Journal of Logic, Language, and Information **4:** 145–167, 1995.

Game-Theoretic Axioms for Local Rationality and Bounded Knowledge*

CRISTINA BICCHIERI[1] and GIAN ALDO ANTONELLI[2]
[1]*Carnegie Mellon University, Department of Philosophy, Pittsburgh, U.S.A.;* [2]*Yale University, Department of Philosophy, New Haven, U.S.A.*

(Received 14 September 1994; in final form 29 May 1995)

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 distinguish between the game theorist's and the players' own "theory of the game." The latter is a theory that is sufficient for each player to infer a certain sequence of moves, whereas the former is intended as a justification of such a sequence of moves. While in general the game theorist's theory of the game is not and need not be axiomatized, the players' theory must be an axiomatic one, since we model players as analogous to automatic theorem provers that play the game by inferring (or computing) a sequence of moves. We provide the players with an axiomatic theory sufficient to infer a solution for the game (in our case, the backwards induction equilibrium), and prove its consistency. We then inquire what happens when the theory of the game is augmented with information that a move outside the inferred solution has occurred. We show that a theory that is sufficient for the players to infer a solution and still remains consistent in the face of deviations must be modular. By this we mean that players have distributed knowledge of it. Finally, we show that whenever the theory of the game is group-knowledge (or common knowledge) among the players (i.e., it is the same at each node), a deviation from the solution gives rise to inconsistencies and therefore 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 induce a revision of the theory.

Key words: Game theory, backwards induction, common knowledge, theory revision

1. Introduction

There are two fundamentally different approaches to modeling cognition within game theory. Understanding the differences between these approaches is of central importance in getting to grips with the problem of backwards induction, and its connection to the consequences of assuming that there is common knowledge of rationality among the players. Our paper makes two important distinctions. The first is the distinction between the game theorist's and the players' own theory of the game. The former provides a justification for a certain solution (e.g., backwards induction), the latter allows the players to infer or compute that solution. The second

* A former version of this paper was presented at the Center for Rationality and Interactive Decision Theory at the Hebrew University of Jerusalem. A subsequent version has been presented at the Nobel Symposium on Game Theory held in Björkborn, Sweden, in June 1993. We would like to thank Martin Dufwenberg, Itzhak Gilboa, Sergiu Hart, Bart Lipman, Dov Samet, Shmuel Zamir and especially Robert Aumann for many useful comments.

distinction is one between the meta-language (i.e., the language of justification) and the object-language (the language of the players).

Meta-theoretic justification is the game theorist's task; computing a solution is the task of the players. While the backwards induction solution has been successfully justified using meta-theoretic arguments (as we show in Section 2), no formal theory of the game from the players' viewpoint has been developed. In this paper we provide the players with a formal (i.e., axiomatic) theory of the game. We want such theory to be the minimal theory sufficient for the players to infer the backwards induction equilibrium. We think of players as theorem provers. If we want them to infer a given solution, we have to provide them with a given axiomatic input. At each node, the relevant axioms will allow the player who chooses at that node to play an optimal move. A player's set of such moves is his equilibrium strategy.

The traditional backwards induction argument is informal, and perfect for the purpose of justification. However, it is not players' argument. The usual implicit premise of the traditional backwards induction argument is that mutual rationality and the structure of the game are common knowledge among the players. It has been argued by Binmore (1987), Reny (1988), and Bicchieri (1989, 1992) that under certain conditions common knowledge of rationality leads to inconsistencies. Their argument is that a player will be unable to explain another player's deviation from the backwards induction equilibrium, since such a deviation is inconsistent with common knowledge of rationality. In this case, it is argued, players become unable to predict future play; as a corollary, what constitutes an optimal choice at a node remains indeterminate. As a consequence of the above criticisms, the usual premises of backwards induction arguments have come to be questioned (Pettit and Sugden (1989), Basu (1990), Bonanno (1991)). However, as Bicchieri (1992) has shown, the problem lies in attributing the traditional backwards induction argument to the players. Part of such argument is an analysis of out-of-equilibrium play, which is typically used to justify a given equilibrium path. When the usual backwards induction argument is formalized in the most straightforward way as the players' own theory, it gives rise to an inconsistency when coupled with information that a deviation has occurred. Thus, the inconsistency is not to be attributed to the players' theory of the game, insofar as this theory is used to infer an equilibrium. Any consistent theory that allows the players to infer an equilibrium becomes inconsistent when combined with a statement to the effect that a deviation from equilibrium takes place. It is only when considering deviations that inconsistencies may arise, but an analysis of deviations need not be part of players' theory of the game. Such theory may thus harmlessly assume that players' rationality is common knowledge: The players would still succeed in computing the backwards induction equilibrium. If we want instead to provide the players with the means to justify a given solution, we have to endow them with an appropriate meta-language within which to express the process of belief revision that out-of-equilibrium play *engenders*. Such language might contain counterfactual conditionals (to talk about

the possibility of deviations), and will be rich enough to express belief revision (which is needed to regain consistency). It is only at the level of belief revision that an assumption of common knowledge of rationality (or of the theory of the game) may be too strong, in that it forces a much more extensive revision of the theory than weaker assumptions.

We choose to make our players rather simple reasoners. We provide them with a theory that is sufficient for them to infer the backwards induction solution directly from the structure of the game alone (we prove the theory to be sufficient in Theorem 2). In order to infer a solution for the game, no counterfactuals are needed. Again, counterfactuals pertain to the realm of justification, and our players just infer (or compute) a solution; we do not ask them to justify it. It must be noted, however, that though we do not ask players to justify a particular solution, we are indeed offering (as game-theorists) a justification for it. A solution to a game is a path through the tree that is consistent with the axioms. Such solution is justified by showing that players can infer it from the axioms they are endowed with.

The players' theory is expressed in a first-order language containing a modal fragment. Our choice is dictated by the fact that we need to express certain facts about beliefs (modally construed) without stepping up to a full-fledged modal predicate logic, which lacks a standard, commonly accepted semantics. Though such hybrid languages are not usually encountered in the literature, they pose no particular conceptual problems.

In Section 3 we first provide a precise definition of the formulas of the language and then, after having identified a class of models, we give a precise definition of the satisfaction relation between a model, a world, and a formula. Having given a model for the theory, we prove the theory's consistency (Theorem 1). In Section 4 we explore what happens if a deviation from the solution occurs. As long as our logic is monotonic, any theory of the game that is sufficient to infer a solution becomes inconsistent when augmented with information that a move outside the solution path takes place. In order to preserve consistency, a revision of the theory is in order. Our players cannot carry out such a revision, since we have not endowed them with a meta-language within which to express belief-revision, nor with a meta-theoretic account of belief revision. Do they need such a meta-language? Not in our model, since the theory of the game is modular. This means that players have just enough knowledge to infer an optimal move at a node or, in other words, that they have "distributed" knowledge of the theory of the game. Distributed knowledge means that – whereas the first player to move has information about all subsequent nodes – the second player has slightly less information. She will have full information about all subsequent nodes, but not about the first node. Similarly for any following player. At every node, the player who chooses at that node has a minimal theory that is just sufficient to infer an optimal move at that node, but does not imply anything about the preceding nodes. If a deviation occurs, it does

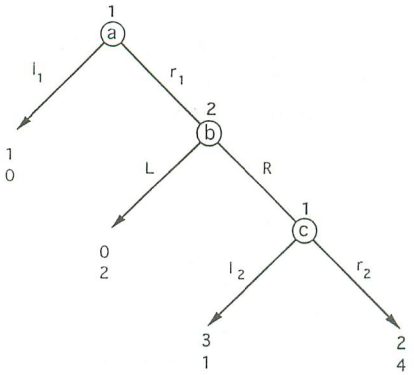

Fig. 1. A simple game.

not force a revision of such a minimal theory. Thus we do not need a model for belief-revision.*

In Section 4, we also examine what happens if we relinquish modularity (distributed knowledge). If the theory of the game is group-knowledge (i.e., each player knows the theory of the whole game), then a deviation makes the theory inconsistent and forces a revision. A fortiori, this also happens if the theory is common knowledge. In this case, we might want to endow the players with a meta-language within which to express belief-revision (but this is stricly outside their theory). Finally, in Section 5 we look at alternative formulations of the players' theory that are still sufficient to infer a backwards induction equilibrium. We assess their merits with respect to how they handle off-equilibrium play, and show that the theory we propose requires the least extensive revision in case it is group-knowledge, and no revision at all when we assume modularity.

2. Justification: An example

As an example of what we mean by a *justification* of backwards induction, let us consider the game of Figure 1.

We assume, as usual, that the structure of the game and players' rationality are common knowledge among them.** By "rationality" is simply meant that a player, when facing a decision under uncertainty, will select that action 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 backwards

* Though in this paper we only give players "local" theories, we are by no means restricted to them. In more recent work, we extend our method to games of imperfect information, and show that it leads quite naturally to the so-called forward induction refinement. The important point is that we give players a mechanical procedure that allows them to infer – from the axioms at their disposal – a solution path consistent with them. See Antonelli and Bicchieri (1994).

** 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 (1969) and Aumann (1976).

induction equilibrium ($l_1 l_2 L$) is justified 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 l_1.

Notice what the above line of reasoning shows: The standard reductio argument that the Nash equilibrium ($l_1 l_2 L$) will be played is valid. Moreover, there is no need to interpret, as it has been repeatedly suggested, the relevant conditionals as counterfactual conditionals.* Any *reductio ad absurdum* proof in mathematics uses material conditionals, and we see no good reason to put subjunctive conditionals in their place.

However, offering a reductio proof amounts to an informal justification of backwards induction. There is a difference between justifying a given solution and actually inferring or computing it: Inferring a solution is the players' task, to be done by means of an appropriate formal (i.e., axiomatic) theory. We may think of players as theorem provers. If we want them to compute a given solution, we have to provide them with a given axiomatic input. At each and every node, the relevant axioms must allow the player who chooses at that node to play an optimal move. In Section 3, we show that all that is needed by the players to infer the backwards induction solution is a minimal theory of the game. Such theory is *modular*, in that for each subgame G' of G, theory T_G must contain just enough information about G' to infer an equilibrium for G'. This means that the level of knowledge relative to G' must not be the same as the level of knowledge relative to G. In other words, players have distributed knowledge of the theory of the game.

* For an explicit discussion and modeling of counterfactuals in game theory, see Bicchieri (1988) and Shin (1992). More recent papers are Aumann (1995) and Stalnaker (1995).

When we formalize the theory of the game in a rigorous way, it becomes evident that the backwards induction solution is compatible with many different levels of players' knowledge of mutual rationality and of the structure of the game. As far as *computing* a solution is concerned, we can indifferently assume, for example, players' common knowledge, group knowledge, or distributed knowledge of rationality.*

A different issue is the following: What happens if a deviation from equilibrium play does occur? An analysis of out-of-equilibrium play is especially important when there are multiple Nash equilibria, some of which might be ruled out if we can show that they are unstable in the face of deviations. A justification of the backwards induction solution might involve showing that the equilibrium thus obtained is stable, as opposed to, say, some other Nash equilibrium that employs weakly dominated strategies. From the viewpoint of a player playing a given equilibrium strategy, an off-equilibrium move is a contrary-to-fact event. An analysis of deviations might thus be cast in terms of counterfactual conditionals.** Such analysis is usually done at a meta-theoretic level, but nothing prevents us from including it in the players' theory of the game. Counterfactuals have been analyzed using the notion of minimally distant or most similar world. If we want to include counterfactuals in the players' own theory in such a way that the solution can be computed, we must do one of the following: (a) Define the similarity relation explicitly and provide a semantic account of counterfactuals; or (b) Give an explicit set of axioms providing a proof-theoretic account of counterfactuals. However, neither (a) nor (b) has been carried out or seems to be forthcoming. Nothing short of this will allow a meta-theoretic justification of backwards induction in terms of counterfactuals to be included in the players' theory of the game.

Note, again, that to compute a solution is a different task than justifying it. Any theory of the game that allows the players to compute the backwards induction equilibrium can dispense with counterfactuals or, for that matter, with any model for belief revision. In Section 3 we present one such theory, and in Section 5 we look at alternative theories that are still sufficient for the players to infer the backwards induction solution. From the viewpoint of computing an equilibrium, all the theories that we discuss are equivalent. They differ, however, in the way they handle deviations. A criterion of choice among them might thus be the extent of the revisions that a deviation induces.

Any theory of the game that employs a monotonic logic and is sufficient for the players to infer a solution becomes inconsistent when augmented with information

* By group knowledge of rationality we mean that each player knows that the other players are rational at every node. By distributed knowledge of rationality we mean that each player knows that the successive players in the game tree are rational, but does not know anything about the preceding players.

** A contrary-to-fact event can be dealt with in several different ways. Instead of using a possible world semantics we could use a syntactical model of belief revision. Alternatively, if the whole theory is expressed in a non-monotonic logic (e.g., default logic), considering contrary-to-fact events does not require the use of belief revision models.

that an off-equilibrium move has been played. In order to preserve consistency, a revision of the theory is in order. When revising the theory of the game, it matters how much the players know, i.e., it matters whether the theory is common knowledge, group knowledge, or distributed knowledge among the players. When the theory of the game is common knowledge, a deviation at any node forces an extensive revision of the whole theory. Such revision will be very costly in terms of lost information: For example, we show in Section 4 that the players have to relinquish the axioms concerning their rationality. When there is much less than common knowledge, a deviation would force a less extensive revision. In Section 4, we consider a case (*Case 2*) in which what has to be revised are the axioms expressing players' beliefs about other players' rationality. Such axioms are less well entrenched than the rationality axioms, so a revision in this case would be less costly. If the theory is distributed knowledge instead, at every node the player who moves there has a *local* theory that allows her to infer an optimal move at that node. Such theory contains no information about the preceding nodes. If a deviation occurs, it has no effect on theories at subsequent nodes. Of course, the statement that a deviation occurs is inconsistent with the theory of the game of the deviating player. This is fairly obvious: Since a player can infer from the theory that a given move(s) is optimal at a node, any alternative moves at that node must be inconsistent with the theory. The important point is that a deviation in such a theory does not force a revision at later nodes. The advantage of the theory of the game we propose is thus that, being modular, it does not require belief revision.

We prove in Section 3 that a theory of the game that is sufficient for the players to infer the backwards induction equilibrium is modular, or its knowledge is distributed among the players. This means that players have just enough knowledge relative to a node to infer an optimal move at that node, but the information relative to a node does not imply anything about previous nodes in the game. In Figure 1, for example, a theory of the game for player 1 at node a must contain information about player 2's move at node b. In turn, player 2's move will depend on her state of knowledge at node b, which includes what player 2 knows of player 1's state of knowledge at node c. In order to choose an optimal move at node a, player 1 has to know what the optimal moves of the subsequent players are. At node c, however, all he needs to know to make a decision are the payoff values at the leaves. Let us now briefly discuss an apparent difficulty with the notion of *local* rationality we employ in our proposal, since this is strictly related to assuming distributed knowedge of the theory of the game.

In Section 3 (Axiom 5), we define rationality to mean that a player maximizes her expected payoff at a node. Furthermore, we argue that, in order to be rational, a player has to know (or believe) that the successive player is rational and knows that the successive player is rational ... and so on up to the end of the game. This requirement has a straightforward explanation. In our model, we do not allow degrees of belief, but just probability-one beliefs (weak knowledge). In order to decide that a given move is optimal at a node, a player has to calculate the

consequences of that move. Unless the node is terminal, the consequences of a given move will depend on what the next player is expected to do in case she is given a chance to choose. Our player has thus to make assumptions about the next player's propensity to choose rationally, as well as her capability of so doing, which may depend on her information about the following player, if there is one. Now, suppose our player has no knowledge about the next player's rationality or information. In this case, she cannot predict the consequences of moving to the next node (at which the other player is choosing). Can she still decide what her optimal move is? If we were to allow degrees of belief, then she could still optimize, given whatever subjective probability she assesses about the next player's rationality and information. However, our model (or, better, the theory we give the players) does not allow for degrees of belief, only knowledge (or probability-one belief). This is consistent with the usual game-theoretic treatment of common knowledge of rationality. If rationality is common knowledge, a player does not attribute a certain probability to an opponent's being rational, he just knows it. Now, if a player is not allowed degrees of belief, in order to be rational (i.e., maximize) at a node she must know whether the next player is rational (or can make a rational choice, for the same reason) at the following node. Otherwise, our player cannot decide what her optimal choice is (i.e., she cannot maximize her expected payoff).

Recall that our goal is to provide the players with the minimum amount of information that is sufficient for them to infer the backwards induction solution. Suppose we wanted to endow them with degrees of belief instead of knowledge. For the players to be able to infer the backwards induction solution, such degrees of belief would have to be pretty specific, i.e., lie within an appropriate range. Take for example the game in Figure 1. At the second node, player 2 will go right for any probability greater than 1/3 that player 1 is not rational (and thus chooses to go right). If we want to give player 2 beliefs that will let her infer that her optimal move is going left, then she must have a degree of belief greater than 2/3 that player 1 is rational (and thus plays left). Could a player be rational with different degrees of belief? Yes, she would obviously still maximize her expected payoff. However, she would no longer play her part in the backwards induction equilibrium.

An axiomatic theory of the game that endows players with degrees of belief would thus have to provide axioms specifying players' probabilistic beliefs, and a formal model for such a theory would be much more complex and cumbersome than the kind of model we provide. A much simpler and manageable axiomatic theory of the game is a theory that employs a (weak or strong) knowledge operator. In this case, in order to maximize one's expected payoff at a node a player has to know what to expect at following nodes, which is tantamount to saying that she has to know that the following players act rationally.

3. The Theory of the Game

A generic finite, extensive form game of perfect information G is represented by a finite tree, having an arbitrary branching factor. For the purposes of this section let us fix a particular tree $(\mathsf{T}, <)$, where $\mathsf{T} = \{n_1, \ldots n_p\}$ is a finite set of partially ordered nodes that satisfy a precedence relation denoted by $<$ *. We will use $n, n', n'', \ldots$ informally as variables ranging over T. Since the tree represents a game, we assume that it is equipped with a function $g : G \rightarrow \{1, \ldots, k\}$ that assigns a player i (for $0 < i \leq k$) to each node. The branching factor of the tree represents the number of choices available to each player at each node. In order to make things interesting, g is also assumed to be *non-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 *vectors* of real numbers whose i-th projections (for $0 < i \leq k$) represent the payoff for player i at that leaf.

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 moves are referred to as "moving left" and "moving right." Morover, we will assume only two players that move in turn in a pre-determined order. Accordingly, payoffs at the leaves are represented by *pairs* 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. The theory will comprise two kinds of axioms: *structural* axioms, describing the game and the payoffs, and *behavioral* axioms that allow the players to infer a move or a sequence of moves.

CONVENTION 1 Assume two players, 1 and 2, of whom player 1 is assumed to move first, so that the root of the tree represents a choice for 1. Call a node *final* if it is non-terminal but all of its children are leaves. Let n be any non-terminal node; then n_{r} and n_{l} denote its right-hand and left-hand child, respectively. Consequently, T_G will be the theory of the game *from the point of view of player* 1.

Before we present the theory of the game, we shall specify a language for the theory and a corresponding class of models. Having done so, we will refer to any formula φ that is true in all models of T_G as a "theorem" or a "consequence" of theory T_G .

---

* The relation $<$ is asymmetric, transitive, and it satisfies the following property: If $n < n''$ and $n' < n''$ and $n \neq n'$, then either $n < n'$ or $n' < n$. The precedence relation is thus only a partial order.

DEFINITION 1 Our language $\mathcal{L}$ will be a *two-sorted first-order* language containg a *modal fragment.* The language comprises two kinds of variables, i.e., $V_1 = b, c, d \ldots$ (which will refer to the nodes of the game tree) and $V_2 = x, y, z, \ldots$ (which will refer to pairs of payoffs, one for each player). Our language will also contain the following: the individual constant a, (which will denote the root of the tree); the propositional constants Rat_n^i, R_n, L_n (for $i = 1, 2$ and n a non-terminal node in the tree T); the function symbols $\max_i(x, y)$ and $\pi(b)$; the predicate symbols $P(b, x)$, $b <^T c$, and $x \leq_i^V y$ for $i = 1, 2$; the connectives $\neg$, $\wedge$, the universal quantifier $\forall$, and the operator K_i $(i = 1, 2)$.

Other connectives and quantifiers are defined from the primitive symbols of the language, so for instance $\varphi \to \psi$ abbreviates $\neg(\varphi \wedge \neg\psi)$ and $\exists x\varphi$ abbreviates $\neg\forall x\neg\varphi$.

DEFINITION 2 We can now give the inductive definition of formulas. To simplify this as well as later definitions, we will use b, c, etc. as meta-variables for *terms* ranging over $V_1 \cup \{a\}$. In this case, we do not need to have separate definitions for formulas containing variables and formulas containing constants.

1. $P(\mathrm{b}, x)$, $\max_i(x, y) = z$, $\pi(\mathrm{b}) = x$, $\mathrm{b} <^T \mathrm{c}$ and $x \leq_i^V y$ are formulas;
2. Rat_n^i, R_n, L_n, $K_j\mathrm{Rat}_n^i$ are formulas;
3. if $K_i\varphi$ is a formula then so is $K_jK_i\varphi$;
4. if φ and ψ are formulas and b, x are variables then $\varphi \wedge \psi$, $\neg\psi$, $\forall x\varphi$, $\forall b\,\varphi$ are formulas;
5. nothing else is a formula.

Observe that we allow the modal operator K_i to be applied only to the propositional constants Rat_n^i.* Next, we specify a class of models for this language.

DEFINITION 3 A *model* for $\mathcal{L}$ is a tuple

$$\mathcal{M} = ((T, <_T), (V, \leq_V^1, \ldots, \leq_V^k), W, R_1, \ldots, R_k, P^*, m_1, \ldots, m_k, p, I),^{**}$$

where:

1. $(T, <_T)$ is a finite tree (representing the game);
2. V is a set of k-tuples of values, and $\leq_V^i$ $(0 < i \leq k)$, where k is the number of players (in our case, $k = 2$), is a reflexive and transitive relation over V (elements of V represent the players' payoff vectors);

* Here K_i is construed as *weak knowledge*, i.e., probability-one belief. The alternative is to employ *strong* knowledge, that is, justified *true* belief. This would mean validating the axiom schema $K_i\varphi \to \varphi$. For instance, this is the approach adopted in Bicchieri (1993), but in the present context it would lead to unnecessary complications.

** While in the object language we denote the partial order relation with the symbol $<^T$, in the model for the language we denote the corresponding relation with the symbol $<_T$. The same holds for the relation $\leq$ as applied to payoff values. Likewise, the symbols P, $\max_i$ and π of the object language are interpreted in the model by P^*, m_i and p, respectively.

3. W is a set of worlds;
4. R_i $(0 < i \leq k)$ are binary accessibility relations over W; the only assumption on R_i is *seriality*: for all $w \in W$ there is a $w' \in W$ such that $R_i(w, w')$;
5. $p : T \rightarrow V$ (p assigns payoff vectors to the terminal nodes of T);
6. P^* is a relation over $T \times V$ (P^* will be used to represent a partial function assigning expected payoff vectors to nodes in T);
7. $m_i : V \times V \rightarrow V$ (m_i represents a generic maximizing function subject to conditions to be specified below);
8. I is a an interpretation function taking as input an individual or propositional constant and a world and returning a node of the tree or a truth value (respectively) as output, satisfying the following conditions:
   (a) for some $m \in T$, $I(a, w) = m$ for every $w \in W$;
   (b) for some $t \in \{\text{true}, \text{false}\}$, $I(L_n, w) = t$ for every $w \in W$;
   (c) for some $t \in \{\text{true}, \text{false}\}$, $I(R_n, w) = t$ for every $w \in W$.
   (d) $I(\text{Rat}_n^i, w) \in \{\text{true}, \text{false}\}$.

It follows from the definition of a model that only the value assigned by I to Rat_n^i depends on the possible world; all the other constants in the language behave as rigid designators. This is due to the fact that in our model the only objects of belief are players' rationality and players' beliefs about other players' beliefs. In particular, possible worlds could be identified with assignments of truth values to the constants Rat_n^i.

DEFINITION 4 If $\mathcal{M}$ is a model for $\mathcal{L}$, an *assignment* (for the variables and the constant a) over $\mathcal{M}$ is a function $s : V_1 \cup V_2 \cup \{a\} \rightarrow T \cup V$ such that $s(a) = I(a)$, and moreover if $b \in V_1$ then $s(b) \in T$, and if $x \in V_2$ then $s(x) \in V$.

DEFINITION 5 If $\mathcal{M}$ is a model, s an assignment for $\mathcal{M}$ and $n \in T$, then s_b^n is the unique assignment such that $s_b^n(\alpha) = s(\alpha)$ if α is not b, and $s_b^n(\alpha) = n$ otherwise. Similarly, if $v \in V$, s_x^v is the unique assignment such that $s_x^v(\alpha) = s(\alpha)$ if α is not x, and $s_x^v(\alpha) = v$ otherwise.

DEFINITION 6 Let $\mathcal{M}$ be a model (with interpretation function I) and s an assignment. We specify when an assignment satisfies a formula in a model at a world, written $\mathcal{M}, w, s \models \psi$. We proceed inductively:

1. $\mathcal{M}, w, s \models \mathsf{b} <^T \mathsf{c}$ if and only if $s(\mathsf{b}) <_T s(\mathsf{c})$;
2. $\mathcal{M}, w, s \models x \leq_i^V y$ if and only if $s(x) \leq_V^i s(y)$;
3. $\mathcal{M}, w, s \models P(\mathsf{b}, x)$ if and only if $P^*(s(\mathsf{b}), s(x))$;
4. $\mathcal{M}, w, s \models \max_i(x, y) = z$ if and only if $m_i(s(x), s(y)) = s(z)$;
5. $\mathcal{M}, w, s \models \pi(\mathsf{b}) = y$ if and only if $p(s(\mathsf{b})) = s(y)$;
6. $\mathcal{M}, w, s \models \text{Rat}_n^i$ if and only if $I(\text{Rat}_n^i, w) = \text{true}$;
7. $\mathcal{M}, w, s \models L_n$ if and only if $I(L_n, w) = \text{true}$;
8. $\mathcal{M}, w, s \models R_n$ if and only if $I(R_n, w) = \text{true}$;
9. $\mathcal{M}, w, s \models K_i\varphi$ if and only if for all w' such that $R_i(w, w')$, $\mathcal{M}, w', s \models \varphi$;

10. $\mathcal{M}, w, s \models \neg\psi$ if and only if $\mathcal{M}, w, s \not\models \psi$;
11. $\mathcal{M}, w, s \models \varphi \wedge \psi$ if and only if $\mathcal{M}, w, s \models \varphi$ and $\mathcal{M}, w, s \models \psi$;
12. $\mathcal{M}, w, s \models \forall b\, \varphi$ if and only if $\mathcal{M}, w, s_b^n \models \varphi$ for every $n \in T$;
13. $\mathcal{M}, w, s \models \forall x\, \varphi$ if and only if $\mathcal{M}, w, s_x^v \models \varphi$ for every $v \in V$.

REMARK Observe that in general the semantic properties we have assumed for the accessibility relation R_i are such that they validate the following axiom schemata:

$$[K_i(\varphi \to \psi) \wedge K_i\varphi] \to K_i\psi$$

(this follows from the satisfaction clause for K_i and implies that $K_i(\varphi \wedge \psi) \to K_i\varphi \wedge K_i\psi$), and

$$K_i\varphi \to \neg K_i\neg\varphi$$

(this is equivalent to the condition of seriality on R_i). Recall that we construe K_i as *belief* and therefore we give up the axiom schema $K_i\varphi \to \varphi$. We require, however, beliefs to be *consistent*, which is precisely what seriality implies.*

CONVENTION 2 We now introduce a particularly important abbreviation. We are going to introduce a "partially defined" function symbol $\pi^*(\mathsf{b})$, representing the expected payoff at node b for the player who moves at b. π^* is not a new primitive symbol of the language: any formula $\Psi(\pi^*(\mathsf{b}))$ in which $\pi^*(\mathsf{b})$ occurs should be regarded as a shorthand for $\forall x(P(\mathsf{b}, x) \Longrightarrow \Psi(x))$. This gives the desired result in the context of our theory because, as we shall see, the theory will contain an axiom to the effect that for every b there is at most one x such that $P(\mathsf{b}, x)$.**

DEFINITION 7 Theory T_G contains the following structural axioms:

$$\max_i(x, y) = x \vee \max_i(x, y) = y \qquad \text{A1.}$$
$$\max_i(x, y) = z \to x \leq_i^V z \wedge y \leq_i^V z \qquad \text{A2.}$$
$$\forall b \forall x \forall y (P(b, x) \wedge P(b, y) \to x = y) \qquad \text{A3.}$$

The first two axioms specify that $\max_i$ is a function that, given two k-tuples of real values, returns the one with the greater i-th projection (if it exists), and chooses arbitrarily otherwise (the axioms do not constrain the behavior of the function on pairs with the same i-th projection). Further, Axiom 3 specifies that P is a functional relation on its domain (but its domain might be strictly smaller than T).

We now give axioms governing the behavior of the players on the basis of the values (among other things) of the propositional constants Rat_n^i. Such constants

* Given our restricted language, such axiom schemata do not apply. But they may be needed within richer languages.

** Recall that the symbol P is interpreted in the model by the relation P^* over $T \times V$ (see Definition 3).

represent player i's rational behavior at node n (given Convention 1, $i = 1$ if and only if n has height m and m is even, and $i = 2$ otherwise; the height of a node n is taken to be the number of links between n and the root). Recall that R_n and L_n are propositional constants representing player i's moving right or left, respectively, at node n (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_n or L_n). Then for each non-terminal node n we have the axiom:

$$Q_a \wedge \ldots \wedge Q_{n'} \rightarrow (R_n \vee L_n), \qquad \text{A4.}$$

where $a \ldots n'$ is the sequence of nodes leading from the root to node n, and each $Q_{n''}$ is $L_{n''}$ or $R_{n''}$ according as the next node in the sequence is the left- or right-hand child of n''. If n is the root, then the antecedent in Axiom 4 becomes empty, and the axiom reduces to $(R_n \vee L_n)$. Axiom 4 says that if a non-terminal node is reached, then the player whose turn it is to move will choose one of the available moves. Of course, we also need to say that such choice is subject to a rationality condition, that is, that the player moves left or right only if by so doing she maximizes her expected payoff (if the two moves carry equal expected payoffs, they are both allowed). The expected payoff at a node is given by the "function" π^*, whose definition we will give shortly. For any non-terminal node n we have the axiom:

$$\begin{aligned} \mathrm{Rat}_n^i \iff & [(R_n \rightarrow \max_i(\pi^*(n_\mathrm{r}), \pi^*(n_\mathrm{l})) = \pi^*(n_\mathrm{r})) \wedge \\ & (L_n \rightarrow \max_i(\pi^*(n_\mathrm{r}), \pi^*(n_\mathrm{l})) = \pi^*(n_\mathrm{l}))]. \end{aligned} \qquad \text{A5.}$$

According to Axiom 5, to be rational at a node n means choosing that move that maximizes the expected payoff at that node. This involves knowing that, at the successor nodes n_r and n_l, $\pi^*(n_\mathrm{r})$ and $\pi^*(n_\mathrm{l})$ are defined. In other words, the player who has to move at node n must know that the successive player(s) are rational at nodes n_r and n_l. In order to define one's expected payoff at a node, one must know what the expected payoff of the following player is at the next node, and at nodes following that, etc., up to the end of the game. Axiom 5 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 further specifying the function $\max_i$, for example assuming that whenever a player is indifferent between m options, he will randomize over them with probability $1/m$. Such behavioral axioms will, however, be ad hoc, and they certainly are not part of the definition of rationality.

By a description of the game we mean a finite conjunction of formulas uniquely characterizing the tree representing the game as well as the structure of the payoffs

associated with the terminal nodes. Such a conjunction is obtained as follows. We assumed that T contains p nodes $n_1, \ldots, n_p$; if q of these nodes are leaves, let $v_1, \ldots, v_q$ be the associated payoffs (k-tuples of values). For brevity, let $\overline{b} = b_1, \ldots, b_p$ and $\overline{x} = x_1, \ldots, x_q$. In our theory, we use the variables $b_1, \ldots, b_p$ to refer to the nodes $n_1, \ldots, n_p$ of T, respectively. Similarly, we use the variables $x_1, \ldots, x_q$ to refer to the payoffs $v_1, \ldots, v_q$. This reflects the distinction between the formal language of the theory and the informal language within which we describe the game. Having established a correspondence between the nodes of T and the variables $\overline{b}$, we will refer from now on to the variable b_j as the variable corresponding to the node n_j. Then the description of the game and of the payoffs is the following conjunction, which we abbreviate by $\delta(\overline{b}, \overline{x})$:*

$$\left[\bigwedge_{n_i <_T n_j} (b_i <^T b_j) \wedge \bigwedge_{0<i\leq k} \bigwedge_{v_n \leq^i_V v_m} (x_n \leq^V_i x_m) \wedge \bigwedge_{p(n)=v} (\pi(b) = x) \right]$$

We now give the behavioral axioms. First we "lift" the function π to a function π^*, with domain $\subseteq T$ and values in V. Function π^* will be an extension of π, but it will *not*, in general, be total. Function π^* is supposed to represent each player's expected payoff at a node, and it will not supply a value unless a player has the "right" amount of knowledge. (Recall that this is achieved formally by taking the notation $\pi^*(\mathsf{b}) = x$ as our metalinguistic shorthand for the relation P.) The behavior of the function is specified by the following axioms:**

$$\pi^*(b) = \pi(b), \qquad \text{A6.}$$

for each variable b corresponding to a terminal node n. Let us abbreviate the conjunction of all such formulas by $\beta_1(\overline{b})$. These sentences say that the expected payoffs at the final nodes are just the payoffs associated with those nodes by the description of the game. For each non-terminal node we have:

$$K_{i_0} \ldots K_{i_q}(\mathrm{Rat}^{3-i}_{n_r} \wedge \mathrm{Rat}^{3-i}_{n_l}) \rightarrow \pi^*(b) = \max_i(\pi^*(b_r), \pi^*(b_l)), \qquad \text{A7.}$$

where n is non-terminal and: b is the variable corresponding to node n; q = the height of n (i.e., its distance from the root); $i = i_0$; $i_0 = 1$ if and only if q is even, and $i_0 = 2$ otherwise; finally, $i_{k+1} = 3 - i_k$, for each $k < q$. Let us abbreviate all such formulas by $\beta_2(\overline{b})$. (Some of the alternatives to this axiom are explored in Section 5.) Note that the string of leading K_i's in the antecedent of Axiom 7 represents the knowledge of the player who moves at n, and if n is the root, it 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

* Note that this conjunction is just the metalinguistic abbreviation of a formal object.

** Recall that in Definition 2 we have used b as a meta-variable for terms ranging over $V_1 \cup \{a\}$.

that the other players behave rationally, it is also necessary that he knows that they so behave at every node. We can now introduce the axiom

$$\exists!\overline{b}\exists!\overline{x}[\delta(\overline{b},\overline{x}) \wedge \beta_1(\overline{b}) \wedge \beta_2(\overline{b})] \qquad \text{A8.}$$

(where $\exists!$ means "there exist exactly"). Axiom 8 completely defines the structure of the game and π^* at every node, provided there is enough knowledge to calculate π^* at subsequent nodes.

Finally, we come to the special axiom specifying to what extent players' rationality is "common knowledge" among them. First, for each node n we specify a sentence Φ_n. We proceed by induction on (the tree representing) the game. If n is a leaf, we let Φ_n be a fixed sentence representing "the true" — for instance, $a = a$ (this is a mere technicality, intendend to take care of "unbalanced" trees); if n is a final node, then Φ_n is just Rat_n^i, where $i = 1$ if and only if the height of n is even, and $i = 2$ otherwise. If n is a non-final, non-terminal node, then

$$\Phi_n \equiv \mathrm{Rat}_n^i \wedge K_i(\Phi_{n_r} \wedge \Phi_{n_l}),^*$$

where, again, $i = 1$ if and only if the height of n is even. Then our last axiom, Axiom 9, is Φ_a (recall that a is the root of the tree).

We claim that the theory T_G, comprising axioms 1-9, is consistent and sufficient to infer the backwards induction equilibrium.

THEOREM 1 Theory T_G is consistent.

The proof of Theorem 1, together with the proof of the following theorem, can be found in the Appendix.

THEOREM 2 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_n or R_n for some node n) representing the branch through G corresponding to an equilibrium.

Theory T_G allows the players to infer an equilibrium path(s). Note that T_G refutes any proposition to the effect that a move off the solution path(s) takes place. Indeed, to choose an off-equilibrium move would mean violating Axiom 5.

4. Deviations

We now turn our attention to the way deviations from equilibrium can be handled in the framework of our theory. First observe that, as long as the logic we adopt is monotonic, any theory of the game that is sufficient to infer a solution becomes inconsistent when augmented with information that a move outside the solution

* Note that each Φ_i is a conjunction. Since we have assumed the operator K_i to apply only to formulas with no propositional connectives, sentences of the form $K_i(a \wedge b)$ should be construed as an abbreviation for $K_i a \wedge K_i b$.

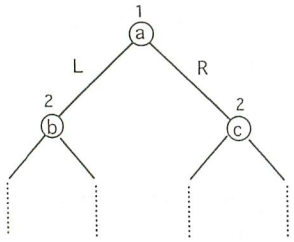

Fig. 2. A schematic game.

path has taken place. In order to preserve consistency, a revision of the theory is in order. However, even if a revision is necessary in any theory of the game whose underlying logic satisfies monotonicity, there are more or less drastic revisions. As we shall presently show, the magnitude of a revision will depend on whether the information relative to a node does or does not imply something about previous nodes in the game.

When we analyze how the theory of the game has to be modified in the face of a deviation from equilibrium, it is necessary to draw a careful 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. As we already mentioned, it is useful to resort to the following metaphor: We imagine each player to be represented by an *automatic theorem prover* that is supplied a 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. It is indeed plausible to assume that players are capable of *revising* their own theory in the face of inconsistencies, but this requires that we endow them with a meta-language within which to express belief revision, as well as with a meta-theoretic account of belief revision. Any talk of belief revision should thus be understood to be taking place at a meta-theoretic level. As it turns out, since our players only have distributed knowledge of the theory of the game, we do not need to endow them with a model for belief revision.

Let us now take some time to explore three possible candidate theories to be assigned to a node n of G. We want to assess their relative merits with respect to the way in which they handle deviations from equilibrium play. Suppose that, for simplicity, we have a game G with root a, whose left- and right-hand children are denoted by b and c (see Figure 2). Player 1 moves at node a. We want to consider some combination of the theories Φ_a, Φ_b, Φ_c. Notice that although Φ_a is recursively defined in terms of Φ_b and Φ_c, it does not entail either one of them. This has to do with our construal of the operators K_i as *belief* operators for which the axiom schemata $K_i\varphi \to \varphi$ are *not* assumed. Then, keeping axioms A1–A8 fixed, we consider assigning a theory of the game to each node n of G as follows.

Case 1: we assign to each node $n \in G$ the same theory Φ_a. That is, we make Φ_a group knowledge among the players.* Suppose that playing R_a is a strictly dominated strategy. Then, as we already know, the theory $\Phi_a \wedge R_a$ 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 an equilibrium for the subgame having c as its root. But

$$\Phi_a \equiv \mathrm{Rat}_a^1 \wedge K_1(\Phi_b \wedge \Phi_c);$$

clearly Φ_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 $\mathrm{Rat}_a^1 \wedge K_1(\Phi_c)$, neither of whose conjuncts is enough to infer an equilibrium. Having rejected Φ_a, player 2 has *no* theory of the game to speak of; what she does at node c is undefined. The situation of player 1 is quite different: Φ_a allows him to calculate $\pi^*(a)$, and thus choose an optimal move at node a. Of course $\Phi_a \wedge R_a$ is inconsistent, but all that player 1 needs to infer his optimal move is just Φ_a. In general, to compute the expected payoff at a node it is not necessary to assume that that node has been reached.

Case 2: we assign to each node $n \in G$ the same theory $\Phi_a \wedge \Phi_b \wedge \Phi_c$. 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 2 cannot but reject Φ_a. However, Φ_c still is sufficient to infer an equilibrium, i.e., to compute a value for $\pi^*(c)$. *Case 3*: we assign to each node $n \in G$ the theory Φ_n. This is the approach we have adopted, which calls for assigning to each node n 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 2 will find herself to choose at node c on the basis of the theory Φ_c (or, perhaps, $\Phi_c \wedge R_a$). But whereas $\Phi_a \wedge R_a$ is obviously inconsistent, $\Phi_c \wedge R_a$ is not. If player 1 were to ask what would happen in case he were to play R_a, he would know that player 2 would play her optimal move at node c, since Φ_c is still sufficient to infer an equilibrium for the subgame starting at c (i.e., $\pi^*(c)$ can be computed). When the theory of the game is modular, a deviation at a node does not force a revision at later nodes, since the theories at later nodes are not inconsistent with a statement to the effect that a deviation at some previous node has occurred.**

Our framework makes it easy to understand in which sense it can be argued that common knowledge of rationality leads to inconsistencies in the face of deviations.‡

---

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

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

‡ See for example Bicchieri (1989), Binmore (1987), and Reny (1988).

As it should now be clear, such inconsistencies arise only at the level of the object-language (i.e., the players' theory of the game); there is no inconsistency at the meta-linguistic level of justification of backwards induction.

To define common knowledge of rationality in our model, recall that the notion we employ is that of rationality *at a node*. 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^* are defined. Conversely, an agent i's being *not rational* at a node n means that $\pi^*(b)$ is not defined (where b is the variable assigned to n in our theory). On a local construal of rationality, assuming common knowledge of rationality amounts to saying that a player's expected payoff at a given node is common knowledge. Since in our axiomatization $\pi^*(b)$ is not defined unless it is defined at all lower nodes in the tree, common knowledge of rationality means that the value of $\pi^*(a)$ is common knowledge among the players (where a is the root of the tree). Equivalently, since such a value is determined by Φ_a, we can identify common knowledge of rationality with common knowledge of Φ_a.

We have shown in *Case 1* above that when the theory of the game is group knowledge among the players such theory becomes inconsistent with the statement that a deviation from equilibrium play has occurred. A fortiori, when the theory is common knowledge among the players, such theory becomes inconsistent when augmented with information that a move outside the solution path has taken place. When the theory is common knowledge (or group knowledge), the inconsistency is not limited to the node where the deviation occurs, but it spreads to the whole game. In order to preserve consistency, a revision of the theory at every node is in order. In both cases, we might want to endow the players with a meta-language within which to express belief-revision. In our model, however, players do not need such a meta-language, nor a theory of belief revision, since they have just enough knowledge to infer an optimal move at a node or, in other words, they have "distributed" knowledge of the theory of the game.

5. Alternative Accounts

Let us now explore two alternatives to our crucial axiom $A7$ (Recall that A5 and A7 are axiom schemas that stand for finite collections of sentences, one for each node). The intended meaning of axiom $A7$ is that *if* player i_0 has the "right" amount of knowledge, *and* function π^* is defined on the children of n, *then* it is defined on node n too. For player i to choose what to do at node n, it is necessary both that the other players behave rationally and that i *knows* that they so behave. Hence, the string of leading K_i's in the antecedent of $A7$. However, all that is needed in order to infer the backwards induction equilibrium is the consequent of $A7$. So it is worth considering what would happen if we were to replace axiom $A7$ by (i) its consequent, thus obtaining the sentence

$$\pi^*(b) = \max_i(\pi^*(b_r), \pi^*(b_l));$$

or (*ii*) replace $A7$ with a sentence in which we drop the leading K_i's from its antecedent, thus obtaining

$$(\mathrm{Rat}_{n_\mathrm{r}}^{3-i} \wedge \mathrm{Rat}_{n_\mathrm{l}}^{3-i}) \rightarrow \pi^*(b) = \max_i(\pi^*(b_\mathrm{r}), \pi^*(b_\mathrm{l}))$$

(and modify Φ_n by analogously dropping the occurrence of K_i).

In both cases our modified theory would still be sufficient to infer a backwards induction equilibrium. This 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 a meta-theoretic level, what kind of belief revisions does this warrant? The two cases crucially differ between themselves, and with our proposal, in the way *deviations from equilibrium* can be handled.

Case (*i*) is simply classical backwards induction: We have replaced $A7$ with the new axiom

$$\pi^*(b) = \max_i(\pi^*(b_\mathrm{r}), \pi^*(b_\mathrm{l}));$$

and axiom Φ_a (Axiom 9) is not needed in this case to infer an equilibrium. The theory thus modified 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 accommodate a deviation.* The only conclusion is that the other player acted against her own best interests for mysterious reasons.

Consider as an example the game in Figure 1, and suppose that player 2 observes r_l. Given our modified theory T_G (we have now changed axiom $A7$), both players are able recursively to define the value of $\pi^*(b)$ at each node b, and in particular both players know that $\pi^*(a)$ is defined. By axiom $A5$, if $\pi^*(a)$ is defined, then $\mathrm{Rat}_a^\mathrm{l}$. Observing r_l forces player 2 to abandon $A5$ for node a, i.e., to abandon the assumption that player 1 is rational at node a.

Case (*ii*) is different: we replace $A7$ by the new axiom

$$(\mathrm{Rat}_{n_\mathrm{r}}^{3-i} \wedge \mathrm{Rat}_{n_\mathrm{l}}^{3-i}) \rightarrow \pi^*(b) = \max_i(\pi^*(b_\mathrm{r}), \pi^*(b_\mathrm{l}))$$

and correspondingly drop occurrences of K_i from A9. 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 $A5$: She can now revise the theory of the game by assuming that player 1 is not rational, at least at node a, *because* $\pi^*(a)$ may not be defined. In turn, $\pi^*(a)$ may be undefined if player 1 does not know that player 2 is rational at node b, or if $\pi^*(b)$ is not defined because player 2 does not know that $\mathrm{Rat}_c^\mathrm{l}$. 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 $A7$ leaves open the possibility that a player's rational behavior at a node is not *known* by

* 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 $A5$ or, for that matter, by changing the structure of the payoffs of the game: we do *not* regard these as *natural* belief revisions.

another player, which would serve equally well to "explain" the latter's deviation at a previous node (this possibility is discussed as *Case 2* in Section 4). Case (*ii*) and our proposal differ, however, in another important respect. First note that in case (*ii*), but *not* in our proposal, if node n' is a descendant of node n in G, then Φ_n implies $\Phi_{n'}$. It follows that if a deviation at node n' is observed, it is not only the theory $\Phi_{n'}$ that needs to be revised, but also Φ_n. This is not the case with $A7$ 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 (again, consider *Case 2* in Section 4).

6. Conclusion

We have argued that there is a distinction between justifying a certain solution (e.g., backwards induction) and inferring or computing that solution. Furthermore, we have suggested that a distinction must also be drawn between the meta-language (i.e., the language of justification) and the object-language (the language of the players). Meta-theoretic justification is the game theorist's task; computing a solution is the task of the players. While the backwards induction solution has been successfully justified (as we did show in Section 2), no formal theory of the game from the players' viewpoint has been developed. Our paper has provided the players with a formal (i.e., axiomatic) theory of the game. We have proved such theory to be consistent, and sufficient for the players to infer the backwards induction solution directly from the structure of the game alone. Finally, we did show that (i) common knowledge of rationality (or of the theory of the game) is not needed by the players in order to infer the backwards induction solution. (ii) If the players' theory of the game is common knowledge (or even group-knowledge) among them, a deviation from equilibrium play forces a global revision. In this case, we should endow the players with a meta-language within which to conduct belief revision. If the theory of the game is modular (or distributed knowledge), a deviation does not force a revision at subsequent nodes. (iii) The game theorist's justification of backwards induction can include a statement to the effect that players have common knowledge of rationality.

Appendix

This Appendix contains proofs of the results in Section 3.

THEOREM 1 Theory T_G is consistent.

Proof. We establish the claim by exhibiting a model $\mathcal{M}$. Such a model will be the formal counterpart of the game represented by T, under the hypothesis that the backwards induction equilibrium is played. To make things simple, suppose that

there are no ties in the payoffs, so that there is only one sequence of moves through T that is consistent with backwards induction. We set:

$$\mathcal{M} = ((T, <_T), (V, \leq_V^1, \ldots, \leq_V^k), W, R_1, \ldots, R_k, P^*, m_1, \ldots, m_k, p, I),$$

where:

1. T is the game tree T for which we are giving the theory;
2. V is the set of k-tuples of real numbers representing the payoff vectors associated with the terminal nodes of T, and the relation $\leq_V^i$ holds between two vectors $\mathbf{x} = x_1 \ldots x_n$ and $\mathbf{y} = y_1 \ldots y_n$ if and only if $x_i \leq y_i$ (as usual, k = the number of players);
3. $W = \{w\}$, i.e., W contains only one world.
4. all the R_i relations $(0 < i \leq k)$ are the same, namely the universal relation over W, which is serial;
5. p is the function that assigns to each terminal node in T the corresponding payoff vector (and assigns arbitrary values to non-terminal nodes — π is never applied to non-terminal nodes in the theory);*
6. m_i is a function taking as input two vectors from V, say $\mathbf{x}$ and $\mathbf{y}$ (in this order), returns the one with the higher i-th projection if it exists, and returns $\mathbf{x}$ otherwise;**
7. P^* is defined inductively on the generation of the tree, beginning with the leaves as a base case; if n is a leaf, $P^*(n, \mathbf{x})$ holds if and only if $\mathbf{x} = p(n)$; if n is not a leaf, $P^*(n, \mathbf{x})$ holds if and only if $\mathbf{x} = m_i(\mathbf{y}, \mathbf{z})$, where $P^*(n', \mathbf{y})$ and $P^*(n'', \mathbf{z})$, and n', n'' are the children of n (this definition implies that for each node n there is an $\mathbf{x}$ such that $P^*(n, \mathbf{x})$);‡
8. I is defined as follows:
   (a) $I(a, w)$ = the root of T, for every $w \in W$;
   (b) $I(L_n, w)$ = true for every $w \in W$, if the left-hand child of n lies along the solution path, and $I(L_n, w)$ = false for every $w \in W$ otherwise;
   (c) $I(R_n, w)$ = true for every $w \in W$, if the right-hand child of n lies along the solution path, and $I(R_n, w)$ = false for every $w \in W$, otherwise;
   (d) $I(\mathrm{Rat}_n^i, w)$ = true for every $w \in W$.

It is immediate to verify that $\mathcal{M}$ is a model for T_G. The structural axioms are obviously satisfied, since they were formulated in such a way as to be true of T. As to the behavioral axioms, $\pi^*(\mathrm{b}) = \pi(\mathrm{b})$ (where b corresponds to a final node) holds since P^* extends p (viewed as a relation). Moreover, since all sentences $\pi^*(b) = \max_i(\pi^*(b_\mathrm{r}), \pi^*(b_\mathrm{l}))$ (where b is non-terminal) are true by definition of P^*, the axioms

$$K_{i_0} \ldots K_{i_q}(\mathrm{Rat}_{n_\mathrm{r}}^{3-i} \wedge \mathrm{Rat}_{n_\mathrm{l}}^{3-i}) \rightarrow \pi^*(b) = \max_i(\pi^*(b_\mathrm{r}), \pi^*(b_\mathrm{l})),$$

* Recall that the symbol π is to be interpreted by p in any model for $\mathcal{L}$.
** Recall that the symbol $\max_i$ is to be interpreted by m_i in any model for $\mathcal{L}$.
‡ Recall that the symbol P is to be interpreted by P^* in any model for $\mathcal{L}$.

are true. Finally, all the Rat_n^i constants are true, and since the accessibility relations R_i are all universal, all sentences of the form

$$K_0 \ldots K_q \mathrm{Rat}_n^i$$

are also true. Therefore the axiom Φ_a is true. To complete the proof it remains to observe that in the model each player moves at least once at each non-terminal node that is reached in the backwards induction equilibrium, and that all and only such nodes are reached. ■

THEOREM 2 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_n or R_n for some node n) representing the branch through G corresponding to an equilibrium.

Proof. It suffices to show that $\pi^*(a)$ is defined. We proceed by induction on (the tree representing) G. If G comprises a unique final node n, then it suffices to invoke Axiom 6 (i.e., the conjunct $\beta_1(\bar{b})$ of Axiom 8.).

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 1 and 2), theories T_{G_b} and T_{G_c} are sufficient to infer that $\pi^*(b)$ and $\pi^*(c)$ are defined.

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 interpretation of K_i as a belief operator, and the way $\beta_2(\bar{b})$ and Φ_a (Axiom 9) have been formulated. It is indeed one of the characteristic features of the present approach that if node n' is a descendant of node n, then Φ_n does *not* imply Φ'_n.

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 n in G_b or G_c, they contain the corresponding instance of Axiom 7, which has the form

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

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

$$K_1 \ldots K_n \varphi.$$

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

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

(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 just as a value for $\pi^*(b)$ or $\pi^*(c)$ can be obtained in T_{G_b} or T_{G_c}, so it will be obtained in T_G, too.

All that is left to observe is that T_G contains the following instance of Axiom 7:

$$K_1(\mathrm{Rat}_b^2 \wedge \mathrm{Rat}_c^2) \Longrightarrow \pi^*(a) = \max_1(\pi^*(c), \pi^*(b)),$$

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

References

Antonelli, A. and Bicchieri, C., 1994, "Forward Induction," Report CMU-PHIL-58.

Aumann, R.J., 1976, "Agreeing to disagree," *Annals of Statistics* **4**, 1236–1239.

Aumann, R.J., 1995, "Backward Induction and Common Knowledge of Rationality", *Games and Economic Behavior* **8**, 6–19.

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

Bicchieri, C., 1988, " Strategic Behavior and Counterfactuals," *Synthese* **30**, 135–169.

Bicchieri, C., 1989, "Self Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntnis* **76**, 69–85.

Bicchieri, C., 1992, "Knowledge-dependent Games: Backwards Induction," in *Knowledge, Belief, and Strategic Interaction*, C. Bicchieri and M.L. Dalla Chiara, eds., Cambridge: Cambridge University Press.

Bicchieri, C., 1993, *Rationality and Coordination*, Cambridge: Cambridge University Press.

Binmore, K., 1987, "Modeling Rational Players, Part I," *Economics and Philosophy* **3**, 179–214.

Bonanno, G., 1991, "The Logic of Rational Play in Games of Perfect Information," *Economics and Philosophy* **7**, 37–61.

Lewis, D., 1969, *Convention*, Cambridge, MA: Harvard University Press.

Pettit, P. and Sugden, R., 1989, "The Backwards Induction Paradox," *Journal of Philosophy* **4**, 1–14.

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

Shin, H., 1992, "Counterfactuals and a Theory of Equilibrium in Games," in *Knowledge, Belief, and Strategic Interaction*, C. Bicchieri and M.L. Dalla Chiara, eds., Cambridge: Cambridge University Press.

Stalnaker, R., *Knowledge, Belief, and Counterfactual Reasoning in Games*, forthcoming in the *Proceedings of the Second Castiglioncello Conference 1995*, C. Bicchieri, R. Jeffrey and B. Skyrms, eds.