ELSEVIER Journal of Mathematical Economics 27 (1997) 325–345

JOURNAL OF Mathematical ECONOMICS

A necessary and sufficient epistemic condition for playing backward induction

Dieter Balkenborg [a,b,*], Eyal Winter [c,d]

[a] *University of Bonn, Bonn, D-53113, Germany*
[b] *University of Southampton, Southampton, SO17 1BJ, UK*
[c] *The Hebrew University of Jerusalem, Jerusalem, Israel*
[d] *Washington University in St. Louis, St. Louis, MO 63130-4899, USA*

Submitted April 1995; accepted February 1996

Abstract

In an epistemic framework due to Aumann we characterize the condition on the rationality of the players that is both necessary and sufficient to imply backward induction in perfect information games in agent form. This condition requires each player to know that the players are rational at later, but not at previous, decision nodes.

JEL classification: C72; D81

Keywords: Extensive games; Backward induction; Subgame perfection; Common knowledge

1. Introduction

The notion of a subgame perfect equilibrium is one of the most important tools in economic applications of game theory; for instance, in the modern theory of industrial organization.

In a finite perfect information game, subgame perfection coincides with the principle of backward induction. Backward induction is the most natural and compelling solution concept for such a game. It yields a unique solution when the game is in 'generic position' and it seems to be based on comparatively weak

* Corresponding author: e-mail: dieter@glider.econ3.uni-bonn.de

PII S0304-4068(96)00776-8

rationality requirements, which do not even invoke the concept of a Nash equilibrium. In perfect information games, backward induction seems to be almost tantamount to a principle that we will call 'forward knowledge of rationality'. The principle requires that the players act rationally at the last decision nodes, that this is known to the players at the second-last decision nodes, that the players act rationally at the second-last decision nodes, and so on. Forward knowledge of rationality requires knowledge of rationality only with respect to succeeding decision nodes, not with respect to preceding ones. While forward knowledge of rationality is a *comparatively* weak rationality condition, it is not an innocuous assumption: a chain of knowledge about the rationality of others as long as the game tree is required.

But can backward induction and forward knowledge of rationality really be identified in perfect information games in generic position? Our objective in this paper is to address this question in a formal model of knowledge. We show that forward knowledge of rationality is indeed the *only* rationality condition (i.e. the only condition that can be stated solely in terms of knowledge and rationality of the players) that is both *necessary and sufficient* to deduce backward induction. We show that any other rationality condition either does not imply backward induction or is logically weaker than forward knowledge of rationality. But, as we will see, some qualification is needed: A rationality condition has to imply forward knowledge of rationality only if it implies backward induction in a sufficiently large class of perfect information games. Counterexamples show that this class may have to be strictly larger than the class of games with a fixed game tree.

Establishing necessary rationality conditions for playing backward induction is of special importance in explaining strategic behavior not consistent with backward induction. By weakening the rationality conditions that imply backward induction we can describe with greater accuracy the possible reasons why backward induction might fail and hence how strong the rationality requirements for backward induction really are. Our condition implies that if backward induction fails, then it must be true that either the first moving player does not act rationally or that he does not know whether all players will act rationally at later decision nodes. Stronger rationality conditions (for instance, common knowledge of rationality, see below) do not have to imply this seemingly obvious conclusion.

Despite the extensive use of subgame perfection in the game-theoretic literature, the logic of subgame perfection and backward induction in perfect information games has been seriously challenged both by experimentalists and by theorists. Several experimental studies have pointed out that real-life strategic behavior is often inconsistent with the principle of backward induction even in extremely simple perfect information games such as the ultimatum bargaining game (see Güth et al., 1982).

On a more theoretical level, the intuition behind backward induction was questioned in examples such as the chain-store paradox (Selten, 1978), and the Centipede game (Rosenthal, 1981), where the backward induction outcome seems implausible. In response, several authors (e.g. Kreps and Wilson, 1982; Milgrom and Roberts, 1982; and Fudenberg et al., 1988) argued that the perfect information game itself might not be the right model to consider. They consider games that are, in a certain sense, close to the original perfect information game, but where the payoff structure is not common knowledge. They show that these nearby games may have Nash equilibria satisfying the strongest refinement criteria that are qualitatively very different from the backward induction solution.

In another approach, Aumann (1992) pointed out that a small lack of common knowledge of rationality (and, actually, of forward knowledge of rationality) might give rise to outcomes very different from the backward induction outcome.

Yet another strand of literature (see Aumann, 1995, for detailed references) questioned directly the logic underlying backward induction. This literature argues that the assumptions on the mutual knowledge of rationality, which are required for deducing that backward induction will be played, are self-contradictory.

In response to the latter literature, Aumann (1995) formulated an elegant model of knowledge and established sufficient epistemic conditions on rationality that imply backward induction in perfect information games in generic position. Aumann shows that common knowledge of rationality implies backward induction. Thus, Aumann's condition requires that all players are rational, that every one of them knows that all players are rational, that every one of them knows that everyone knows that all players are rational, and so on. [1]

We will use the model and the methods introduced by Aumann. In his framework we show that backward induction is implied by a condition that is intrinsically weaker than common knowledge of rationality and we identify the weakest such condition.

Results similar to ours might still hold in frameworks where – unlike in Aumann's model – common knowledge of rationality would be inconsistent. One might expect this since, first of all, forward knowledge of rationality is weaker than common knowledge of rationality. Secondly, the contradictions that several authors obtained with common knowledge of rationality seem to be caused by the knowledge of the players about the rationality of other players at past decision nodes. They might not occur with forward knowledge of rationality, which refers only to the player's knowledge of rationality at future decision nodes. We conjecture, for instance, that Samet's (1993) notion of 'common hypothesis of rationality' is necessary and sufficient for backward induction in his model of hypothetical knowledge.

Aumann's approach, which we are using, implicitly assumes that the game structure, and in particular the payoff structure, are common knowledge. The work by Kreps and Wilson (1982), Milgrom and Roberts (1982) and Fudenberg et al.

[1] Aumann uses the semantic approach to model knowledge. Kramerz (1993) obtains related results in a syntactic approach.

(1988) referred to above has shown that this is a very strong assumption. As much as common knowledge of rationality is not needed to deduce backward induction, common knowledge of the payoffs is not needed either. In a slight extension of the framework used here in which the payoffs can vary with the state, it can be shown that forward knowledge of rationality *and payoffs* implies backward induction. We do not know what the *minimal* requirements on the knowledge of rationality and payoffs would be in order to deduce backward induction, but the techniques we are using in this paper should be helpful in addressing this question, too.

In Section 2 we review Aumann's (1995) epistemic framework and introduce the notion of forward knowledge of rationality. We will assume throughout that each player has only one decision node. As discussed in Subsection 5.3, this is not a serious restriction of our analysis. We also explain in Section 2 what we mean by a minimal (i.e. a necessary and sufficient) rationality condition for backward induction. Our notion of minimality refers to a class of perfect information games that can be strictly larger than the class of games with a fixed game tree. We illustrate in Subsection 5.1 why this is necessary.

The argument as to why forward knowledge of rationality is indeed the minimal rationality condition for backward induction is illustrated in Section 3 for games with two players and two decision nodes. The general proof is given in Section 4. Section 5 contains extensions and additional remarks. In the appendix we explain in more detail how our notion of minimality relates to logical deduction in epistemic logic.

2. The model

Following Aumann (1995) we restrict ourselves to finite perfect information games in generic position. Moreover, we will assume that every player has only one decision node. A game is said to be in generic position if no player receives the same payoff at any two terminal nodes. Henceforth, we will refer by the word 'game' to any perfect information game with the properties just mentioned.

Let N be a finite set of players. We consider a partial order $\succ$ on the set of players that is induced by a tree with a root whose nodes are the players in N; that is, we write $j \succ i$ when player j follows player i in that tree. The player r at the origin of the tree is called the *root player*. Player j is an *immediate successor* of player i if $j \succ i$ and there is no player j' with $j \succ j' \succ i$. Let $S(i)$ denote the set of immediate successors of player i.

We will say that a game Γ has the *order of moves* $(N, \succ)$ if it has the set of players N and if, whenever $j \succ i$ for $i, j \in N$, j comes after i in Γ. That is, each path through j's decision node also goes through i's decision node. The order of moves defines an equivalence relation on the set of all games. Two games with the same game tree and the same assignment of players to decision nodes always have the same order of moves. But games with the same order of moves can have different payoff functions and different numbers of terminal nodes.

For a given game Γ let A_i be the set of actions available to player i and let $A := \prod_{i\in N} A_i$ be the set of combinations of actions for the players. The *backward induction solution* $b = (b_1, \ldots, b_n) \in A$ is defined inductively as follows. If there is no player j with $j \succ i$, then b_i is simply i's optimal action conditional on his decision node being reached. Otherwise, b_i is inductively defined as the action that, conditional on his decision node being reached, maximizes player i's payoff, given that each player j with $j \succ i$ plays b_j.

We now describe the model in which epistemic conditions for playing backward induction can be given.

A *knowledge space* is a pair $(\Omega, \mathscr{P})$, where Ω is a set of states of the world and $\mathscr{P} = (\mathscr{P}_i)_{i\in N}$ are partitions of Ω for each player i in N, called *knowledge partitions*. $\mathscr{P}_i(\omega)$ denotes the set of player i's knowledge partition that contains the state $\omega \in \Omega$. When a state of the world ω is realized, each player i knows that the event $\mathscr{P}_i(\omega)$ has happened. For an arbitrary event E the event 'i knows E', denoted by $K_i(E)$, is the union of all elements of $\mathscr{P}_i$ that are included in E. The event 'E is common knowledge' is defined by

$$CK(E) = E \cap K(E) \cap K(K(E)) \cap K(K(K(E))) \cap \ldots,$$

where $K(E) := \bigcap_{i\in N} K_i(E)$ for any event E.

A *knowledge system* (or simply a *model*) is a quadruple $\psi = (\Omega, \mathscr{P}, \Gamma, f)$, where $(\Omega, \mathscr{P})$ is a knowledge space, Γ is a game and $f = \{f_i\}_{i\in N}$ assigns to each state of the world ω a combination of actions for all the players. [2] We assume that the map $f_i : \Omega \to A_i$ is *measurable* with respect to player i's knowledge partition, i.e. $f_i(\omega') = f_i(\omega)$ holds for all $\omega' \in \mathscr{P}_i(\omega)$. We call a model ψ *compatible* with the order of moves $(N, \succ)$ if it is based on a game Γ with the order of moves $(N, \succ)$.

We use the notion of 'conditional rationality' introduced in Aumann (1995), i.e. player i's rationality is defined as the set of all states of the world in which i does not know of any action that will yield him a conditional payoff higher than the one obtained by playing f_i. To define this formally, it is convenient to make use of the following notation. Given a statement θ, we denote by $[\theta]$ the event that consists of all elements in Ω for which θ holds true. For example, the event $[f_i = a_i]$ is the set of all states of the world ω where $f_i(\omega) = a_i$. The event 'player i is rational' is now defined as

$$R_i = \bigcap_{a_i\in A_i} \neg K_i\left[h_i(a_i, f^{i}) > h_i(f_i, f^{i})\right],$$

[2] We will say that a player takes action a_i at ω if $f^i(\omega) = a_i$. We will do so in our verbal discussions even if the decision node of the player is not reached in the state ω. Krumerz (1993) shows that in a formal semantic approach it is important to distinguish carefully between taking an action (if the decision node is reached) and *planning* to take an action (regardless of whether the decision node is reached or not).

where h_i denotes player i's payoff function for the subgame starting with player i's move, $\neg$ denotes complementation in Ω and $f^i := \{f_j\}_{j \succ i}$. We set $R := \bigcap_{i \in N} R_i$.

The event '*player i plays backward induction*' is written $I_i := [f_i = b_i]$. The event '*the backward induction solution is played*' is simply $I := \bigcap_{i \in N}[f_i = b_i]$. Given a strategy combination $a \in A$ in Γ, let z_a be the end point reached in Γ when a is played. Thus z_f is a function from Ω to the set Z of all end points of the game. Finally, the event '*the backward induction outcome is reached*' is given by $IP = [z_b = z_f]$.

2.1. Forward knowledge of rationality

In Aumann (1995) it is shown that common knowledge of rationality implies backward induction, i.e. $CK(R) \subseteq I$.

Our sufficient condition is based on the idea that to deduce backward induction it is enough that knowledge of rationality 'flows' in only one direction. More specifically, each player's knowledge of rationality applies only to his successors, and he need not know anything about the rationality of those playing before him. Formally, the event '*forward knowledge of rationality*' is defined inductively as follows.

For each player i we define the event Q_i: if i has no successor, then $Q_i := R_i$. Furthermore, if i has a successor, then Q_i is inductively defined by

$$Q_i := R_i; \cap K_i\left(\bigcup_{j \in S(i)} Q_j\right).$$

The event 'forward knowledge of rationality' is now given by $Q := Q_r$, where r is the root player.

Note that while the formula for common knowledge of rationality depends only on the set of players N, forward knowledge of rationality also depends on the order $\succ$.

Aumann's result can now be strengthened as follows.

Theorem 2.1. Forward knowledge of rationality implies backward induction, i.e. $Q \subseteq I$.

The proof is a straightforward adaptation of Aumann's (1995) proof and given in Section 4.

2.2. The notion of minimality

To define our notion of a minimal epistemic condition for backward induction, we will regard all epistemic expressions (such as I, R_i, Q, etc.) as operators that assign to each model ψ an event in the relevant set of states of the world. Thus, an epistemic expression F without a model is merely a string of formal symbols that becomes an event whenever we assign it to a model ψ. From now on, we write $I(\psi)$, $R_i(\psi)$, $Q^f(\psi)$, etc. for the events determined by the epistemic expressions I, R_i, Q, etc. in a given model ψ.

An epistemic expression F is said to be an (*epistemic*) *rationality condition* if it can be obtained from the expressions $\{R_i\}_{i \in N}$ by means of the regular set theory operators, i.e. $\cup$ (union), $\cap$ (intersection) and $\neg$ (complementation), and the knowledge operators K_i. (The expression R_i will only be applied to the event that player i is rational.) Observe that the expression 'the backwards induction is played' $I(\psi)$ is *not* a rationality condition.

A rationality condition F is said to be a *minimal rationality condition for backward induction* with respect to the order of moves $(N, \succ)$ if:

(1) $F(\psi) \subseteq I(\psi)$ for every model ψ compatible with the order of moves $(N, \succ)$ (i.e. F implies backward induction for a sufficiently rich class of games).

(2) for any other rationality condition F' satisfying Condition 1, we have $F'(\psi) \subseteq F(\psi)$ for every model ψ compatible with the order of moves $(N, \succ)$ (i.e. F' implies F in every model ψ).

If F satisfies only Condition 1, then we call it a *rationality condition for backward induction* with respect to the order of moves $(N, \succ)$.

A minimal rationality condition for backward induction is thus *sufficient* for playing backward induction. It is *necessary* for playing backward induction insofar as any other rationality condition for backward induction 'implies' the minimal condition.

Theorem 2.2. For a given order of moves $(N, \succ)$ there exists a unique minimal rationality condition for backward induction. This condition is the forward knowledge of rationality.

The result follows from Theorem 2.1 and Proposition 4.1 in Section 4. To obtain the result that forward knowledge of rationality is the *minimal* rationality condition for backward induction, it is important that a rationality condition for backward induction is defined with respect to all games with the same order of moves and not with respect to some smaller class of games. This will be illustrated in Subsection 5.1.

3. Illustration of the result

This section illustrates our result for the simplest interesting class of games. We consider the class of games with just two players, and two decision nodes, where, without loss of generality, player 1 moves first. Forward knowledge of rationality then means that player 1 is rational and that he knows (and hence, that it is indeed true) that player 2 is rational. In short

$$R_1 \cap K_1 R_2 .$$

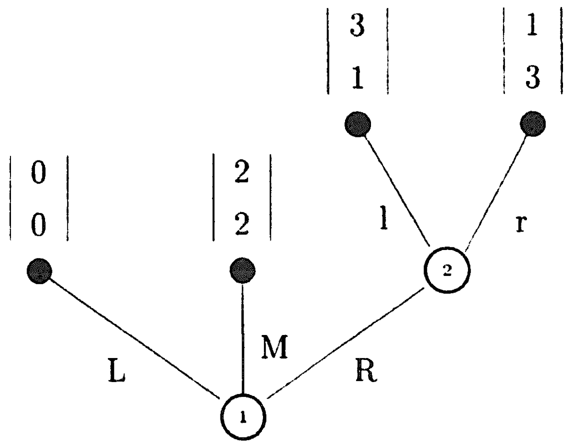

Fig. 1.

This rationality condition implies backward induction. Intuitively, player 2 is rational if and only if he plays his backward induction choice. If player 1 knows that player 2 takes the backward induction choice but player 1 would himself take an action different from his backward induction choice then player 1 would know that there is an action different from the one he chooses which must yield a higher payoff. A rational player 1 therefore has to choose his backward induction choice in this situation. The formalization of this argument is given in Aumann (1995).

Of course, forward knowledge of rationality is strictly weaker than common knowledge of rationality. Consider, for instance, the following two-state model ψ based on the game in Fig. 1. Player 2 takes his action 'right' in both states and does not know in which of the two states he is. Player 1 knows the state. In one his choice is 'left' and in the other his choice is 'right'. In both states player 2 does not know that player 1 is rational, since he considers it possible that player 1 takes his left, strictly dominated, choice. Therefore, common knowledge of rationality yields the empty set. The event 'forward knowledge of rationality' is, however, not empty since it contains the state where player 1 takes his right choice.

Our main concern is to show that any rationality condition for backward induction F implies forward knowledge of rationality in every model ψ, i.e.

$$F(\psi) \subseteq (R_1 \cap K_1 R_2)(\psi).$$

The crucial observation for proving this is the fact that if we alter the model ψ by changing Γ and $\{f_i\}_{i \in \{1,2\}}$, so that $R_1(\psi)$ and $R_2(\psi)$ remain the same, then $F(\psi)$ will not change either.

Now let $(\Omega, \mathscr{P}_1, \mathscr{P}_2)$ be the knowledge space of ψ and let $E_i := R_i(\psi)$ be the event where player $i \in \{1, 2\}$ is rational.

We construct a new model $\psi' = (\Omega, \mathscr{P}_1, \mathscr{P}_2, \Gamma, f_1', f_2')$ on the knowledge space $(\Omega, \mathscr{P}_1, \mathscr{P}_2)$, which is again based on the game Γ in Fig. 1 with the maps f_1', f_2' defined as follows:

Player 2 chooses 'right' at each state in E_2 and 'left' at each state in $\neg E_2$.

Player 1 chooses 'middle' in $E_1 \cap K_1 E_2$, 'right' in $E_1 \cap \neg K_1 E_2$ and 'left' in $\neg E_1$.

The important properties of the new model ψ' are:

(1) Players 1 and 2 are rational exactly in the events E_1, E_2, respectively, i.e. $R_i(\psi') = E_i$ for $i \in \{1, 2\}$.

(2) The event 'the backward induction solution is played' is *identical* to the event 'forward knowledge of rationality', i.e.

$$I(\psi') = (R_1 \cap K_1 R_2)(\psi').$$

By assumption, F implies backward induction for all models based on the game in Fig. 1. Thus, $F(\psi') \subseteq I(\psi')$ and hence

$$F(\psi') \subseteq (R_1 \cap K_1 R_2)(\psi').$$

Since both F and $R_1 \cap K_1 R_2$ are rationality conditions, we conclude for the original model ψ:

$$F(\psi) \subseteq (R_1 \cap K_1 R_2)(\psi).$$

Thus, $R_1 \cap K_1 R_2$ is indeed a minimal rationality condition for backward induction according to our definition.

We now turn to the proof of the general case.

4. Proofs

Theorem 2.1. Forward knowledge of rationality implies backward induction, i.e.

$$Q(\psi) \subseteq I(\psi),$$

for all models ψ based on games with a given order of moves $(N, \succ)$.

Proof. For the sake of simplicity we drop ψ from the formulas throughout the proof, as we did prior to Subsection 2.2. It will be shown by induction that

$$Q_i \subseteq \bigcup_{j \succeq i} I_i, \tag{4.1}$$

for all $i \in N$. Take some $i \in N$ and suppose that (4.1) is satisfied for each immediate successor $j \in S(i)$. Then

$$Q_i = R_i \cap K_i\Big(\bigcap_{j \in S(i)} Q_j\Big) \subseteq R_i \cap K_i\Big(\bigcap_{j \succ i} I_j\Big) = R_i \cap K_i\big[f^i = b^i\big].$$

Now Aumann (1995) shows in the proof of his Theorem A that if i is rational and knows that all his successors are rational then i plays backward induction. Formally,

$$R_i \cap K_i[f^i = b^i] \subseteq I_i.$$

Therefore,

$$Q_i \subseteq R_i \cap K_i[f^i = b^i] \subseteq \bigcap_{j \succcurlyeq i} I_i.$$

□

Proposition 4.1. Let F be a rationality condition with

$$F(\psi) \subseteq I(\psi),$$

for all models ψ compatible with a given order of moves $(N, \succ)$. Then

$$F(\psi) \subseteq Q(\psi),$$

for all models ψ compatible with $(N, \succ)$.

Two further results are needed before we can prove Proposition 4.1.

Define for a model ψ and an action $a_i \in A_i$:

$$R_i(\psi, a_i) = \bigcap_{a'_i \in A_i \setminus \{a_i\}} \neg K_i[h_i(a'_i, f^i) > h_i(a_i, f^i)].$$

$R_i(\psi, a_i)$ is the event where it is rational for player i to take action a_i.

Lemma 4.2. For any model ψ:

$$R_i(\psi) = \bigcup_{a_i \in A_i} [f_i = a_i] \cap R_i(\psi, a_i).$$

Proof. We have

$$R_i(\psi) = \left(\bigcup_{a_i \in A_i} K_i[a_i = f_i]\right) \cap \left(\bigcap_{a'_i \in A_i} \neg K_i[h_i(a'_i, f^i) > h_i(f_i, f^i)]\right)$$

$$= \bigcup_{a_i \in A_i} \left(\bigcap_{a'_i \in A_i} (K_i[a_i = f_i] \cap \neg K_i[h_i(a'_i, f^i) > h_i(f_i, f^i)])\right).$$

As in Aumann (1995), Formula (13) and the preceding formulas in the proof of Theorem A, it follows that

$$K_i[a_i = f_i] \cap \neg K_i[h_i(a'_i, f^i) > h_i(f_i, f^i)]$$

$$= K_i[a_i = f_i] \cap \neg K_i[h_i(a'_i, f^i) > h_i(a_i, f^i)].$$

Using $K_i[a_i = f_i] = [a_i = f_i]$ and $\neg K_i[h_i(a_i, f^i) > h_i(a_i, f^i)] = \Omega$, the claim follows. □

Lemma 4.3. Fix an order of moves $(N, \succ)$. For every knowledge space $(\Omega, (\mathscr{P}_i)_{i\in N})$ and every family of events $\{E_i\}_{i\in N}$ in Ω with $K_i E_i = E_i$ for all $i \in N$ there exists a model

$$\psi = (\Omega_i, (\mathscr{P}_i)_{i\in N}, \Gamma, (f_i)_{i\in N})$$

compatible with the order of moves $(N, \succ)$ on the given knowledge space $(\Omega, (\mathscr{P}_i)_{i\in N})$ for which

$$Q(\psi) = I(\psi) = IP(\psi).$$

and

$$R_i(\psi) = E_i$$

for all $i \in N$. $IP(\psi)$ denotes hereby the event where the backward induction path is played.

Proof. Throughout the proof we work with ordinal preferences instead of payoff functions. Thus, for $i \in N$ and two end points z, z' we write $z \ \mathrm{p}_i \ z'$ for $h_i(z) > h_i(z')$. For the subgame starting with player i's move, a combination of actions of the players in the subgame (including player i) is denoted by $a(i)$ and the induced outcome in the subgame by $z_{a(i)}$.

The proof is by induction. Let $i = 1$ be the root player for the order $\succ$. Let $S(1) = \{2, \ldots, j^*\}$ be the set of immediate successors of player 1. We assume that for every immediate successor $j \in S(1)$ we have found a model,

$$\psi_j = \left(\Omega, (\mathscr{P}_k)_{k\in N_j}, \Gamma_j, (f_k)_{k\in N_j}\right),$$

as follows:

(1) Γ_j is a game with a set of players $N_j := \{k \mid k \succcurlyeq j\}$ compatible with the order induced by $\succ$ on this set.

(2) The knowledge space $(\Omega, (\mathscr{P}_k)_{k\in N_j})$ is obtained from $(\Omega, (\mathscr{P}_k)_{k\in N})$ by omitting all knowledge partitions for the players not in N_j.

(3) ψ_j satisfies

$$Q_j(\psi_j) = I(\psi_j) = IP(\psi_j).$$

and

$$R_k(\psi_j) = E_k,$$

for all players $k \in N_j$.

(The induction assumption consisting of these three conditions is satisfied if $S(1) = \emptyset$, i.e. $N = \{1\}$.)

Let $b(j)$ denote the backward induction solution and let $z_{b(j)}$ denote the backward induction outcome for the game Γ_j.

Consider now the game Γ constructed as follows: Player 1 is the root player. His set of actions is

$$A_1 = \{a_1^-, a_1^+\} \cup \{a_1^j \mid j \in S(1)\},$$

where a_1^- leads to the terminal node z^- and a_1^+ leads to the terminal node z^+. By an action a_1^j, player 1 gives the move to his immediate successor j after which the game Γ_j with set of players N_j is played as a subgame.

The game tree of Γ and the assignment of players to decision nodes is thereby completely specified and it is clear that Γ will be compatible with $(N, \succ)$. We next have to define the players' preferences on the terminal nodes.

For a player $k \in N_j$, $j \in S(1)$, we choose preferences on the terminal nodes of Γ that extend those for the subgame Γ_j and are in generic position.

We will select the preferences of player 1 such that his outside option a_1^+ becomes his backward induction choice b_1 and hence z^+ becomes the backward induction outcome of Γ. We therefore impose the following restrictions on player 1's preferences:

- He is not indifferent between any two terminal nodes.
- He prefers any outcome to the outcome z^-:

$$z \, \mathrm{p}_1 \, z^-$$

for any outcome $z \neq z^-$.

- He prefers the outcome z^+ to the backward induction outcome in any subgame Γ_j, $j \in S(1)$:

$$z^+ \, \mathrm{p}_1 \, z_{b(j)}$$

for all $2 \leqslant j \leqslant j^*$.

- But for an outcome other than the backward induction outcome in the subgame Γ_j he prefers this outcome to z^+ and to any outcome in the subgames $\Gamma_{j'}$ with lower index j':

$$z_{a(j)} \, \mathrm{p}_1 \, z^+,$$

and

$$z_{a(j)} \, \mathrm{p}_1 \, z_{a(j')},$$

for any $2 \leqslant j' < j \leqslant j^*$, any outcome $z_{a(j')}$ of the subgame $\Gamma_{j'}$ and any outcome $z_{a(j)} \neq z_{b(j)}$ of the subgame Γ_j.

- Finally, among the backward induction outcomes of the subgames, he prefers the outcomes for the subgame Γ_j with lower index, i.e.

$$z_{b(j)} \, \mathrm{p}_1 \, z_{b(j')},$$

for all $2 \leqslant j < j' \leqslant j^*$.

Next we construct the model ψ by specifying the functions $f_j : \Omega \to A_j$. For any player $k \in N_j$, $j \in S(1)$, we take the function f_k as specified by the model ψ_j.

The function f_1 for player 1 is now defined by

$$[f_1 = a_1^-] := \neg E_1$$

$$[f_1 = a_1^+] := E_1 \cap \bigcap_{j \in S(1)} K_1[z_{f(j)} = z_{b(j)}],$$

and

$$[f_1 = a_1^j] := E_1 \cap \bigcap_{j<j'\leqslant j^*} K_1[z_{f(j')} = z_{b(j')}] \cap \neg K_1[z_{f(j)} = z_{b(j)}],$$

for $2 \leqslant j \leqslant j^*$.

Thus, player 1 uses the strictly dominated choice a_1^- on $\neg E_1$. On E_1 he uses the outside option a_1^+ if he knows that the backward induction outcome results in all subgames Γ_j. If he does not know on E_1 that the backward induction outcome results in all these subgames, he is optimistic and gives the move to the next player in the subgame Γ_j with the highest index j, where the backward induction outcome might not result.

The construction and the induction assumptions imply $R_k(\psi) = E_k$ for every player $k \in N_j$, $j \in S(1)$. We have to show that player 1 is rational exactly on E_1: If player 1 uses a strategy other than a_1^-, then a terminal node other than z^- will result, which he always prefers. Hence, for all $a_1 \in A_1$, $a_1 \neq a_1^-$,

$$[z_{(a_1,f^1)} \mathrm{p}_1 z_{(a_1^-,f^1)}] = \Omega$$
$$\Rightarrow \neg K_1[z_{(a_1,f^1)} \mathrm{p}_1 z_{(a_1^-,f^1)}] = \emptyset,$$

which implies

$$R_1(\psi, a_1^-) = \emptyset,$$

and

$$[f_1 = a_1^-] \cap R_1(\psi) = \neg E_1 \cap R_1(\psi) = \emptyset.$$

Player 1 is rational when choosing a_1^+. We obtain for all a_1^j, $j \in S(1)$,

$$[z_{f(j)} = z_{b(j)}] \subseteq \neg[z_{(a_1^j,f^1)} \mathrm{p}_1 z_{(a_1^+,f^1)}]$$
$$\Rightarrow K_1[z_{f(j)} = z_{b(j)}] \subseteq K_1 \neg [z_{(a_1^j,f^1)} \mathrm{p}_1 z_{(a_1^+,f^1)}]$$
$$\Rightarrow K_1[z_{f(j)} = z_{b(j)}] \subseteq \neg K_1[z_{(a_1^j,f^1)} \mathrm{p}_1 z_{(a_1^+,f^1)}],$$

which says that if player 1 knows that the backward induction outcome results in any subgame, then he is better off choosing the outside option a_1^+. Since

$$\neg K_1[z_{(a_1^-,f^1)} \mathrm{p}_1 z_{(a_1^+,f^1)}] = \Omega,$$

we conclude that

$$[f_1 = a_1^+] \subseteq \bigcap_{j \in S(1)} K_1[z_{f(j)} = z_{b(j)}] \subseteq R_1(\psi, a_1^+).$$

Player 1 is rational when taking action a_1^j ($2 \leqslant j \leqslant j^*$): For any action $a_1^{j'}$ with $j < j'$, player 1 is better off by giving the move to player j rather than to player j'

with a higher index if he knows that the backward induction outcome results in the subgame $\Gamma_{j'}$, i.e.

$$\begin{aligned}&[z_{f(j')} = z_{b(j')}] \subseteq \neg [z_{(a_1^{j},f^{1})} \mathrm{p}_1 z_{(a_1^{j'},f^{1})}] \\ &\Rightarrow K_1[z_{f(j')} = z_{b(j')}] \subseteq K_1 \neg [z_{(a_1^{j},f^{1})} \mathrm{p}_1 z_{(a_1^{j'},f^{1})}] \\ &\Rightarrow K_1[z_{f(j')} = z_{b(j')}] \subseteq \neg K_1 [z_{(a_1^{j},f^{1})} \mathrm{p}_1 z_{(a_1^{j'},f^{1})}] \\ &\Rightarrow [f_1 = a_1^{j}] \subset K_1[z_{f(j')} = z_{b(j')}] \subseteq \neg K_1 [z_{(a_1^{j},f^{1})} \mathrm{p}_1 z_{(a_1^{j'},f^{1})}],\end{aligned}$$

while for $a_1^{j'}$ with $j > j'$ he might gain more by giving the move to player j rather than to player j' with a lower index if he considers it possible that the backward induction outcome does not result in Γ_j:

$$\begin{aligned}&[z_{f(j)} = z_{b(j)}] = [z_{f(j')} \mathrm{p}_1 z_{f(j)}] = [z_{(a_1^{j'},f^{1})} \mathrm{p}_1 z_{(a_1^{j},f^{1})}] \\ &\Rightarrow \neg K_1[z_{f(j)} = z_{b(j)}] = \neg K_1 [z_{(a_1^{j'},f^{1})} \mathrm{p}_1 z_{(a_1^{j},f^{1})}] \\ &\Rightarrow [f_1 = a_1^{j}] \subseteq \neg K_1[z_{f(j)} = z_{b(j)}] = \neg K_1 [z_{(a_1^{j'},f^{1})} \mathrm{p}_1 z_{(a_1^{j},f^{1})}].\end{aligned}$$

Similarly

$$[f_1 = a_1^{j}] \subseteq \neg K_1[z_{(a_1^{+},f^{1})} \mathrm{p}_1 z_{(a_1^{j},f^{1})}],$$

and since

$$[f_1 = a_1^{j}] \subseteq \Omega = \neg K_1[z_{(a_1^{-},f^{1})} \mathrm{p}_1 z_{(a_1^{j},f^{1})}],$$

it follows that

$$[f_1 = a_1^{j}] \subseteq R_1(\psi, a_1^{j}).$$

Therefore,

$$E_1 = \bigcup_{a_1 \neq a_1^{-}} [f_1 = a_1] \subseteq R_1(\psi),$$

and since $\neg E_1 \cap R_1(\psi) = \emptyset$:

$$E_1 = R_1(\psi).$$

In the model ψ, player 1 takes his backward induction choice $b_1 = a_1^{+}$ and hence the backward induction path results if and only if he knows that the backward induction outcome results in every subgame. By the induction assumption the latter is equivalent to the knowledge that the backward induction solution is played in every subgame, i.e.

$$\begin{aligned}&[f_1 = b_1] = K_1\Big(\bigcap_{j \in S(1)} IP(\psi_j)\Big) \\ &\Rightarrow IP(\psi) = [f_1 = b_1] = [f_1 = b_1] \cap \bigcap_{j \in S(1)} IP(\psi_j) \\ &\qquad = [f_1 = b_1] \cap \bigcap_{j \in S(1)} I(\psi_j) = I(\psi).\end{aligned}$$

Finally,

$$Q(\psi) = Q_1(\psi) = R_1(\psi) \cap K_1\left(\bigcap_{j \in S(1)} Q_j(\psi)\right)$$
$$= R_1(\psi) \cap K_1\left(\bigcap_{j \in S(1)} IP(\psi_j)\right) = [f_1 = b_1] = IP(\psi),$$

which completes the proof. □

Proof of Proposition 4.1. Let $\psi = (\Omega, (\mathscr{P}_i)_{i \in N}, \Gamma, (f_i)_{i \in N})$ be a model compatible with $(N, \succ)$. By Lemma 4.3 we can find a game Γ' compatible with $(N, \succ)$ and a model $\psi' = (\Omega, (\mathscr{P}_i)_{i \in N}, \Gamma', (f'_i)_{i \in N})$ with the same knowledge space as ψ such that for all $i \in N$:

$$R_i(\psi') = R_i(\psi), \tag{4.2}$$

and

$$Q(\psi') = I(\psi').$$

By assumption on F, $F(\psi') \subseteq I(\psi') = Q(\psi')$. But (4.2) implies that any rationality condition yields for both models ψ and ψ' the same event in Ω. Hence, $F(\psi) \subseteq I(\psi)$, which was to be shown. □

5. Remarks

5.1. On our notion of minimality

In this paper we considered the minimal rationality condition for backward induction with respect to all models based on games with the same order of moves $(N, \succ)$. Alternatively, we could have asked for the minimal rationality condition with respect to all models based on a fixed game or with respect to all models based on games with a fixed game tree. This subsection illustrates that the results change significantly if such alternative notions of minimality are considered. In particular, the existence of a minimal rationality condition is no longer obvious. We will use the framework of the two-player games discussed in Section 3 for illustration.

In the game of Fig. 2 the player who moves first, player 1, has a strictly dominant move. Player 1's only rational choice is to choose 'left', regardless of what he knows about player 2's action. The backward induction solution will hence be played if both players are rational, i.e. if the formula

$$R_1 \cap R_2$$

holds. Obviously, the event 'forward knowledge of rationality' is for some models based on the game a strict subset of the event 'both players are rational'.

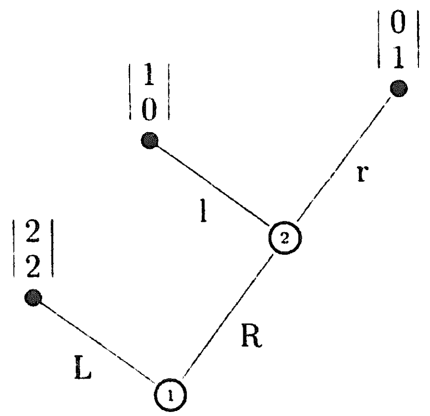

Fig. 2.

Consequently, *forward knowledge of rationality is not the minimal rationality condition for backward induction with respect to all models based on the game in Fig. 2.*

For the game in Fig. 3, where player 1 does not have a dominating move, it can actually be shown that *forward knowledge of rationality is a minimal rationality condition for backward induction with respect to all models based on this game.* However, the following peculiarity of Aumann's definition of rationality has an interesting implication for this game. Suppose player 1 does not know *whether* player 2 is rational, i.e. $\neg J_1 R_2$ holds with

$$J_1 R_2 := K_1 R_2 \cup K_1 \neg R_2.$$

Then player 1 considers it both possible that player 2 might be playing 'left' or 'right'. Since player 1 does not have a dominating move, he is rational regardless

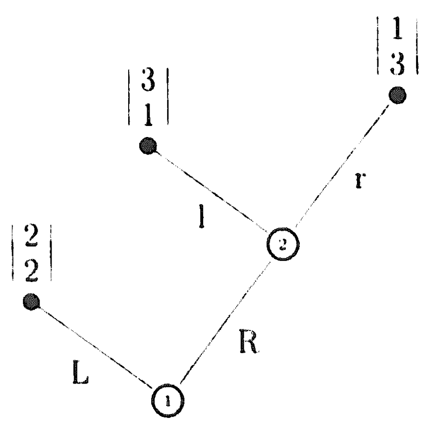

Fig. 3.

of what he does. Hence $R_1 \cup J_1 R_2$ is always true and common knowledge among the players. Consider now the following expression:

$$F := (R_1 \cap K_1 R_2) \cup (R_1 \cap R_2 \cap \neg CK(R_1 \cup J_1 R_2)). \qquad (5.1)$$

For the game in Fig. 3, the part of the expression following the union yields the empty set. Therefore, F and forward knowledge of rationality yield the same event for every model based on this game. However, it is intuitively clear, and also follows from the next paragraph, that F describes a statement that is logically strictly weaker than forward knowledge of rationality. Consequently, *forward knowledge of rationality is not the only minimal rationality condition for backward induction with respect to the models based on the game in Fig. 3. There are also others, which are strictly weaker in a logical sense.*

We have constructed F such that it implies $R_1 \cap R_2$. Therefore, it is also a rationality condition for backward induction with respect to all models based on the game in Fig. 2. Indeed, it is easy to see that F is a rationality condition for backward induction with respect to all models based on games *with the game tree* in Fig. 2. Moreover, one can construct a model with four states based on the game in Fig. 2, where both players are rational in a single state and where a player never knows whether his opponent is rational or not. In this model, forward knowledge of rationality yields the empty set. However, one has $CK(R_1 \cup J_1 R_2) \subseteq CK(R_1) \subseteq K_2(R_1) = \emptyset$ and hence F does not yield the empty set. Consequently, *forward knowledge of rationality is not the minimal rationality condition for backward induction with respect to all models based on games with the game tree in Fig. 2.*

We do not know whether F is the minimal rationality condition or whether a minimal rationality condition exists here. [3] The reason why this problem is resolved in our approach is that one can always add to a game a dominated action leading to a terminal node without changing the order of moves. Adding such an action for each player makes it possible that each player can be irrational without knowing anything about his opponents.

These difficulties appear in our model because we do not use a Bayesian approach. Our models do not specify the *beliefs* of the players in each state and in that sense the notion of rationality we use is *weaker* than Bayesian rationality. With specified beliefs it is, in the game of Fig. 3, not sufficient for Bayesian rationality that player 1 does not know anything about the actions planned by his opponent. For instance, player 1 may believe that player 2 chooses 'left' with probability 1/4 and 'right' with probability 3/4. Then player 1 is not Bayesian rational if he chooses 'right'.

[3] F is not a finite expression and hence does not correspond to a formula in the modal logic S5.

5.2. The minimal order-independent rationality condition for backward induction

Our notion of forward knowledge of rationality does not only depend on the set of players N but also on the order of moves $(N, \succ)$. We may be interested in a minimal condition on the knowledge of rationality for the players that implies backward induction for *all* games with player set $N = \{1, \ldots, n\}$ regardless of the order $\succ$. Using Lemma 4.3 it can be shown that this condition is simply the intersection of all forward knowledge of rationality conditions with respect to all possible orders $\succ$. It is easy to see that this intersection equals:

$$\bigcap_{\pi \in \Pi(n)} K_{\pi(1)} K_{\pi(2)}, \ldots, K_{\pi(n-1)} R_{\pi(n)}, \tag{5.2}$$

where $\Pi(n)$ is the set of all permutations of $\{1, \ldots, n\}$.

Condition (5.2) is still much weaker than common knowledge of rationality. For instance, due to the agent form this condition never requires a player to know what an opponent knows about him.

5.3. Games not in agent form

Our results can be adapted in a straightforward manner to more general perfect information games, where a player may have more than one decision node. The expression:

$$R_1 \cap K_1 R_2,$$

can, for instance, be read as 'the player at the first decision node takes a rational action at this node and knows that the player at the second decision node takes a rational action at that node'. Provided we know what the first and the second decision node is, this sentence is meaningful even if there is only a single player at both decision nodes. More generally, we can interpret R_i as the statement 'the player *at node i* takes a rational action at that node' and $K_i(\ldots)$ as 'the player *at node i* knows...'. [4] With this reading forward knowledge of rationality and any rationality condition is meaningful. Clearly, rationality in this altered sense in no longer an attribute of a player. It is an attribute of a decision. A player may act rationally at some decision nodes and irrationally at others.

To proceed formally, a game Γ consists now of a game tree, a set of players $M = \{1, \ldots, m\}$, an assignment $\kappa: N \to M$ of decision nodes to players and a map from terminal nodes to payoffs of the players. In particular, we have a well-defined ordering $\succ$ on the set of decision nodes $N = \{1, \ldots, n\}$. A model ψ for such a game specifies, in addition, a state space Ω, the knowledge partitions

[4] But notice again that we are using a notion of conditional rationality, i.e. for deciding whether it is rational for a player to take an action at some node it is not relevant whether this node can be reached or not.

$\mathscr{P}_1, \ldots, \mathscr{P}_m$ for the players and an assignment $j_i : \Omega \to A_i$ of states to actions for each decision node $i \in N$. From such a model we obtain a model ψ' for the agent form of the game if we duplicate the knowledge partitions for each decision node of a player, i.e. if we write $K_i(E)$ instead of $K_{\kappa(i)}(E)$ for a decision node i of a player $\kappa(i)$, and if we similarly duplicate the payoffs. The following statement is then obvious given our result.

Forward knowledge of rationality is the minimal rationality condition for backwards induction with respect to all models based on the games with the same order of moves $(N, \succ)$. (However, the number of players and the assignment of decision nodes to players can vary here.)

Acknowledgements

We wish to thank Sergiu Hart for many valuable discussions and helpful suggestions. We are also grateful to Bob Aumann, Christian Ewerhart and Dov Samet for valuable comments. The first author would like to thank the Center for Rationality and Interactive Decision Theory at the Hebrew University of Jerusalem for their hospitality and for financial support. Financial support from the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303 at the University of Bonn, is also gratefully acknowledged. The second author is grateful to the German Israeli Foundation for financial support.

Appendix: Minimality and logical implication

In this appendix we describe in which sense the statement 'every rationality condition for backward induction implies forward knowledge of rationality' holds in the epistemic (or modal) logic S5. For a full description of this logic, see, for example, Chellas (1980). Here we remind ourselves only of the following facts.

Propositions of the epistemic logic S5 are – in the case of several players – inductively formed from a list of basic propositions $a_1, a_2, \ldots,$ by applying a finite number of times the logical operators $\vee$, $\wedge$ and $\neg$ and the knowledge operators $k_i (i \in N = \{1, \ldots, n\})$. A proposition in S5 is a theorem if it can be deduced from the axioms of standard propositional logic, together with the axioms on knowledge, by using the inference rules modus ponens and necessitation. This syntactic approach is linked with the semantic approach based on knowledge spaces via the notion of an interpretation. An *interpretation* consists of a knowledge space $(\Omega, (\mathscr{P}_i)_{i \in N})$ and a mapping that assigns to every basic proposition a_k an event $E_k \subseteq \Omega$. Every proposition f defines a well-defined set theoretic formula and hence an event $[f] \subseteq \Omega$ if we replace the logical operators $\vee$, $\wedge$ and $\neg$ by the set-theoretic operators $\cup$, $\cap$ and $\neg$, the k_i by the knowledge operators

K_i, and the elementary propositions a_k by the events E_k.[5] Strong completeness of S5 implies that a proposition g can be deduced from a proposition f by the axioms and inference rules of S5 if and only if $[f]\subseteq[g]$ holds for every interpretation.

We translate the notion of a rationality condition into this setting. A rationality condition was constructed inductively from the expressions R_i. The definition of rationality, together with the measurability assumption, implies in our framework that a player always knows when he is rational, i.e. $K_i(R_i)=R_i$. For the completeness result of S5 it is important that a basic proposition can also be interpreted as an event that a player does not know. Therefore, we cannot take basic propositions as the statements 'player i is rational'. We take instead the propositions $r_1 := k_1(a_1),\ldots, r_n := k_n(a_n)$, where the $a_1,\ldots,a_n$ denote n distinct fixed basic propositions. We call a proposition a 'rationality condition' if it is built inductively from these propositions by using the logical operators and the knowledge operators. Given an order of moves $(N, \succ)$ 'forward knowledge of rationality' yields a well-defined rationality condition q in this sense.

For a given order of moves $(N, \succ)$ we mean by the statement 'the rationality condition f is a condition for backward induction' that the following holds for every model ψ based on a game with the given order of moves:[6] if an interpretation assigns to the basic propositions $a_1,\ldots,a_n$ (and hence to $r_1,\ldots,r_n$) the events $R_1(\psi),\ldots,R_n(\psi)\subseteq\Omega$, then $[f]\subseteq I(\psi)$ holds. This condition is satisfied if f corresponds to a set-theoretic formula that is a rationality condition for backward induction as defined in Section 2.

Suppose now that f is a rationality condition for backward induction. Suppose we are given any interpretation with a knowledge space $(\Omega, (\mathscr{P}_i)_{i\in N})$. The propositions $r_1,\ldots,r_n$ are then mapped to events $E'_1 := K_1(E_1),\ldots,E'_n := K_n(E_n)$ satisfying $E'_i = K_i(E'_i)$. Using Lemma 4.3 (with E'_i instead of E_i) it follows that $[f]\subseteq[q]$. Since this holds for every interpretation, strong completeness of S5 implies: forward knowledge of rationality q can be deduced via the axioms and inference rules of S5 from every rationality condition for backward induction f.

References

Aumann, R.J., 1992, Irrationality in game theory, in: O. Hart, P. Dasgupta, D. Gale and E. Maskin, eds., Economic analysis of markets and games (MIT Press, Cambridge, MA) 282–290.

Aumann, R.J., 1995, Backward induction and common knowledge of rationality, Games and Economic Behavior 8, 6–19.

[5] However, not every well-defined set-theoretic formula corresponds to a proposition in S5 because S5 allows only for *finite* formulas. For instance, common knowledge cannot be formulated in S5 (see Kaneko and Nagashima, 1993). For a non-finitary logic where common knowledge is well-defined see Kaneko and Nagashima (1991).

[6] For brevity we do not give a purely syntactic description of this statement

Chellas, B., 1980, Modal logic (Cambridge University Press, Cambridge).

Fudenberg, D., D.M. Kreps and D.K. Levine, 1988, On the robustness of equilibrium refinements, Journal of Economic Theory 44, 354–380.

Güth, W., R. Schmittberger and B. Schwarze, 1982, An experimental analysis of ultimatum bargaining, Journal of Economic Behavior and Organization 3, 376–388.

Kaneko, M. and T. Nagashima, 1991, Game logic, parts I, II and III, Mimeo.

Kaneko, M. and T. Nagashima, 1993, Indefinability of common knowledge concept in finitary logics, Discussion Paper 551, Institute of Socio-Economic Planning, University of Tsukiuba.

Kramerz, F., 1993, How agents plan their actions in games: A model of player's reasoning with common knowledge of rationality, Discussion Paper 9325, INSEE, Paris.

Kreps, D., and R. Wilson, 1982, Reputation and imperfect information, Journal of Economic Theory 27, 253–279.

Milgrom, P., and J. Roberts, 1982, Predation, reputation and entry deterrence, Journal of Economic Theory 27, 280–312.

Rosenthal, R.W., 1981, Games of perfect information, predatory pricing and the chain store paradox, Journal of Economic Theory 25, 92–100.

Samet, D., 1993, Hypothetical knowledge and games with perfect information, Mimeo.

Selten, R., 1978, The chain-store paradox, Theory and Decision 9, 127–159.