Backward Induction without Common Knowledge
Author(s): Cristina Bicchieri
Reviewed work(s):
Source: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1988, Volume Two: Symposia and Invited Papers (1988), pp. 329-343
Published by: The University of Chicago Press on behalf of the Philosophy of Science Association
Stable URL: http://www.jstor.org/stable/192895
Accessed: 26/12/2011 20:30

Backward Induction without Common Knowledge

Cristina Bicchieri

Carnegie-Mellon University

1. Information and meta-information

Game theory studies the behavior of rational players in interactive situations and its possible outcomes. For such an investigation, the notion of players' rationality is crucial. While notions of rationality have been extensively discussed in game theory, the epistemic conditions under which a game is played — though implicitly presumed — have seldom been explicitly analyzed and formalized. These conditions involve the players' reasoning processes and capabilities, as well as their knowledge of the game situation.[1] Game theory treats some aspects of information about chance moves and other players' moves by means of information partitions in extensive form games. But a player's knowledge of the structure, for example, of information partitions themselves is different from his information about chance moves and other players' moves. The informational aspects captured by the extensive form games have nothing to do with a player's knowledge of the structure of the game.

Game theorists implicitly assume that the structure of the game is *common knowledge* among the players. By 'common knowledge of p' is meant that p is not just known by all the players in a game, but is also known to be known, known to be known to be known, ... *ad infinitum*.[2] The very idea of a Nash equilibrium is grounded on the assumptions that players have common knowledge of the structure of the game and of their respective priors. These assumptions, however, are always made *outside* the theory of the game, in that the formal description of the game does not include them.[3]

The assumptions about players' rationality, the specification of the structure of the game, and the players' knowledge of all of them should be part of the theory of the game. Recent attempts to formalize players' knowledge as part of a theory of the game include Bacharach (1985, 1987), Gilboa (1986), Mertens and Zamir (1985), Brandenburger and Dekel (1985), Kaneko (1987) and Samet (1987). In these works a common knowledge axiom is explicitly introduced, stating that the axioms of logic, the axioms of game theory, the behavioral axioms and the structure of the game are all common knowledge among the players.

Is it always necessary for the players to have common knowledge of the theory of the game for a solution to be derived? Different solution concepts may need different amounts of knowledge on the part of the players to have predictive validity at all. For example, while

---

PSA 1988, Volume 2, pp. 329-343

common knowledge is necessary to attain an equilibrium in a large class of normal form games, it may lead to inconsistencies in finite, extensive form games of perfect information (Reny 1987, Bicchieri 1989).[4] More generally, if players' epistemic states and their degree of information about other players' epistemic states are included in a theory of the game, which solutions to non-cooperative games can be derived? I believe the consequences of explicitly modeling players' knowledge as part of the theory of the game are far-reaching.

In this paper I examine finite, extensive form games of perfect and complete information. These games are solved working backwards from the end, and this procedure yields a unique solution. It is commonly assumed that backward induction can only be supported by common knowledge of rationality (and of the structure of the game). In section 2 it is proved instead that the levels of knowledge of the theory of the game (hence, of players' rationality) needed to infer the backward induction solution are finite.

That limited knowledge is sufficient to infer a solution for this class of games does not mean it is also a necessary condition. In section 3, I introduce the concepts of *knowledge-dependent games* and *knowledge-consistent play*, and prove that knowledge has to be limited for a solution to obtain. More specifically, it is proved that for the class of games considered here backward induction equilibria are knowledge-consistent plays of knowledge-dependent games. Conversely, every knowledge-consistent play of a knowledge-dependent game is a backward induction equilibrium.

For the class of games considered, there exist knowledge-dependent games that have no knowledge-consistent play. For example, a player might be unable — given what she knows — to 'explain away' a deviation from equilibrium on the part of another player, in that reaching her information set is inconsistent with what she knows.

If the theory of the game were to include the assumption that *every* information set has a small probability of being reached (because a player can always make a mistake), then no inconsistency would arise. In this case, the solution concept is that of *perfect equilibrium* (Selten 1975), which requires an equilibrium to be stable with respect to 'small' deviations. The idea of perfect equilibrium (like other 'refinements' of Nash equilibrium) has the defect of being *ad hoc*, as well as of assuming — as Selten himself has recognized — less than perfect rationality.[5]

The present paper has a different goal. What I want to explore here is under which epistemic conditions a rationality axiom can be used to derive a unique prediction about the outcome of the game. As it will be made clear in the example of section 2, a small variation in the amount of knowledge possessed by the players can make a big difference, in that higher levels of knowledge of the theory of the game may make the players unable to 'explain away' deviations from the equilibrium path. The idea is that of finding the minimal set of axioms from which a solution to the game can be inferred.

Since the players (as well as the game theorist) have to reason to an equilibrium, the theory must contain a number of meta-axioms stating that the axioms of the theory are known to the players. In particular, the theory of the game T can contain a meta-axiom A_n stating that the set of game-theoretic ('special') axioms A_1-A_{n-1} is k-level group-knowledge among the players, but not a meta-axiom A_{n+1} saying that A_n is group-knowledge among the players. If A_{n+1} is added to T, it becomes group-knowledge that the theory is inconsistent at some information set. In this case, the backward induction solution cannot be inferred.

2. Backward induction equilibrium

In this section non-cooperative, extensive form games of perfect information are defined and it is proved that the levels of knowledge needed to infer the backward induc-

tion equilibrium are *finite,* contrary to the common assumption that only an infinite iteration of levels of knowledge (i.e., common knowledge) can support the solution.

Definition 2.1. A non-cooperative game is a game in which no precommitments or binding agreements are possible.

Definition 2.2. A finite n-person game Γ of perfect information in extensive form consists of the following elements:

(i) A set $N = \{1, 2, ..., n\}$ of players.

(ii) A finite tree (a connected graph with no cycles) T, called the game tree.

(iii) A node of the tree (the root) called the first move. A node of degree one and different from the root is called a terminal node. Ω denotes the set of all terminal nodes.

(iv) A partition $P^1, ..., P^n$ of the set of non-terminal nodes of the tree, called the player partition. The nodes in P^i are the moves of player i. The union of $P^1, ..., P^n$ is the set of moves for the game.

(v) For each $i \in N$, a partition $I^{i1}, ... , I^{ik}$ of P^i (I^{ij} denotes the j-th information set ($j \geq 1$) of player i) such that for each $j \in \{1, ..., k\}$:
(a) each path from the root to a terminal node can cross I^{ij} at most once, and
(b) since there is perfect information, I^{ij} is a singleton set for every i and j.

(vi) For each terminal node t, an n-dimensional vector of real numbers, $f^1(t), ... , f^n(t)$ called the payoff vector for t.

Every player in Γ knows (i)-(vi).

Definition 2.3. A *pure strategy* s^i for player i is a k-tuple that specifies, for each information set of player i, a choice at that information set. The set of i's pure strategies is denoted by $S^i = \{s^i\}$. Let $\mathbf{S} = S^1 \times .... \times S^n$. A *mixed strategy* x^i for player i is a probability distribution over player i's pure strategies.

Definition 2.4. The function $\pi^i : S^1 \times .... \times S^n \rightarrow \Re$ is called the *payoff function* of player i. For an n-tuple of pure strategies, $s = (s^1, ... , s^n) \in S$, the expected payoff to player i, $\pi^i(s)$, is defined by

$$\pi^i(s) = \sum_{t \in \Omega} p_s(t)\, \pi^i(t)$$

where $p_s(t)$ is the probability that a play of the game ends at the terminal node t, when the players use strategies $s^1, ... , s^n$.

Definition 2.5. A pure strategy n-tuple $s = (s^1, ... , s^n) \in \mathbf{S}$ is an *equilibrium point* for Γ if

$$\pi^i(s|\, y^i) \leq \pi^i(s) \text{ for all } y^i \in S^i$$

where $s|\, y^i = (s^1, ..., s^{i-1}, y^i, s^{i+1}, ..., s^n)$.
We also say that $s^i \in S^i$ is a *best reply* of player i against s if $\pi^i(s|\, s^i) = \max_{y^i \in S^i} \pi^i(s|\, y^i)$.

Definition 2.6. A subgame $\Gamma_j \in \Gamma$ is a collection of branches of the game that start from the same node and the branches and node together form a game tree by itself.

Theorem 2.1. (Kuhn 1953) A game Γ of perfect information has an equilibrium point in pure strategies.

Proof. By induction on the number of moves in the game. Suppose the game has only one move. Then the player who has to move, in order to play an equilibrium strategy, should choose the branch which leads to a terminal node with the maximum payoff to him. Therefore the theorem is true when Γ has one move. Suppose the theorem is true for games with at least K moves ($K \geq 1$). Let Γ be a game with at most K+1 moves, where T is the game tree for Γ, r the root of T, and k the number of branches going out of r (these branches are numbered from 1 to k). The node at the end of of the j-th branch from r is the root of a subtree T_j of T, where T_j is the tree for a subgame Γ_j of Γ (since the game is one of perfect information).

For each $s^i \in S^i$, let s^{ij} be the actions recommended by s^i at the information sets in Γ_j. Let $S^{ij} = \{s^{ij}\}$ and let $\pi^{ij}(s^j)$ be the expected payoff of player i in the subgame Γ_j, when the players play the combination of strategies $s^j = (s^{1j}, \ldots, s^{nj})$ ($j = 1, \ldots, k$). Each subgame Γ_j is a game of perfect information with K moves or less and by assumption it has an equilibrium $\bar{s}^j = (\bar{s}^{1j}, \ldots, \bar{s}^{nj})$, so that

$$(*)\ \pi^{ij}(\bar{s}^j | t^{ij}) \leq \pi^{ij}(\bar{s}^j) \text{ for all } t^{ij} \in S^{ij}$$

Consider now the root r of T. For some player $n \in N$, $r \in P^n$. Let $l \in \{1, \ldots, k\}$ be a branch departing from r such that

$$\pi^{nl}(\bar{s}^l) = \max_{1 \leq j \leq k} \pi^{nj}(\bar{s}^j)$$

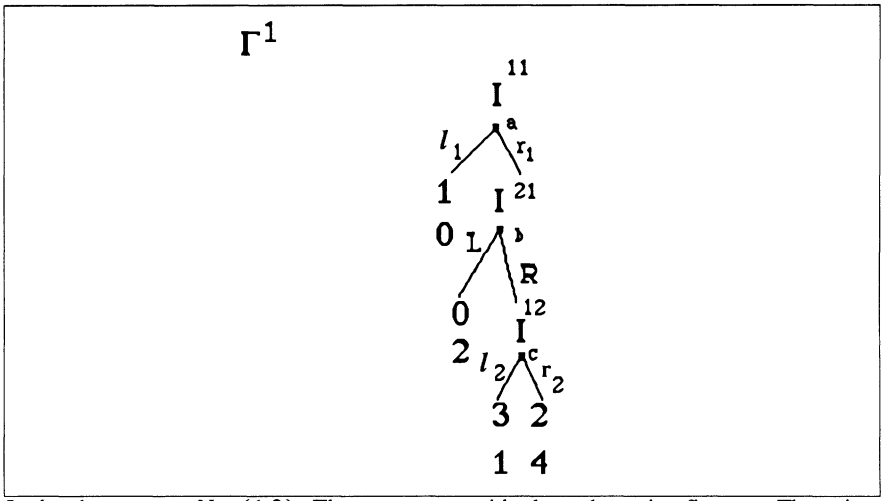
$\bar{s}^i \in S^i$ is defined as follows:

$$(a)\ \bar{s}^i = \prod_{j=1}^{k} \bar{s}^{ij} \quad \text{for } i \neq n \qquad (b)\ \{l\} \times \prod_{j=1}^{k} \bar{s}^{nj} \quad \text{for } i = n$$

Thus at information set I^i, the equilibrium strategy $\bar{s}^i$ of player i tells him to choose branch l if $I^i = \{r\}$, and to play $\bar{s}^{ij}(I^i)$ if I^i is an information set in Γ_j. We have to prove that $\bar{s} = (\bar{s}^1, \ldots, \bar{s}^n)$ is an equilibrium point for Γ.

For $i \neq n$, and $t^i = \prod_{j=1}^{k} t^{ij} \in S^i$, $\pi^i(\bar{s} | t^i) \leq \pi^i(\bar{s})$ by (*) and since by assumption player n chooses branch l at node r under $\bar{s}^n$. For $i = n$, and $t^n = \{m\} \times \prod_{j=1}^{k} t^{ij} \in S^n$, where m is one of the branches going out of r, $\pi^n(\bar{s} | t^n) \leq \pi^n(\bar{s})$ by (*) and since, by assumption, $\pi^{nl}(\bar{s}^l) = \max_{1 \leq j \leq k} \pi^{nj}(\bar{s}^j)$. Hence $\bar{s}$ is a pure strategy equilibrium for Γ.

The equilibrium can be found by working backwards from the terminal nodes to the root. At each information set, a player chooses the branch which leads to the subtree yielding him the highest equilibrium payoff. To illustrate this method, consider the following two-person extensive form game of perfect information with finite termination.



In the above game, N = {1,2}. The game starts with player 1 moving first at a. The union $P^1 \cup P^2$ is the set of moves {a, b, c}. P^1 = {a, c}, P^2 = {b}. I^{11} = {a}, I^{21} = {b}, I^{12} = {c}. Each player has two pure strategies: either to play left, thus ending the game, or to play right, in which case it is the other player's turn to choose. $S^1 = \{l_1l_2, l_1r_2, r_1r_2, r_1l_2\}$. S^2 = {L, R}. The payoffs to the players are represented at the endpoints of the tree, the upper number being the payoff of player 1, and each player is assumed to be rational (i.e., to wish to maximize his expected payoff).

The equilibrium described above for such games is obtained by backward induction as follows: at node I^{12} player 1, if rational, will play l_2, which grants him a maximum payoff of 3. Note that player 1 does not need to assume 2's rationality in order to make his choice, since what happened before the last node is irrelevant to his decision. Thus node I^{12} can be substituted by the payoff pair (3, 1). At I^{21} player 2, if rational, will only need to know that 1 is rational in order to choose L. That is, player 2 need consider only what she expects to happen at subsequent nodes (i.e., the last node) as, again, that part of the tree coming before is now strategically irrelevant. The penultimate node can thus be substituted by the payoff pair (0, 2). At node I^{11}, rational player 1, in order to choose l_1, will have to know that 2 is rational *and* that 2 knows that 1 is rational (otherwise, he would not be sure that at I^{21} player 2 will play L). From right to left, nonoptimal actions are successively deleted, and the conclusion is that player 1 should play l_1 at his first node. Thus $\bar{s}^1(I^{11}) = l_1$, $\bar{s}^2(I^{21})$ = L, $\bar{s}^1(I^{12}) = l^2$, and $(\pi^1(\bar{s}), \pi^2(\bar{s})) = (1, 0)$.

In the classical account of this game, $(l_1 L\ l_2)$ represents the only possible pattern of play by rational players because the game is one of *complete information*, i.e., the players know each other's rationality, strategies and payoffs. Player 1, at his first node, has two possible choices: 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 rational for him to play l_1 at the first node; otherwise he may play r_1. His conjecture about player 2's choice at the second node is based on what he thinks player 2 believes would happen if she played R. Player 2, in turn, has to conjecture what player 1 would do at the third node, given that she played R. Indeed, both players have to conjecture each other's conjectures and choices at each possible node, until the end of the game.

In our example, complete information translates into the conjectures $p(l_1) = 1$, p(R) = 0 and $p(r_2) = 0$. The notion of complete information does not specify any particular level of

knowledge that the players may possess, but it is customarily assumed by game theorists that the structure of the game and players' rationality are *common knowledge* among them.

Note, again, that *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.* The central idea is that if a player's strategy is to be part of a rational solution, then it must prescribe a rational choice of action in all conceivable circumstances, even those which are ruled out by some putative equilibrium. An equilibrium is thus endogenously determined by considering the implications of deviating from the specified behavior. The backward induction requirement calls for considering equilibrium points which are in equilibrium in each of the subgames and in the game considered as a whole. This means that it only matters where you are, not how you arrived there, as history of past play has no influence on what individuals do.

Since a strategy specifies what a player should choose in every possible contingency (i.e., at all information sets at which he may find himself), and a player's contingency plan ought to be rational in the contingency for which it was designed, it is necessary to give meaning to the idea of a choice conditional upon a given information set having being reached. Does it make sense to talk of a choice contingent upon other choices that may never occur? What counts as 'rational' behavior at information sets not reached by the equilibrium path depends on how a player explains the fact that a given information set is reached, since different explanations elicit different choices. For example, it has been argued that at I^{21} it is not evident that player 2 will only consider what comes next in the game (Binmore 1987; Reny 1987). Reaching I^{21} may not be compatible with backward induction, since I^{21} *can only be reached if 1 deviates from his equilibrium strategy, and this deviation stands in need of explanation.* When player 1 considers what player 2 would choose at I^{21}, he has to have an opinion as to what sort of explanation 2 is likely to give for being called to play, since 2's subsequent action depends on it. Binmore's criticism rightly points out that *a solution must be stable also with respect to forward induction.* In other words, if equilibrium behavior is determined by behavior off the equilibrium path, a solution concept must allow the players to 'explain away' deviations.

Selten's 'trembling hand' model (Selten 1975) provides the canonical answer. According to Selten, we must suppose that whenever a player wants to make some move a, he will have a small positive probability ε of making a different and unintended move $b \neq a$ instead by 'mistake'. If any move can be made with a positive probability, all information sets have a positive probability of being reached.

What relates Selten's theory of mistakes to backward induction? Since the backward induction argument relies on the notion of players' rationality, one has to show that rationality and mistakes are compatible. Admitting that mistakes can occur means drawing a distinction between deciding and acting, but a theory that wants to maintain a rationality assumption is bound to make mistakes entirely random and uncorrelated. Systematic mistakes would be at odds with rationality, since one would expect a rational player to learn from past actions and modify his behavior. If a deviation tells that a player made a mistake (i.e., his hand 'trembled'), but not that he is irrational, a mistake must not be the product of a systematic bias in favor of a particular type of action, as would be the case with a defective reasoning process.

In our example, when player 2 finds she has to move, she will interpret 1's deviation as the result of an unintended, random mistake. So if 1 plays (but did not choose to play) r_1, 2 knows that the probability of r_2 being successively played remains vanishingly small, viz. $p(r_2) = p(r_2|r_1) = \varepsilon$. This makes 2 choose strategy L, which is a best reply to player 1's strategy after allowing for the possibility of trembles. Player 1 knows that, were he to play r_1, player 2 would want to respond with L, and that there is only a vanishingly small probability that R is played instead. For $p(R) = \varepsilon$, player 1's best reply is l_1.

Thus (l_1L l_2) remains an equilibrium in the new 'perturbed' game that differs from the original game in that any move has a small positive probability of being made.

According to Binmore (1987, 1988), this characterization of mistakes is necessary for the backward induction argument to work, in that it makes out of equilibrium behavior compatible with players' rationality. Otherwise, Binmore argues, a deviation would have to be interpreted as proof of a player's 'irrationality'. Is this conclusion warranted? If common knowledge of rationality is assumed, then one must also offer some argument to explain how a player, facing a deviation, can still be able to maintain without contradiction that the deviator is rational. Selten's 'trembling hand' hypothesis is not the only plausible one, but certainly it is an answer.[6] But is common knowledge of rationality at all needed to get the backward induction solution?

Binmore and Reny have been skeptical of the classical solution precisely because they did not question the common knowledge assumption. In what follows, I show that the backward induction solution can be inferred from a set of assumptions that include a specification of players' knowledge. *The levels of knowledge needed for the solution to obtain are finite, and their number depends on the length of the game.*

A play of the game we have just described makes a number of assumptions about players' rationality and knowledge, from which the backward induction solution necessarily follows. Let us consider them in turn. First of all, the players know their respective strategies and payoffs. Second, the players are rational, in the sense of being expected utility maximizers. Third, the players have group-knowledge of rationality and of the structure of the game. This means that each player knows that the other player is rational, and knows the other player's strategies and payoffs. Is this information sufficient to infer a solution to the game?

It is easy to verify that in the above game *different levels of knowledge* are needed at different stages of the game for backward induction to work. For example, if R_1 stands for 'player 1 is rational', R_2 for 'player 2 is rational', and K_2R_1 for 'player 2 knows that player 1 is rational', R_1 alone will be sufficient to predict 1's choice at the last node, but in order to predict 2's choice at the penultimate node, one must know that rational player 2 knows that 1 is rational, i.e. K_2R_1. K_2R_1, in turn, is not sufficient to predict 1's choice at the first node, since 1 will also have to know that 2 knows that he is rational. That is, $K_1K_2R_1$ needs to obtain. Moreover, while R_2 only (in combination with K_2R_1) is needed to predict L at the penultimate node, K_1R_2 must be the case at I^{11}.

Theorem 2.2. In finite extensive form games of perfect and complete information, the backward induction solution holds if the following conditions are satisfied for any player i at any information set I^{ik} : (α) player i is rational and knows it, and knows his available choices and payoffs, and (β) for every information set I^{jk+1} that immediately follows I^{ik}, player i knows at I^{ik} what player j knows at information set I^{jk+1}.

Proof. The proof is by induction on the number of moves in the game. If the game has only one move, the theorem is vacuously true since at information set I^{i1}, if player i is rational and knows it, and knows his available choices and payoffs, he will choose that branch which leads to the terminal node associated with the maximum payoff to him and this is the backward induction solution. Suppose the theorem is true for games involving at most K moves (some K$\geq$ 1). Let Γ be a game of perfect and complete information with K+1 moves and suppose that conditions α and β are satisfied at every node of game Γ. Let r be the root of the game tree T for Γ. At information set I^{ir}, player i knows that conditions α and β are satisfied at each of the subgames starting at the information sets that immediately follow I^{ir}. Then at I^{ir} player i knows that the outcome of play at any of those subgames would correspond to the backward induction solution for that subgame. Hence at I^{ir} if player i is rational, he will choose the branch going out of r which leads to the subgame whose backward induction solution is best for him, and this is the backward induction solution for game Γ.

3. Knowledge-dependent games

Theorem 2.2 tells that, for the backward induction solution to hold, we do not need to assume common knowledge but only *limited knowledge* of rationality and of the structure of the game. All that is needed is that a player, at any of her information sets, knows what the next player to move knows. Thus the player who moves first will know more things than the players who move immediately after, and these in turn will know more than the players who follow them in the game. However, if the same player has to move at different points in the game, we want that player's knowledge to be the same at all of his information sets. This requirement has a natural interpretation in the normal form representation of such games.

Consider the normal form equivalent of game Γ^1

		2	
		L	R
1	l_1	1,0	1,0
	$r_1 l_2$	0,2	3,1
	$r_1 r_2$	0,2	2,4

In this game, strategy $r_1 l_2$ weakly dominates $r_1 r_2$, so if 2 knows that 1 is rational 2 will expect 1 to eliminate $r_1 r_2$. In the extensive form representation, this corresponds to player 2 knowing that rational player 1, at the last node, will choose l_2. *In order to eliminate his weakly dominated strategy*, player 1 need not know whether 2 is rational. This corresponds to the last node of the extensive form representation, where 1 does not need to consider what happened before, since it is now strategically irrelevant. Player 1 needs to know that 2 is rational only when, having eliminated $r_1 r_2$, he has made L weakly dominant over R. Note that player 1, in order to be sure that 2 will choose L, has to know that 2 is rational *and* that 2 knows that 1 is rational, otherwise there would be no weakly dominated strategy for player 2 to delete. Having thus deleted R, 1's best reply to L is l_1. And this corresponds to the first node, where player 1 has to know that 2 is rational and that 2 knows that 1 is rational. Evidently player 1 needs to know *more* than player 2, even in the normal form, since the order of iterated elimination of dominated strategies starts with player 1's strategy $r_1 r_2$. In the extensive form the backward induction argument makes player 1's previous knowledge irrelevant at his subsequent node, but this does not mean that player 1 knows less. This point becomes even clearer if we remember that we are dealing with static games: a player can plan a strategy in advance and then let a machine play on his behalf.

Given that the solution for this class of games depends upon the information possessed by the players, we may want to know whether variations in the level of knowledge would make a difference. Since only limited knowledge is sufficient to infer the backward induction solution, is it also a necessary condition? We know that assuming common knowledge leads to an inconsistency (Reny 1987; Bicchieri 1989), but is an inconsistency produced by simply assuming levels of knowledge higher than those which are sufficient to infer the solution? In particular, it is worth exploring what would happen were the players to know what the players preceding them know, i.e., what would happen were knowledge to go in both directions.

In order to address this issue, we have to explicitly model players' knowledge of the game, as well as the reasoning process that leads them to choose a particular sequence of actions. The theory of the game will have to include a set of assumptions specifying what the players know about the structure of the game and the other players. The main result of this section is that, for any finite extensive form game of perfect and complete

information, *the levels of knowledge that are sufficient to infer the backward induction solution are also those which are necessary to infer it.* Higher levels of knowledge make the theory of the game inconsistent at some information set.

More formally, if we have n players, and some propositions $p_1, \ldots, p_m$, we can construct a knowledge language L by closing under the standard truth-functional connectives and the rule that says that if p is a formula of L, then so is $K_i p$, (i = 1, ... , n), where $K_i p$ stands for 'i knows that p'. Since we are interested in modeling collective knowledge, we add the group-knowledge operator E_G, where $E_G p$ stands for 'everyone in group G knows that p'. If G = {1, 2, ... , n} $E_G p$ is defined as the conjunction $K_1 p \wedge K_2 p \wedge \ldots \wedge K_n p$. K-level group-knowledge of p can be expressed as $E^k_G p \equiv \bigwedge_{i_j \in G,\, 1 \le j \le k} K_{i1} K_{i2} \ldots\ldots K_{1k} p$.

If p is E^k_G-knowledge for all $k \geq 1$, then we say that p is common knowledge in G, i.e., $C_G p \equiv p \wedge E_G p \wedge E^2_G p \wedge \ldots\ldots \wedge E^m_G p \wedge \ldots\ldots$ $C_G p$ implies all formulas of the form $K_{i1} K_{i2} \ldots\ldots K_{in} p$, where the i_j are members of G, for any finite n, and is equivalent to the infinite conjunction of all such formulas.

In order to reason about knowledge, we must provide a semantics for this language. Following Hintikka (1962), we use a possible-worlds semantics. The main idea is that there is a number of possible worlds at each of which the propositions p_i are stipulated to be true or false, and all the truth functions are computed at each world in the usual way. For example, if w is a possible world, then $p \wedge q$ is true at w iff both p and q are true at w. An individual's state of knowledge corresponds to the extent to which he can tell what world he is in, so that a world is possible relative to an individual i. In a given world one can associate with each individual a set of worlds that, given what she knows, could possibly be the real world. Two worlds w and w' are equivalent to individual i iff they create the same evidence for i. Then we can say that an individual i knows a fact p iff p is true at all worlds that i considers possible, i.e., $K_i p$ is true at w iff p is true at every world w' which is equivalent to w for individual i. An individual i does not know p iff there is at least one world that i considers possible where p does not hold.

The following set of axioms and inference rules provides a complete axiomatization for the notion of knowledge we use

A1 : All instances of tautologies
A2 : $K_i p \Rightarrow p$
A3 : $(K_i p \wedge K_i(p \Rightarrow q)) \Rightarrow K_i q$
A4 : $K_i p \Rightarrow K_i K_i p$
A5 : $\sim K_i p \Rightarrow K_i \sim K_i p$
MP : If p and $p \Rightarrow q$, then q
KG : If $\vdash p$, then $\vdash K_i p$

Some remarks are in order. A_2 tells that if i knows p, then p is true. A_3 says that i knows all the logical consequences of his knowledge. This assumption is defensible considering that we are dealing with a very elementary (decidable) logical system. A_4 says that knowing p implies that one knows that one knows p. Intuitively, we can imagine providing an individual i with a database. Then i can look at her database and see what is in it, so that if she knows p, then she knows that she knows it. A_5 is more controversial, since it says that not knowing implies that one knows that one does not know. This axiom can be interpreted as follows: individual i can look at her database to see what she does not know, so if she doesn't know p, she knows that she does not know it. Rule KG says that if a formula p is provable in the axiom system A_1 -A_5, then it is provable that $K_i p$. A formula is provable in an axiom system if it is an instance of one of the axiom schemas, or if it follows from one of the axioms by one of the inference rules MP or KG. Also, a formula p is consistent if $\sim p$ is not provable.

It is easy to verify that the rule KG makes all provable formulas in the axiom system A_1-A_5 *common knowledge* among the players. Suppose q is a theorem, then by KG it is a theorem that $K_i q$ (i = 1, ... , n). If $K_i q$ is a theorem, then it is a theorem that $K_j K_i q$ (for all $j \neq i$), and it is also a theorem that $K_i K_j K_i q$, and so on. In the system A_1-A_5, if $\vdash p$ then $\vdash Cp$. We call the class of axioms A_1-A_5 *general axioms*.

Beside logical axioms, a theory of the game will include game-theoretic solution axioms, behavioral axioms, and axioms describing the information possessed by the players. This second class of axioms we call *special axioms*.

Let us consider as an example game Γ^1:

A6 : The players are rational (i.e., $R_1 \wedge R_2$)
A7 : At node I^{11}, $(r_1 \vee l_1) \wedge \sim(r_1 \wedge l_1)$
A8 : At node I^{21}, $(L \vee R) \wedge \sim (L \wedge R)$
A9 : At node I^{12}, $(r_2 \vee l_2) \wedge \sim(r_2 \wedge l_2)$
A10 : $\pi^1(l_1) = 1$, $\pi^2(l_1) = 0$
A11 : $\pi^1(L) = 0$, $\pi^2(L) = 2$
A12 : $\pi^1(r_2) = 2$, $\pi^2(r_2) = 4$
A13 : $\pi^1(l_2) = 3$, $\pi^2(l_2) = 1$
A14 : At node I^{12}, $R_1 \Rightarrow l_2$
A15 : At node I^{21}, $[R_2 \wedge K_2 R_1] \Rightarrow L$
A16 : At node I^{11}, $[R_1 \wedge K_1 (R_2 \wedge K_2 R_1)] \Rightarrow l_1$
A17: E^2_G (A_6 -A_{16})

A_6 is a behavioral axiom: it tells that the players are rational in the sense of being expected utility maximizers. A_7 -A_9 specify the choices available to each player at each of his information sets, and say that a player can choose only one action. A_{10} -A_{13} specify players' payoffs. A_{14} -A_{16} are solution axioms, and specify what the players should do at any of their information sets if they are rational and know a) that the next player to move is rational and b) what the next player to move knows. A_{17} says that each player knows that each player knows A_6-A_{16}. We call these axioms 'special' since, even if every player knows that every player knows the axioms A_6 -A_{16}, no common knowledge is assumed.

From A_1- A_{17}, the players are able to infer the equilibrium solution l_1. To verify that this level of knowledge is compatible with a deviation from equilibrium, consider in turn the reasoning of both players. In order to decide which strategy to play, player 1 must predict how player 2 would respond to his playing r_1. The main stages of 1's reasoning can be thus described:

r_1	1
By assumption	
$K_1 K_2 ([R_1 \wedge K_1 (R_2 \wedge K_2 R_1)] \Rightarrow l_1)$	2
By axioms A_{16}, A_{17}	
$K_1 K_2 (\sim l_1 \Rightarrow \sim [R_1 \wedge K_1 (R_2 \wedge K_2 R_1)])$	3
By 1, 2, A_1, KG	
$K_1 K_2 (r_1 \vee l_1) \wedge \sim(r_1 \wedge l_1)$	4
By A_{17}, A_7	
$K_1 K_2 (\pi^1(l_1) = 1, \pi^2(l_1) = 0)$	5
By A_{17}, A_{10}	
$K_1 (R_2 \wedge K_2 R_1)$	6
By A_6, A_{17}	
$K_1 \sim K_1 K_2 (K_1 (R_2 \wedge K_2 R_1))$	7
By A_5, 3, A_{17}	

For all that player 1 knows, his playing r_1 can be 'explained away' by player 2 as due to $\sim K_1(R_2 \wedge K_2R_1)$. In other words, what player 1 knows of player 2 does not conflict with his knowledge that $K_2R_1 1$. Since

$K_1 [R_2 \wedge K_2R_1] \Rightarrow L$ 8
By A_{17}, A_{15}

player 1 knows that 2 will respond with L to r_1, hence he plays l_1.

What would player 2 think facing a deviation on the part of player 1?

r_1 1
By assumption
$K_2 (r_1 \vee l_1) \wedge \sim(r_1 \wedge l_1)$ 2
By A_{17}, A_7
$K_2 (\pi^1(l_1) = 1, \pi^2(l_1) = 0)$ 3
By A_{17}, A_{10}
$K_2 (\sim l_1 \Rightarrow \sim [R_1 \wedge K_1 (R_2 \wedge K_2R_1)])$ 4
By A_{17}, A_{16}, A_1
$K_2 (R_1 \wedge K_1R_2)$ 5
By A_{17}, A_6
$K_2 (\sim l_1 \Rightarrow \sim K_1 K_2R_1)$ 6
By 4, 5

player 2 can 'explain' why r_1 was played, and since this explanation does not conflict with K_2R_1, she will choose strategy L.

What would happen if further levels of knowledge were added? Suppose the following axiom is added to the theory

A_{18} : $E^2_G (A_6 - A_{17})$

Since there is one more level of knowledge, now both players know that $K_1K_2R_1$ and $K_2K_1R_2$ obtain. This level of information implies that — were r_1 to be played — player 2 would face an inconsistency. As before,

$K_2 (\sim l_1 \Rightarrow \sim [R_1 \wedge K_1 (R_2 \wedge K_2R_1)])$ 1
By A_1, A_{16}, A_{18}
$K_2 [R_1 \wedge K_1 (R_2 \wedge K_2R_1)]$ 2
By A_6, A_{18}
$K_2 l_1$ 3
By 2, A_{16}
r_1 4
By assumption
$K_2 (r_1 \vee l_1) \wedge \sim(r_1 \wedge l_1)$ 5
By A_7, A_{18}
$K_2 \sim [R_1 \wedge K_1 (R_2 \wedge K_2R_1)]$ 6
By 1, 4
$[R_1 \wedge K_1 (R_2 \wedge K_2R_1)]$ 7
By A_2
$\sim [R_1 \wedge K_1 (R_2 \wedge K_2R_1)]$ 8
By A_2

Since the conjunction of the formulas 7 and 8 is false, and in classical logic one can deduce anything from a false statement, player 2 can use this conjunction to construct a proof that "r_1". Adding axiom A_{18} makes the theory of the game *inconsistent for player* 2, therefore 2 is unable to use it to predict how player 1 would respond if she were to play R. Which leaves 2 uncertain as to how to play herself.

Is the theory of the game also inconsistent for player 1? It is easy to verify that the state of information of player 1 does not let him realize that — were he to play r_1— player 2 would face an inconsistency. By A_{18}, player 1 knows $K_2K_1R_2$. But the levels of knowledge assumed in A_{18} do not let 1 know that K_2 $(K_1K_2R_1)$. Therefore player 1 can believe that 2 will explain a deviation by assuming $\sim (K_1K_2R_1)$. If so, he can predict that 2's response will be L, which makes him play l_1. Hence a theory of the game that includes axiom A_{18} supports the backward induction solution.

The backward induction equilibrium cannot be inferred only in the case in which $K_1(K_2K_1K_2R_1)$ obtains. This level of knowledge is brought forth by the additional axiom

A_{19} : E^2_G $(A_6$ -$A_{18})$

In this case player 1 would know that playing r_1 makes the theory of the game inconsistent for player 2 at I^{21}. If so, player 2 would be unable to predict what would happen were she to play R and 1, knowing that, would be unable to predict what would happen were he to play r_1.

Since a solution concept for the class of games we are examining depends upon the levels of knowledge possessed by the players, we have to introduce a few new definitions:

Definition 3.1. A *knowledge-dependent game* is a quadruple $\Gamma = (N, S^i, K^i, \pi^i)$ where $N = \{1, \ldots, n\}$ is the number of players; S^i is the set of strategies of player i; K^i is the knowledge possessed by player i and is defined as the union of what i knows at each of his information sets, i.e., $K^i = \bigcup_{1 \le j \le k} K^i I^{ij}$; π^i is player i's payoff.

Definition 3. 2. An n-tuple of strategies $(s^1, \ldots, s^n)$ is a *knowledge-consistent* play of a knowledge-dependent game if, for each player i, every choice s^{ij} that strategy s^i recommends at each information set $I^{ij} \in P^i$ satisfies the following conditions: (i) reaching I^{ij} is compatible with K^i and (ii) it can be proven from K^i that s^{ij} is a best reply for player i at I^{ij}.

Theorem 3.1. For every finite, extensive form game of perfect and complete information, the backward induction equilibrium is a knowledge-consistent play of some knowledge-dependent game and, conversely, every knowledge-consistent play of a knowledge-dependent game is a backward induction equilibrium.

Proof. The first part of the proof is trivial, since Theorem 2. 2 illustrates a specification of the knowledge of each player that makes the backward induction equilibrium a knowledge-consistent play. The second part of the theorem can be proven by induction on the number of moves in the game. Suppose the game has only one move. In order to make a choice, the player who has to move must know his available strategies and payoffs. A rational player knows that he should choose that branch which leads to a terminal node with the maximum payoff to him. Then if the player knows his strategies and payoffs, he can infer his payoff-maximizing solution, which is the backward induction solution. Assume the theorem is true for all games involving at most K moves (some $K \ge 1$). Then it follows that the knowledge-consistent play $(s^1, \ldots, s^n)$, restricted to any of the subgames of Γ having no more than K moves, corresponds to the backward induction solution for that subgame. Let Γ be a knowledge-dependent game with K+1 moves and

let r be the root of the game tree T for Γ. At information set I^{ir} there is a recommendation of play s^{ir} for player i that can be inferred from K^i. Let $K = \bigcup_{1 \leq m \leq k} K_j^{jr+m}$ be the union of the knowledge possessed by each player j which has to play at an information set that immediately follows I^{ir}. Then player i's knowledge of K implies the choice of the move that is the backward induction solution at I^{ir}. Therefore the union of K^i and K allows one to derive both the backward induction solution for I^{ir} and the strategy s^{ir}. The two must coincide since the union of K^i and K cannot lead to an inconsistent system.

Notes

[1]Recent attempts to analyze and model the players' reasoning process that leads to the selection of an equilibrium include Harsanyi's 'tracing procedure' (Harsanyi 1977), Skyrms' 'deliberational dynamics' (Skyrms 1986), Harper's application of the notion of 'ratifiable choice' to games (Harper 1988) and models of counterfactual reasoning in games (Shin 1987; Bicchieri 1988). Other studies of players' reasoning that focus on internal consistency of beliefs have led to the notion of 'rationalizability' (Bernheim 1984; Pearce 1984).

[2]The iterative notion of common knowledge was introduced by Lewis (1969), and a different definition, based on the notion of knowledge partition, was applied to game theory by Aumann (1976). Tan and Werlang (1986) have shown the equivalence of the two notions.

[3]Bayesian game theory has the same problem: the players' incomplete information about the structure of the game is simply *described* in the form of an extensive form game with chance moves (Harsanyi 1967, 1968). In this case, too, some basic assumptions of the theory are not treated as part of the theory.

[4]More recently, Gilboa and Schmeidler (1988) proved that in information-dependent games a common knowledge axiom is inconsistent with a rationality axiom.

[5]I have shown elsewhere (Bicchieri 1988) that the various refinements of Nash equilibrium can be uniformly treated as different rules for belief change, and that such rules can be inferred from a richer theory of the game that includes epistemic criteria that allow an ordering of the rules in terms of epistemic importance. In the class of games I am considering, a theory of the game that contains a model of belief change would *always* let the players 'explain away' any deviation from equilibrium (Bicchieri 1988a).

[6]If the players were endowed with a model of belief-change (Bicchieri 1988, 1988a), there would be other hypotheses beside Selten's that make common knowledge of rationality compatible with out of equilibrium behavior.

References

Aumann, R. J. (1976), "Agreeing to disagree", *The Annals of Statistics* 4: 1236-1239.

Bacharach, M. (1987), "A theory of rational decision in games", *Erkenntnis* 27: 17-55.

_ _ _ _ _ _ _ _. (1985), "Some extensions of a claim of Aumann in an axiomatic model of knowledge", Journal of Economic Theory 37: 167-55.

D. Bernheim (1984), "Rationalizable strategic behavior", *Econometrica* 52: 1007-1028.

Bicchieri, C. (1988), "Strategic behavior and counterfactuals", *Synthese* 76: 135-169.

_____. (1988a), "Common knowledge and backward induction: a solution to the paradox", in M. Vardi (ed.) *Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann Publishers, Los Altos.

_____. (1989), "Self-refuting theories of strategic interaction: a paradox of common knowledge", *Erkenntnis* 30: 69-85.

Binmore, K. (1987), "Modeling rational players I", *Economics and Philosophy* 3: 179-214.

_____. (1988), "Modeling rational players II", *Economics and Philosophy* 4: 9-55.

_____ and A. Brandenburger (forthcoming), "Common knowledge and game theory", *Journal of Economic Perspectives.*

Bonanno, G. (1987), "The logic of rational play in extensive games", Disc. Paper no. 16, Nuffield College, Oxford.

A. Brandenburger (forthcoming), "The role of common knowledge assumptions in game theory", in F. Hahn (ed.) *The Economics of Information, Games, and Missing Markets*, Cambridge University Press.

_____. and Dekel, E. (1985a), "Common knowledge with probability", Research Paper no. 796R, Graduate School of Business, Stanford University.

_____. (1985b), "Hierarchies of beliefs and common knowledge", Research Paper no. 841, Graduate School of Business, Stanford University.

Gilboa, I. (1986), "Information and meta-information", Working paper no. 30-86, Tel-Aviv University.

_____ and D. Schmeidler (1988), "Information dependent games", *Economics Letters* 27: 215-221.

Halpern, J. and Fagin, R. (1988), *Modelling knowledge and action in distributed systems*. Technical Report, IBM.

_____. and Moses, Y. (1987), "Knowledge and common knowledge in a distributed environment", IBM Research Report RJ 4421.

Harper, W. (1988), "Causal decision theory and game theory", in Harper and Skyrms (eds.), *Causation in Decision, Belief Change and Statistics*, Reidel.

Harsanyi, J. (1967-68), "Games with incomplete information played by 'Bayesian' players", Parts I, II, and III. *Management Science* 14: 159-182, 320-332, 468-502.

_____. (1975), "The tracing procedure: a Bayesian approach to defining a solution for n-person non-cooperative games", *International Journal of Game Theory* 4: 61-94.

_____. and R. Selten (1988), *A General Theory of Equilibrium Selection in Games*. The MIT Press, Cambridge.

Hintikka, J. (1962), *Knowledge and Belief*. Cornell University Press, Cornell.

Kaneko, M. (1987), "Structural common knowledge and factual common knowledge", RUEE Working Paper no. 87-27, Hitotsubashi University.

Kuhn, H.W. (1953), "Extensive games and the problem of information", in H.W. Kuhn and A.W. Tucker (eds.) *Contributions to the Theory of Games*. Princeton University Press, Princeton.

Lenzen, W. (1978), "Recent work in epistemic logic", *Acta Philosophica Fennica* 30: 1-219.

Lewis, D. (1969), *Convention*. Harvard University Press, Cambridge.

Luce, R. and Raiffa, H. (1957), *Games and Decisions*. Wiley, New York.

Mertens, J.-F. and Zamir, S. (1985), "Formulation of Bayesian analysis for games with incomplete information", *International Journal of Game Theory* 14: 1-29.

Parikh, R. and Ramanujam, R. (1985), "Distributed processes and the logic of knowledge", *Proceedings of the Workshop on Logics of Programs*: 256-268.

Pearce, D. (1984), "Rationalizable strategic behavior and the problem of perfection", *Econometrica* 52: 1029-1050.

Reny, P. (1987), "Rationality, common knowledge, and the theory of games", Working paper, Department of Economics, University of Western Ontario.

Samet, D. (1987), "Ignoring ignorance and agreeing to disagree", mimeo, Northwestern University.

Selten, R. (1975), " Re-examination of the perfectness concept for equilibrium points in extensive games", *International Journal of Game Theory* 4: 22-55.

Shin, H.S. (1987), "Counterfactuals, common knowledge and equilibrium", mimeo, Nuffield College, Oxford.

Skyrms, B. (1989), "Deliberational dynamics and the foundations of Bayesian game theory", in J. E. Tomberlin (ed.) *Epistemology*. Ridgeview, Northridge.

_____. (1986), "Deliberational equilibria", *Topoi* 1.

Tan, T. and Werlang, S. (1986), "On Aumann's notion of common knowledge—an alternative approach", Working paper no. 85-26, University of Chicago.

Van Damme, E.E.C. (1983), *Refinements of the Nash Equilibrium Concept*. Springer-Verlag, Berlin.