Game-Theoretic Axioms for Local Rationality and Bounded Knowledge

by

Cristina Bicchieri and Gian Aldo Antonelli

May 1993

Report CMU-PHIL-37

Philosophy
Methodology
Logic

Pittsburgh, Pennsylvania 15213-3890

Game-Theoretic Axioms for Local Rationality and Bounded Knowledge[1]

Cristina Bicchieri
Carnegie Mellon University

Gian Aldo Antonelli
Yale University

May 1993

[1]**We would like to thank the participants of the Rationality seminar at the Center for Rationality and Interactive Decision Theory at the Hebrew University of Jerusalem, as well as Itzhak Gilboa, Sergiu Hart, Shmuel Zamir and especially Robert Aumann for many useful comments.**

Abstract

We present an axiomatic approach for a class of finite, extensive form games of perfect information that makes use of notions like "rationality at a node" and "knowledge at a node." We show that, in general, a theory that is sufficient to infer an equilibrium must be modular: for each subgame G^1 of a game G the theory of game G must contain just enough information about the subgame G' to infer an equilibrium for G'. This means, in general, that the level of knowledge relative to any subgame of G must not be the same as the level of knowledge relative to the original game G. We show that whenever the theory of the game is the same at each node, a deviation from equilibrium play forces a revision of the theory at later nodes. On the contrary, whenever a theory of the game is modular, a deviation from equilibrium play does not cause any revision of the theory of the game.

1 Introduction

It is generally agreed that rational choice often yields counterintuitive results in finite, extensive form games of perfect information, as well as in finitely iterated noncooperative games involving simultaneous moves. In both cases, a backward induction argument leads to the paradoxical result. Even if they grant that the conclusions of backward induction arguments can be paradoxical, game theorists have widely accepted the formal validity of the conclusions. In this light, a typical solution to the paradox involves contrasting the unbounded rationality usually attributed to the players with a more realistic view of boundedly rational agents. For example, Selten [11] has suggested that "limited rationality" explains why people play tit-for-tat in the finitely repeated prisoner's dilemma, and Kreps et al. [8] explain the same cooperative behavior by introducing some uncertainty about the rationality of the players.

An alternative way of dealing with the paradox is to question the very validity of the backward induction argument itself by questioning the usual premises of such argument. This is the line of argument adopted by Reny [12], Binmore [7], Bicchieri [3], [4], Pettit and Sugden [10], Basu [2] and Bonanno [6]. The present paper belongs to this latter tradition, as it questions and modifies the usual epistemic assumptions made in backward induction arguments. By "epistemic assumptions" we mean the assumptions about what players know about each other and the structure of the game. Such assumptions are often made only implicitly, in that the formal description of the game does not include them. For example, it is usually implicitly assumed that players have common knowledge of the structure of the game and of their being rational. Reny [12] and Bicchieri [3] have argued that under certain conditions common knowledge of rationality leads to inconsistencies. Typically, this has to do with a player's inability to explain another player's deviation from equilibrium, since such a deviation is inconsistent with common knowledge of rationality and of the theory of the game.

In this paper we assign to each game G a theory T_G such that: (i) theory T_G is sufficient to infer an equilibrium; and (ii) theory T_G is not too strong, in the sense that it does not give rise to contradictions as a consequence of containing "too many" levels of knowledge as indicated by Bicchieri [3]. In doing this, it will be crucial that T_G be *modular*: For each subgame G' of G, theory T_G must contain just enough information about G' to infer an

equilibrium for G' (from the point of view of G). This means, in general, that the level of knowledge relative to G' must not be the same as the level of knowledge relative to G. Whenever the level of knowledge at a subgame is the same as the level of knowledge relative to the original game, a deviation from equilibrium play forces a revision of the theory.

Note that the theory T_G is the theory that the players themselves adopt to infer a solution. In this respect, too, we depart from the game-theoretic tradition. Game-theoretic models usually do not specify the players' reasoning, nor do they attribute a "theory" to the players, since in such models it is the outside observer (the game theorist) who reasons to a solution; the players themselves do not. By endowing the players with a theory of the game, our aim is to explicitly model the reasoning that leads them to play the backward induction equilibrium. In our model, a player is like an automatic theorem prover that, from a finite set of axioms and inference rules, can infer an action at the node at which he has to move.

In the present paper we prove that the theory of the game which is sufficient for the players to infer a solution cannot be the *same* theory that the game theorist uses to prove the backward induction result. In our model, the players are only "locally rational" and, at each node, have only just enough knowledge to choose an optimal action at that node. As a corollary, in order to deal with the classic "paradoxes of backward induction" it is no longer necessary to assume that players are less than perfectly rational. It is just enough to assume that the players have, at each node, an amount of knowledge that is not sufficient to infer an optimal choice at that node. In which case the players' behavior is not determinate.

2 The Received View: An example

As an example of what we mean by the "usual epistemic assumption" underlying backward induction arguments, let us consider the game of Figure 1. It is usually assumed that the structure of the game and players' rationality are common knowledge among them.[1] By "rationality" is simply meant that a player, when facing a decision under uncertainty, will select that action

[1] By "common knowledge" of p is meant that everybody knows that p, and everybody knows that everybody knows that p, and so on ad infinitum. For a definition of common knowledge, see Lewis [9] and Aumann [1].

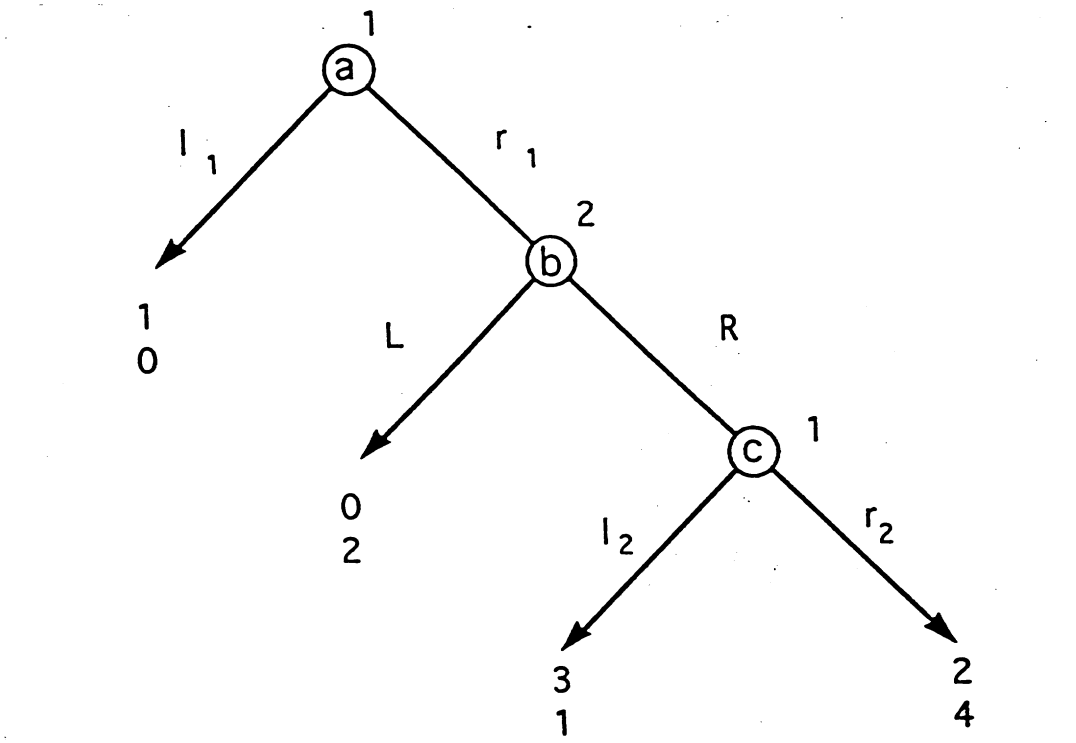

Figure 1:

that maximizes her expected utility with respect to her subjective probability over the uncertain events (in this case, the other player's moves). In the game of Figure 1, the equilibrium $(l_1 l_2 L)$ is obtained by backward induction as follows: Given common knowledge of rationality (CKR), the following proposition must be true

(i) "If node c is reached, player 1 will play l_2."

By CKR, the truth of proposition (i) is common knowledge. Now suppose node b is reached. Player 2 knows that proposition (i) is true, hence she knows that if she plays R, 1 will play l_2. We then have proposition

(ii) "If node b is reached, player 2 will play L".

By CKR, proposition (ii) is common knowledge. Consider now node a. Player 1 knows that proposition (ii) is true, so he knows that if he were to play r_1, player 2 would play L. We then have proposition

(iii) "At node a, player 1 will play l_1".

Note that proposition (iii) does not falsify (i) or (ii). (i) and (ii) are conditional propositions with a false antecedent, therefore they are trivially true. They have a false antecent because, given CKR, the nodes b and c

will never be reached. In deriving proposition (i), we assume that node *c* is reached. And given CKR, node c must be reached by rational play. So (i) is a hypothetical statement that is used as part of a proof that node *c* cannot be reached by rational play. It is then proved (by reductio) that player 1, being rational, will play /]. Given our interpretation of rationality, this means that *li* provides player 1 with greater expected utility than n. But given common knowledge of rationality, this can never be established.

In other words, the answer to the question "What would happen if player 1 were to play *r{r* is indeterminate. For in order to answer this question, we have to predict what player 2 will do at node t, after observing r_{lB} What player 2 will do depends on what she expects 1 to do if she plays *R.* Since we have proved that *R* is an event that will not occur, and this is common knowledge, it must also be common knowledge that both statements "If J?, then player 1 plays r_2" and "If /?, then player 1 plays $/_2$" are trivially true, being material implications with a false antecedent. So at node 6, what player 2 will do remains indeterminate. Similarly at node a; since we proved that *ri* is an event that will not occur, and this proof is common knowledge, it must also be common knowledge that both statements "If n, then player 2 plays L" and "If ri, then player 2 plays *IT* are trivially true. It follows that it can be proved both that player l's equilibrium choice is *r* and that his equilibrium choice is \\.

Notice what the above line of reasoning shows: The standard argument that the Nash equilibrium *(hhL)* will be played is valid. But were player 1 to ask whether he should play $/_{1?}$ given that *1* is part of the equilibrium, he would have to ask himself what would happen were he to play *r* instead, and that question can only be answered if he can predict player 2's reaction to ri. And as we have just said, this is something he cannot do.

We may question whether it makes sense to ask "What would happen if player 1 were to play *r{V* Our claim is that it is a meaningful question if we want to model players' reasoning and the deliberational process that leads them to infer a solution for the game. Backward induction as a reductio proof is a proof given outside the game by an external observer. If we instead want to model how the players themselves reason to an equilibrium, we have to model how they come to decide that a given strategy is optimal for them. In our example, player 1 cannot decide what to choose at node *a* because we have implicitly assumed that, at each node, he reasons according to the game theorist's own theory of the game. Since this theory considers the game as

a whole, a player endowed with it will know the whole game at any node at which she has a choice; this means that at any node the player who has to choose at that node will know what the players choosing at previous nodes know. In the example of Figure 1, at node b player 2 knows that l_1 has been proved to be the optimal choice for player 1 at node a. Because of that, player 2 can infer anything from observing r_1. By giving the players the *same* theory as the game theorist's, it is no longer possible to consider the subtree starting at b as an independent game because what happened before node b is no longer strategically irrelevant.

Note that the very definition of a strategy as a contingent plan of action involves considering what to do at nodes that may never be reached, as is the case with strategy $l_1 l_2$ in the game of Figure 1. To model players' choice of an equilibrium strategy profile must then involve modeling their deliberation at every possible node. In other words, a specification of the solution requires a description of what both agents expect to happen at each node, were it to be reached, even though in equilibrium play no node after the first is ever reached. It is thus important to provide the players with just enough knowledge to decide, at each possible node, what is the optimal choice at that node. Player 1, for example, has two possible choices at his first node: l_1 or r_1. What he chooses depends on what he expects player 2 to do afterwards. If he expects player 2 to play L at the second node, then it is optimal for him to play l_1 at the first node; otherwise he may play r_1. In making a decision, what matters to player 1 is player 2's state of knowledge at node b, which includes what player 2 thinks of player 1's state of knowledge at node c. Note that it is player 1's knowledge of player 2's state of knowledge that determines his choice, whereas what 1 knows that he would know were he to reach node c is completely irrelevant to his decision problem. "What player 1 knows at node c" can therefore be interpreted as what player 1 knows at node a about what player 2 knows at node b about player 1 at node c.

In this paper, we show that the amount of knowledge that is sufficient to infer the backward induction solution is limited: At any node, it is sufficient that the player who has to move at that node knows that the successive player is rational. In our interpretation of rationality, this means knowing that the successive player knows that the next player is rational ... and so on up to the end of the game. It follows that "player i's knowledge at node x" must be interpreted as the *intersection* of what the preceding players know

about player i's state of knowledge at node x.[2]

This definition of knowledge at a node may look strange, as a player would seem to know less and less as further nodes down the tree are reached. And indeed, were the game a truly dynamic one, this definition of knowledge would be inappropriate. But the kind of games we are considering are static; before playing the game, a player has to consider what to do in any possible contingency, and chooses a strategy on the ground of such consideration. As we already mentioned, what matters to a player is what he thinks is going to happen at successive nodes; in this sense, every subgame can be conceived as an independent game. In such a static context, it is then quite intuitive to think that players come to know more and more as they play larger games, as knowledge is defined bottom-up. We now make all this precise.

3 The Theory of the Game

In general, a finite, extensive form game of perfect information G is represented by a finite tree, having an arbitrary branching factor, equipped with a function $p : G \to \{0, \ldots, k)$ that assigns a player i (for $i \leq k$) to each node. The branching factor of the tree is supposed to represent the number of choices available to each player at each node. In order to make things interesting, p is also assumed *not* to be injective, thereby ensuring that at least one player gets to move more than once. Payoffs at the terminal nodes (leaves) of the tree are represented by *real-valued vectors*, whose i-th projections (for $i \leq k$) represent the payoff for player i at that node.

However, there is nothing conceptual to gain in representing such generality, while there is much to lose in notational perspicuity. All the points that we want to make can be made equally well for a restricted class of games. Consequently, we make the following simplifying assumptions. We will restrict ourselves to games represented by *binary* trees, i.e., games in which each player has precisely two choices at each node. Conventionally, these are referred to as "moving left" and "moving right." Morover, we will assume

[2] Strictly speaking, in our model it is not true that a player's knowledge at a node is the intersection of what the preceding players know about his state of knowledge at that node, because we use a belief operator. It is however important to stress that at each node the players are required to move on the basis of an amount of knowledge that varies according to the stage of the game.

only two players that move in turn in a pre-determined order. Accordingly, payoffs at the leaves are represented by *paisr* of real values.

In what follows, we will be employing a notion of *limited* rationality: rather than presupposing that an agent's rationality is an absolute notion, an all-or-nothing affair, we will focus on the idea of player *i* being rational *at a given node*, and *not absolutely*. We are now ready to provide our theory of the game, in the form of axioms *A1-A7* below.

CONVENTION Assume two players, 0 and 1, of whom player 0 is assumed to move first, so that the root of the tree represents a choice for 0. Consequently, *TQ* will be the theory of the game *from the point of view of player* 0. In what follows, *a* always denotes the *root* of G.

DEFINITION Theory *TQ* comprises the rule of *Modus Ponens* (from *ip* and *ip —> %!>* infer T/?), and the following *logical* axiom schemata (for *i* = 0,1), where *Ki* is a *belief* operator:[3]

A	*all instances of tautologies*
A_2	$(K_i\varphi \wedge K_i(\varphi \rightarrow \psi)) \rightarrow K_i\psi$
A_3	$K_i\varphi \rightarrow K_iK_i\varphi$

Beside these axioms, we let *TQ* contain a description of the *structure of payoffs* of the game G. This is a finite conjunction of statements of the form 7r(x) = y, where *x* is a terminal node (leaf) of the tree, and *y* = (21,22) is, say, a *pair* of real valued payoffs, representing the payoff for player 0 and player 1, respectively.

DEFINITION Call a node *final* if it is non-terminal but all of its children are leaves. Let *x* be any non-terminal node; then x_r and *x* denote its right-hand and left-hand child, respectively. It will be useful to characterize a class of functions max; (for *i* = 0,1), having as input two pairs (i.e., vectors of length two) of real-valued payoffs and returning as output a pair of real-valued payoffs. This class is defined as the class of all functions satisfying

[3]**Here *Ki* is construed as the game-theorist's *weak knowledge,* i.e., probability-one belief. The alternative is to employ *strong* knowledge, in the philosopher's sense—that is, at least since Plato, justified *true* belief. This would mean adding an axiom schema to the effect that *Knp* —• <*p.* For instance, this is the approach adopted in Bicchieri [5], but in a context such as the present one, it seems to lead to unnecessary complications.**

the following clauses:

$$\max;(u,i?) \in \{u,t>\},$$
$$\max_t(u, v) = z \Rightarrow (z)i = \max((u)_t\text{-}, (t>);)$$

where $(\bullet)$,- is the projection function: $((u_u \ldots, u_n))\{ = u_{\cdot}$-. The behavior of any of the functions max,- is totally determined when there areno ties in player t's payoffs at a node; when such ties occur, we leave open the possibility of adopting several different choice policies, as embodied in the different functions satisfying the above conditions.

Suppose that x is a non-terminal node. We now "lift" the function *IT* to a function TT*, with domain $\subseteq G$ and values in the real numbers. Function 7r* will be an extension of 7r, but it will *not,* in general, be total.[4] Function 7r* is supposed to represent each player's expected utility at a node, and it will not supply a value unless a player has the "right" amount of knowledge. The behavior of the function is specified by the following axioms:

A\ . **TT*(X) = TT(X),**

for each terminal node x, and

A% $K_{io} \ldots \text{tf}_{f}\text{-}_{n}(\text{Rati7}^{\text{l}\text{''}\,A}\ \text{RatiT''}) \Rightarrow **(*) = \text{niax}_{t}\text{-}(7r*(x_r), 7r*(xO),$

for each non-terminal node x, where: n = *h(x); i* = io", *io* = 0 if and only if *h(x)* is even, and io = 1 otherwise; and $ik+\backslash = 1 — i*$, for each $k < n$. (Some of the alternatives to this axiom are explored below.) Note that the string of leading AVs in the antecedent of *A*\ represents the knowledge of the player who moves first in the game. The reason is straightforward: For the first player to decide what to do, it is not only necessary that the others players behave rationally, it is also necessary that he know that they so behave at every node.

Define Rat^. as a propositional constant representing player ż's rational behavior at x (given our convention, *i* = 0 if and only if x has height *h(x)* = *n* and n is even, and *i* = 1 otherwise; the height h(x) of a node x is defined as the

[4]**The reader who is made unconfortable by the existential import usually associated with the functional notation, might want to introduce a payoff *predicate* constant P^m instead, along with an axiom to the effect that for every x there is at most one *y* such that ^'(*>y)> and regard any context *(^*(x)) as an abbreviation for $Vy\{P^m\{x,y) \Rightarrow \Psi(y)\}$.**

number of links between x and the root). The *definition* or intended meaning of such a constant is given in axiom A_6^x below. Let R_x and L_x be propositional constants representing player i's moving right or left, respectively, at node x (since the player whose turn it is to move is determined by the height of the node, it doesn't need to be explicitly indicated in R_x or L_x). Then we have the axiom:

$$\begin{aligned} A_6^x \qquad & (R_x \iff \neg L_x) \wedge \\ & (\mathrm{Rat}_x^i \iff [(R_x \iff \max_i(\pi^*(x_r), \pi^*(x_l)) = \pi^*(x_r)) \wedge \\ & \qquad\qquad (L_x \iff \max_i(\pi^*(x_r), \pi^*(x_l)) = \pi^*(x_l)]. \end{aligned}$$

In our definition, to be rational at a node x involves knowing that, at the successor nodes x_r and x_l, $\pi^*(x_r)$ and $\pi^*(x_l)$ are defined. In other words, the player who has to move at node x must know that the successive player(s) are rational at nodes x_r and x_l. So, in order to be rational at a node, the player who chooses at that node must know that the successive player is rational and knows that the next player is rational ..., up to the end of the game. A_6^x also says that, whenever there are ties, rationality is relative to a choice policy. When there is a tie, a player can adopt any of several choice rules, but precisely which one is not part of a rigorous definition of rationality. For rational choice to be defined also in the case of ties, one might add a behavioral axiom that singles out one of the possible $\max_i$ functions specifying, for example, that whenever a player is indifferent between n options, he will randomize over them with probability $1/n$. Such behavioral axioms will, however, be ad hoc, and they certainly are not part of the definition of rationality.

Finally, we come to the special axiom specifying precisely to what extent the players' rationality is "common knowledge" among them. First, for each node x we specify a sentence Φ_x. We proceed by induction on (the tree representing) the game. If x is a leaf, we let Φ_x be a propositional constant $\top$ representing "the true" (this is a mere technicality, intendend to take care of "unbalanced" trees); if x is final, then Φ_x is just Rat_x^i, where $i = 0$ if and only if $h(x)$ is even, and $i = 1$ otherwise. If x is a non-final, non-terminal node, then

$$\Phi_x \equiv \mathrm{Rat}_x^i \wedge K_i(\Phi_{x_r} \wedge \Phi_{x_l}),$$

where, again, $i = 0$ if and only if $h(x)$ is even. Then A_7, our last axiom, is Φ_a.

We claim that the theory $T_G = \{A_1, \ldots, A_7\}$ is sufficient to infer the equilibrium and not so "strong" as to give rise to inconsistencies.

THEOREM For each game G, theory T_G is sufficient to infer $E_1 \vee \ldots \vee E_n$, where each E_i is a conjunction of "moves" $M_{i_1} \wedge \ldots \wedge M_{i_m}$ (where each M_{i_j} is of the form L_x or R_x for some node x) representing the branch through G corresponding to an equilibrium.

Proof. It suffices to show that $\pi_0^*(a)$ is defined. We proceed by induction on (the tree representing) G. If G comprises a unique final node x, then it suffices to invoke axiom A_4.

Now consider a game G, with root a, and let b and c be its children. Let G_b and G_c be the subtrees of G with roots b and c, respectively. By inductive hypothesis (modulo a permutation of 0 and 1), theories T_{G_b} and T_{G_c} are sufficient to infer that $\pi_1^*(b)$ and $\pi_1^*(c)$ are defined.

Now, if theories T_{G_b} and T_{G_c} were subtheories of T_G, then the desired conclusion would easily follow from the inductive hypothesis. However, this is not so, given our construal of K_i as a belief operator, and the way our axioms A_5^x and A_7 have been formulated. It is indeed one of the characteristic features of the present approach that if node y is a descendant of node x then Φ_x does *not* imply Φ_y.

There is a way around this difficulty. Theories T_{G_b} and T_{G_c} allow us to derive a value for $\pi^*(b)$ and $\pi^*(c)$ because for each node x in G_b or G_c, they contain the corresponding instance of axiom A_5^x, which has the form

$$\underbrace{K \ldots K}_{n \text{ times}} \varphi \Longrightarrow \pi^*(x) = \cdots,$$

and Φ_b or Φ_c (according as x is in G_b or G_c) provides the antecedent

$$K_0 \ldots K_n \varphi.$$

Now it is easy to verify that for each node x in G_b or G_c, theory T_G contains the axiom

$$\underbrace{K \ldots K}_{n+1 \text{ times}} \varphi \Longrightarrow \pi^*(x) = \cdots$$

(with *one more* occurrence of the K operator with respect to T_{G_b} or T_{G_c}). Correspondingly, Φ_a will now supply the antecedent of the above formula. It

follows that a derivation of a value for $\pi^*(b)$ or $\pi^*(c)$ in T_{G_b} or T_{G_c} can be reproduced in T_G, which therefore will supply a value for $\pi^*(b)$ and $\pi^*(c)$ as well.

All that is left to observe is that T_G contains the following instance of axiom A_5^x:

$$K_0(\mathrm{Rat}_b^1 \wedge \mathrm{Rat}_c^1) \Longrightarrow \pi^*(a) = \max_0(\pi^*(c), \pi^*(b)),$$

whose antecedent, in turn, is supplied by Φ_a. This allows us to derive a value for $\pi^*(a)$. ∎

4 Alternative Accounts of Deviations

Let us take the time to explore two alternatives to our crucial axiom A_5^x. Clearly its intended meaning is that *if* player i_0 has the "right" amount of knowledge, *and* function π^* is defined on the children of x, *then* it is defined on x too. This seems to us conceptually correct: for player i to choose what to do, it is necessary not only that the other players behave rationally, it is also necessary that i *know* that they so behave. Hence, the string of leading K_i's in the antecedent of A_5^x. However, all that is needed in order to infer the equilibrium is the consequent of A_5^x. So it is worth considering what would happen if we were to replace axiom A_5^x by (i) its consequent; or (ii) the result of dropping the leading K_i's from its antecedent (and modify Φ_x by analogously dropping the occurrence of K_i). In both cases, as is clear, we would still be able to infer an equilibrium. But the two cases would differ between themselves, and with the current proposal, in the way *deviations from the equilibrium* can be handled.

In what follows, we will analyze how the theory assigned to each node has to be modified in the face of a deviation from equilibrium. In doing so, it is necessary to distinguish carefully between the player whose turn it is to move at each node, and the game-theorist who observes the game "from the outside." As already mentioned, we want to give an idealized account of the game *from the players' own point of view.* Now it is indeed plausible to assume that the players be capable of *revising* their own beliefs (theories) in the face of inconsistencies, but this requires that the distinction between the level of the theory of the game (on the basis of which each player is making a choice) and the *meta-level* at which belief revision takes place be drawn as sharply as possible.

As we already mentioned, it is useful to resort to the following metaphor: We shall imagine that each player is represented by an *automatic theorem prover* that is supplied some theory of the game as input, and returns as output one of the two possible moves "left" or "right." When faced with inconsistencies, there is nothing a player can do: It is only at the meta-level that we can start talking about belief revision. In principle, a player could well be equipped with a meta-linguistic component, but for clarity it is best to keep the issues distinct for the time being.

Having said this, we can now go back to the alternative formulations of axiom *Af.* Recall that we considered replacing the axiom by: (i) its consequent; or *(ii)* the result of dropping the leading AYs from its antecedent (and modify $\$_x$ by analogously dropping the occurrence of A*). Both cases, as is clear, are sufficient to infer an equilibrium. This precisely means that in either case the theory, when augmented with information to the effect that a deviation has taken place, is simply *inconsistent* But at the meta-level, what kind of belief revisions does this warrant?

Case (i) is simply classical backward induction: axiom *A7* is not needed in this case to infer an equilibrium. This theory does not leave a player much room to maneuver in case a deviation from equilibrium is observed: There is no *natural* way of revising a player's beliefs in order to accomodate a deviation.[5] The only conclusion is that the other player acted against her own best interests for totally mysterious reasons.

Consider again the game in Figure 1, and suppose that player 2 observes rj. Given our modified theory *TQ* (we have now changed axiom ^4f), both players are able recursively to define the value of 7r*(x) at each node x, and in particular both players know that 7r*(a) is defined. By axiom *A%,* if ?r*(a) is defined, then Rat^*_a. Observing n forces player 2 to abandon *A%,* i.e., to abandon the assumption that player 1 is rational at node *a.*

Case (ii) is different: in the presence of an observed deviation from equilibrium, a tentative "explanation" is available for the other player. When player 2 observes a deviation, she is not forced to give up axiom *A%:* she can now revise the theory of the game by assuming that player 1 is not rational, at least at node a, *because 7r*(a)* may not be defined. In turn, ?r*(a) may be

[5]It is certainly possible to modify the theory to account for a deviation by giving up the very definition of rationality at a node *as* given in axiom *A%* or, for that matter, by changing the structure of the payoffs of the game: we do *not* regard these as *natural* belief revisions.

undefined if player 1 does not know (or believe) that player 2 is rational at node b, or if $\pi^*(b)$ is not defined because player 2 does not know (or believe) that Rat_c^1. A deviation in case (ii) is therefore less costly in terms of revisions than a deviation in case (i).

Case (ii) is on a par with the present proposal, since our version of A_5^π leaves open the possibility that a player's rational behavior at a node is not *known* or *believed* by another player, which would serve equally well to "explain" the latter's deviation at a previous node. Case (ii) and our proposal differ, however, in another, important respect. First note that in case (ii), but *not* in our proposal, if node y is a descendant of node x in G, then Φ_x implies Φ_y. It follows that if a deviation at node y is observed, it is not only the theory Φ_y that needs to be revised, but also Φ_x. This is not the case with A_5^π as we defined it. In our theory, deviations from equilibrium play can be dealt with *locally*: they might force a revision of the theories assigned to *later* nodes in the game, but never of theories assigned to *earlier* nodes.

This extra feature, we believe, is of some import, since it gives our theory a certain *modularity*. Although the theory assigned to a subgame G' of G is *not* a subtheory of T_G (given our construal of the belief operator), still it contains enough information to allow the player that moves first in G to infer an equilibrium for G'. Far from being a *negative* feature, this fact allows us to *insulate* a deviation from equilibrium, preventing it from spreading *upwards*. Its consequences are confined to *later* moves in the game, and prior moves are unaffected.

5 Deviations in the Present Account

We now turn our attention to the way deviations from equilibrium can be handled in the framework of our theory. First observe that there is a sense in which theory T_G, beside being sufficient for inferring an equilibrium, is also necessary. Suppose we were to assign the same amount of information T_G not only to a game G, but also to any subgames G' of G. Then, *precisely because T_G is sufficient to infer an equilibrium*, it would make it impossible consistently to explain a *deviation* from equilibrium.

Suppose for instance that we were to assign T_G as the theory of any subgame G' of G, and let G'' be such a game, which represents however a deviation from equilibrium play. As in the above theorem, let $M_1 \wedge \ldots \wedge M_m$

be a. sequence of moves from the root of G to the root of G'' (this ran *he* thought as representing the "previous history" of the game). Then the theory

$$r_C; U\{M_1 A \ldots A M_m\},$$

would be inconsistent. The player who moves first at G'' has no explanation available for such a deviation. Moreover, even from the point of view of the player who moves first at G, the situation is totally unassessable. Indeed, in order to determine that $M\backslash$ A ... A M_m is dominated, player 0 has to infer what would happen, were she to move that way. This means that the value of $\pi^*(6)$ (where $i = 0,1$ as appropriate, and b is the root of $G'')$ has to be determined on the basis of the theory assigned to G'', and this value in unavailable if this theory is TQ U {MI A ... A M_m}.

This leads to the idea of assigning to each node of a tree representing a game, as is already argued in Bicchieri [5] an amount of knowledge that is the *intersection*, and not the *union*, of any amounts of knowledge assigned to nodes higher up in the tree. In other words, the amount of knowledge assigned to each node has to be defined *bottom up* from the leaves, precisely as is accomplished in the recursive definition of $\$_x$ above.[6]

Let us also recall that, as already mentioned, among the advantages of this way of proceeding there is also the possibility of *explicitly* defining *rationality*, a possibility that is not available if rationality is taken to be an absolute notion. In the latter sense, as an all-or-nothing affair, rationality amounts to an agent's always choosing the most profitable live option at each stage of any game. Such generality is simply not expressible in our language, a fact that only comes to the foreground when one sets out to write down the necessary axioms explicitly.

On the contrary, the notion we employ is that of rationality *at a node*. When cast in these terms, local rationality simply amounts to an agent's choosing an action with the highest expected utility, and this is always possible as long as our functions π^M are defined. Conversely, an agent Vs being *not rational* at a node x means that $\pi'(x)$ is not defined. As is often the case with many philosophically intriguing ideas, the notion of rationality has certainly lost some of its metaphysical clout, gaining however in perspicuity and rigor.

[6]Note that, whenever every node is reached in equilibrium, it makes no difference whether theory TQ is assigned to every node or whether we let the theory vary according to the node to which it is assigned.

Similar considerations apply to the players' having *common knowledge* of rationality, a common assumption in game theory. On a local construal of rationality, assuming this kind of common knowledge amounts to saying that the payoff of a given player at a given node is common knowledge. Since in our axiomatization 7r*(x) is not defined unless it is defined at all lower nodes in the tree, common knowledge of rationality means that the value of ?r*(a) is common knowledge among the players (where a is the root of the tree). Equivalently, since such a value is determined by $\$_a$, we can identify common knowledge of rationality with common knowledge of $\$_a$.

As Bicchieri [3], Binmore [7], and Reny [12] have argued, under certain conditions common knowledge of rationality leads to inconsistencies. As already mentioned, this has to do with a player's inability to explain another player's deviation from equilibrium, since such a deviation is inconsistent with common knowledge of rationality and of the theory of the game.

In our framework, inconsistencies arise even when much less than common knowledge is assumed. As we shall presently see, it is sufficient that the theory of the game is group knowledge[7] among the players for that theory to become inconsistent with the statement that a deviation from equilibrium play has occurred. There are at least three possible candidate theories to be assigned to a node *x* of *G*. Before we describe these candidates and assess their merits, let us suppose that, for simplicity, we have a game *G* with root a, whose left- and right-hand children are denoted by *b* and c. As before, player 0 moves at node a. We want to consider some combination of the theories $\$_a$,$\$fe,<l>_c$. Notice that although $\$_a$ is recursively defined in terms of $\$*$, and $\langle\Phi\rangle_G$ it does not entail either one of them. This has to do with our construal of the operators A': as *belief* operators for which the axiom schemata *Knp* —+ *ip* are *not* assumed. Then we could consider assigning a theory of the game to each node *x* of *G* as follows.

Case 1: we assign to each node *x* G *G* the same theory $\$_a$. That is, we make $\langle\Phi\rangle_a$ group knowledge among the players. Suppose that playing R_a is a strictly dominated strategy. Then, as we already know, the theory $\$_a$ A R_n is inconsistent, and therefore of no use for the second player, were she to find herself playing at c. Consequently, the second player has to *revise* her theory of the game in such a way that the resulting theory is still sufficient to infer

[7]By group knowledge of *p* we mean that every member of the group knows *p*.

an equilibrium for the subgame having c as its root. But

$$\$_a = \text{Rat}^\circ_a \wedge A'_o(\$6 \wedge *_c);$$

clearly \$b is of no use for the second player, since it contains information relative to a subgame that is no longer accessible. So the theory that must be revised is $\text{Rat}^\circ_a \wedge Ko\{\$c)$, neither of whose conjuncts is enough to infer an equilibrium. Having rejected $\$_a$, player 1 has *no* theory of the game to speak of; what she does at node c is undefined.

Case 2: we assign to each node $x \in G$ the same theory $\$_a \wedge \$\& \wedge \$_c$. Again, this theory is group knowledge among the players, and as before it is inconsistent with R_a. Finding herself in the position of having to revise her theory, player 1 cannot but reject $\$_a$. However, $\$_c$ still is sufficient to infer an equilibrium, i.e., to compute a value for $ir\backslash(c)$.

Case 3: we assign to each node $x \in G$ the theory $\$_x$. This is the approach sketched above, which calls for assigning to each node x a minimal theory that is sufficient to infer an equilibrium for the corresponding subgame. Thus, each player finds himself choosing at each successive node on the basis of weaker and weaker theories. In our example, this means that player 1 will find herself to choose at node c on the basis of the theory $\$_c$ (or, perhaps, $\$_c \wedge i?_a$). No inconsistency arises, no theory revision is required.

A corollary of our model is that if player i at node x has not enough knowledge to infer an optimal choice, then the backward induction equilibrium cannot be inferred, and the outcome of G remains indeterminate. This consideration suggests that the experimental results that are often at odds with the predictions of backward induction arguments may be formally modelled as due to insufficient knowledge on the part of the players, rather than being the result of the players' being less than fully rational.

References

[1] R.J. Aumann, *Agreeing to disagree,* **Annals of Statistics** 4 (1976), p. 1236-9.

[2] K. Basu, *On the Non-Existence of a Rationality Definition for Extensive Games,* **International Journal of** Game **Theory** 19 (1990), pp. 33-44.

[3] C. Bicchieri, *Self Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge*, **Erkenntnis** 30 (1989), pp. 69–85.

[4] C. Bicchieri, *Knowledge-dependent Games: Backward Induction*, in Bicchieri & Dalla Chiara, **Knowledge, Belief, and Strategic Interaction**, Cambridge University Press, Cambridge 1992.

[5] C. Bicchieri, *Rationality and Coordination*, Cambridge University Press, Cambridge, forthcoming 1993.

[6] G. Bonanno *The Logic of Rational Play in Games of Perfect Information*, **Economics and Philosophy** 7 (1991), p. 37–61.

[7] K. Binmore, *Modeling Rational Players, Part I*, **Economics and Philosophy** 3 (1987), pp. 179–214.

[8] D. Kreps, P. Milgrom, J. Roberts, and R. Wilson, *Rational cooperation in the Finitely Repeated Prisoner's Dilemma*, **Journal of Economic Theory** 27 (1982), pp. 245-52.

[9] D. Lewis, **Convention**, Harvard University Press, Cambridge , MA 1969.

[10] P. Pettit and R. Sugden, *The Backward Induction Paradox*, **Journal of Philosophy** 4 (1989), pp. 1–14.

[11] R. Selten *The Chain-Store Paradox*, **Theory and Decision** 9 (1978), pp. 127–159.

[12] P. Reny, *Rationality, Common Knowledge and the Theory of Games*, unpublished manuscript, Dept. of Economics, University of Western Ontario, London, Ontario, 1988.