrm{I}} \wedge S_{\mathrm{II}} \wedge \ldots \wedge S_n)$ *is a* rational solution *of the game described by* $\Gamma$ *if and only if the formula*

$$\Gamma \wedge R_{\mathrm{I}} \wedge R_{\mathrm{II}} \wedge \ldots \wedge R_n \Rightarrow S_{\mathrm{I}} \wedge S_{\mathrm{II}} \wedge \ldots \wedge S_n \qquad \textbf{(A.1)}$$

*is a theorem of* $\mathfrak{L}^*$.

As in propositional calculus, we say that formula $\Phi$ is a *theorem* of $\mathfrak{L}^*$ if there exists a finite sequence of formulas $\Phi_1, \Phi_2, \ldots, \Phi_m$, such that: (1) $\Phi_m = \Phi$, (2) each $\Phi_j$ $(j = 1, \ldots, m)$ is either an axiom or is obtained from previous elements in the sequence by means of the rule of inference modus ponens or the rule(s) of inference of individual rationality. We say that the sequence $\Phi_1, \Phi_2, \ldots, \Phi_m$ is a *proof* of $\Phi$. It is clear that $\mathfrak{L}^*$ is an extension of propositional calculus and therefore all the theorems of propositional calculus are theorems of $\mathfrak{L}^*$. Recall that a formula is a theorem of propositional calculus if and only if it is a tautology. The rule of inference of individual rationality thus extends the set of theorems to include formulas that are not tautologies, such as formula **(A.1)**.

1. For example, for the game of Figure 1, $V = \{A, B, X, Y, \pi_{\mathrm{I}} = t, \pi_{\mathrm{II}} = u, R_{\mathrm{I}}, R_{\mathrm{II}}, \wedge, \vee, \neg, \Rightarrow, \Leftrightarrow\}$, where $t$ and $u$ are real numbers.
2. For example, in the game of Figure 1, $\Sigma_{\mathrm{I}} = \{(A \vee B) \Rightarrow A, (A \vee B) \Rightarrow B\}$ and $\Sigma_{\mathrm{II}} = \{(X \vee Y) \Rightarrow X, (X \vee Y) \Rightarrow Y\}$.

# APPENDIX 2

In this appendix we show that the axiom scheme **(3)** and the rule of inference **(4)–(5)** are inconsistent. Consider an extensive-form game of perfect information and let $A_1, A_2, \ldots, A_m$ ($m > 1$) be the actions available to the player (call him player I) who moves at the root of the tree. Then part of the description of the game-tree, $\Gamma$, will be the formula:

$$(A_{I1} \vee A_{I2} \vee \ldots \vee A_{Im}) \wedge -(A_{I1} \wedge A_{I2}) \wedge \neg(A_{I1} \wedge A_{I3}) \wedge \ldots \wedge \neg(A_{Im-1} \wedge A_{Im}) \qquad \textbf{(B.1)}$$

Assume that the game has a rational solution. Then the following formula must be a theorem for some $k = 1, \ldots, m$:

$$\Gamma \wedge R_I \wedge R_{II} \wedge \ldots \wedge R_n \Rightarrow A_{Ik} \qquad \textbf{(B.2)}$$

Let $\beta$ be the largest payoff in the game. Fix an arbitrary $j$ different from $k$. From **(B.1)** we have that $(A_{Ik} \Rightarrow \neg A_{Ij})$ and therefore, since $[\neg A_{Ij} \Rightarrow (A_{Ij} \Rightarrow \pi_I = \beta + 1)]$ is a tautology, it follows from **(B.1)** and **(B.2)** that the following is a theorem:

$$\Gamma \wedge R_I \wedge R_{II} \wedge \ldots \wedge R_n \Rightarrow (A_{I1} \vee \ldots \vee A_{Im}) \wedge (A_{Ik} \Rightarrow \pi_I \leq \beta) \wedge (A_{Ij} \Rightarrow \pi_I = \beta + 1).$$

Hence, using axiom **(3)** or rule of inference **(4)–(5)** we obtain

$$\Gamma \wedge R_I \wedge R_{II} \wedge \ldots \wedge R_n \Rightarrow (R_I \Rightarrow \neg A_{Ik}) \qquad \textbf{(B.3)}$$

which is equivalent to

$$\Gamma \wedge R_I \wedge R_{II} \wedge \ldots \wedge R_n \Rightarrow \neg A_{Ik} \qquad \textbf{(B.4)}$$

Now, the conjunction of **(B.2)** and **(B.4)** is equivalent to

$$\neg(\Gamma \wedge R_I \wedge R_{II} \wedge \ldots \wedge R_n).$$

# APPENDIX 3

In this appendix we show, by means of an example, that the definition of rational solution *can* be applied to games of imperfect information and that axiom NERD may be sufficient to eliminate Nash equilibria that are subgame-perfect but not sequential.

Consider the game of Figure 5. This game has two pure-strategy Nash equilibria: $(R, A)$ and $(M, B)$. Both are (trivially) subgame-perfect, but only $(M, B)$ is sequential (because $A$ is a strictly dominated action for player II at her information set).

The logical description of this game, $\Gamma^5$, is given by the conjunction of the following propositions:

$$L \vee M \vee R, \qquad \neg(L \wedge M), \qquad \neg(L \wedge R), \qquad \neg(M \wedge R),$$

$$L \Rightarrow (A \vee B), \qquad M \Rightarrow (A \vee B), \qquad \neg(A \wedge B),$$

$$R \Rightarrow (\pi_{\mathrm{I}} = 2 \wedge \pi_{\mathrm{II}} = 0),$$

$$L \wedge A \Rightarrow (\pi_{\mathrm{I}} = 0 \wedge \pi_{\mathrm{II}} = -2),$$

$$L \wedge B \Rightarrow (\pi_{\mathrm{I}} = 1 \wedge \pi_{\mathrm{II}} = 1),$$

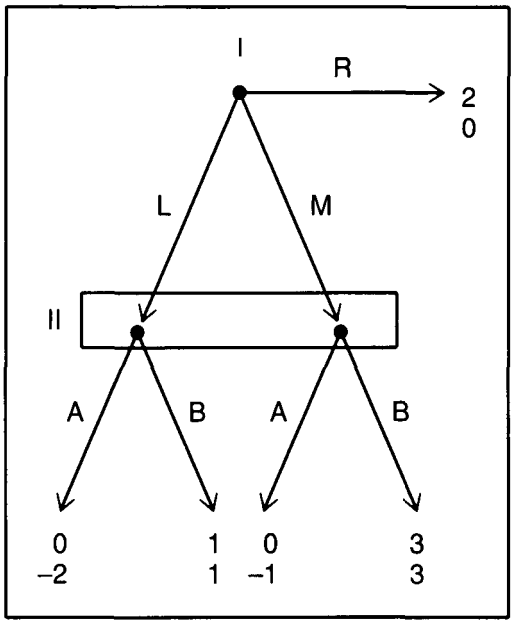


FIGURE 5.

$$M \wedge A \Rightarrow (\pi_{\mathrm{I}} = 0 \wedge \pi_{\mathrm{II}} = -1),$$

$$M \wedge B \Rightarrow (\pi_{\mathrm{I}} = 3 \wedge \pi_{\mathrm{II}} = 3).$$

*Claim:* $(M, B)$ is the unique NERD-rational solution.
*Proof:*

$$\Gamma^5 \wedge (A \vee B) \Rightarrow (A \vee B) \wedge (A \Rightarrow \pi_{\mathrm{II}} \leq -1)$$
$$\wedge (B \Rightarrow \pi_{\mathrm{II}} \geq 1) \wedge (-1 < 1) \qquad \text{(T)}$$
$$\Gamma^5 \wedge (A \vee B) \Rightarrow (R_{\mathrm{II}} \Rightarrow \neg A) \qquad \text{(NERD)}$$
$$\Gamma^5 \wedge R_{\mathrm{II}} \Rightarrow [(A \vee B) \Rightarrow B] \qquad \text{(IM)}$$
$$\Gamma^5 \wedge R_{\mathrm{II}} \Rightarrow (M \Rightarrow B) \wedge (L \Rightarrow B) \qquad \text{(IM)}$$
$$\Gamma^5 \wedge R_{\mathrm{II}} \Rightarrow (L \vee M \vee R) \wedge (R \Rightarrow \pi_{\mathrm{I}} = 2)$$
$$\wedge (L \Rightarrow \pi_{\mathrm{I}} = 1) \wedge (M \Rightarrow \pi_{\mathrm{I}} = 3) \qquad \text{(IM)}$$

Applying NERD twice to the last sentence gives

$$\Gamma^5 \wedge R_{\mathrm{I}} \wedge R_{\mathrm{II}} \Rightarrow \neg R \wedge \neg L \Rightarrow M.$$

This proves that $(M, B)$ is a NERD-rational solution. Uniqueness follows from proposition 2. ∎