36 Backward Induction and Common Knowledge of Rationality

1. Introduction

Backward induction, the oldest idea in game theory, has maintained its centrality to this day. Cornerstone of Zermelo's (1913) proof that chess has optimal pure strategies, it subsequently played a vital role in the development of perfect equilibrium (Selten, 1965, 1975). In turn, this led to the modern "refinement" literature, which seeks to identify particularly "significant" equilibria in a game.

Quite apart from its relation to Zermelo's theorem and to refinements,

[1] Research support under NSF grant SES-9209277 at the economics department, Stanford University, is gratefully acknowledged. Also gratefully acknowledged are numerous discussions with the participants of the following conferences, at which earlier versions of this paper were presented: the Nobel Symposium on Game Theory, Bjorkborn, June 1993; the First Summer Workshop on Game Theory, Valencia, June 1993; the Fourth International Conference on Game Theory, Stony Brook, July 1993; and the annual summer workshop of the Stanford Institute for Theoretical Economics, Stanford, August 1993. The individuals whose comments materially affected the current form of the paper are too numerous to mention here, but three who were especially helpful should be singled out. Long conversations with Dov Samet enabled the author to sort out many of the conceptual problems raised by the analysis herein; Phil Reny kept the author on his toes by shooting down countless previous attempts to get this thing right; and a long correspondence with Ken Binmore helped us to focus our presentation.

This chapter originally appeared in *Games and Economic Behavior* 8 (1995): 6–19. Reprinted with permission.

backward induction has a compelling logic of its own, especially in perfect information (PI) games (like chess). The last player, who must choose between leaves of the game tree, makes a choice that maximizes his payoff; taking this as given, the previous player makes a choice maximizing *his* payoff; and so on, until the beginning of the game is reached. Nothing seems simpler or more natural.

Yet it is precisely this logic that has come under increasing recent scrutiny (e.g., Basu, 1990; Ben Porath, 1997; Bicchieri, 1989, 1992; Bicchieri and Antonelli, 1995; Binmore, 1987; Binmore and Brandenburger, 1990; Bonanno, 1991; Pettit and Sugden, 1989; Reny 1992). Indeed, it is not obvious just what assumptions on the rationality of the players would justify it. Simple rationality on the part of each player—each player's being a utility maximizer—is certainly not enough; in some sense, the players must also ascribe rationality to each other. It has been suggested that one assume "common knowledge of rationality": that all players know that all are rational, all know this, all know *this,* and so on ad infinitum (or at least, for a number of levels no less than the maximum duration of the game). But as the works cited above show, in extensive games there are serious difficulties in formulating the relevant concepts (knowledge, rationality, etc.).

The aim of this paper is to present a coherent formulation and proof of the principle that "in PI games, common knowledge of rationality implies backward induction." Some of its features are as follows:

(i) Each player chooses a *strategy,* in the usual game theoretic sense of the term (Kuhn, 1953, say); that is, he decides what to do at each of his vertices v in the game tree, whether or not v is reached.

(ii) When deciding what to do at v, the player considers the situation *from that point on:* he acts *as if* v is reached. It is this feature that distinguishes the current analysis from a strategic form analysis.

(iii) *Rationality* of a player means that he is a habitual payoff maximizer: that no matter where he finds himself—at which vertex—he will not knowingly continue with a strategy that yields him less than he could have gotten with a different strategy.

(iv) The result is not empty; common knowledge of rationality is possible in every PI game.

(v) Knowledge means absolute certainty, not probability 1 belief. Probability plays no role in the model.[2] However, probability may easily be introduced, and then the results hold with the usual expected-utility-maximizing definition of rationality.

[2] Thus one does not need Savage's (1954) axioms or any similar axioms.

(vi) The *time* of the players' knowledge is the start of play, before any actions are taken.

The plan of the paper is as follows: Section 2 contains the formal statements of our results, Section 3, the proofs. The formalism is interpreted in Section 4. Finally, Section 5 is devoted to discussion.

2. The Results

Let a PI (perfect information) game (Kuhn, 1953) with n players be given. Assume that it is in "general position," i.e., that the payoffs to each player at different leaves of the game tree are different. Say that vertex w comes *after* vertex v (written $w > v$) if each play through w also goes through v.

To define the concept of inductive (short for backward inductive) choice, let v be a vertex of Player i, and suppose that the inductive choice is defined at all vertices w after v. Then the *inductive choice* b^v at v is defined as the action that maximizes i's payoff when all players make the inductive choices at all vertices after v. In particular, when there are no vertices after v—when the action at v directly determines the outcome—then the inductive choice is simply the action maximizing i's payoff. When the inductive choices are made at all vertices, the resulting outcome is called the *inductive outcome*.

To formalize knowledge, we use the standard partition model. A *strategy* of Player i is a function s_i that assigns to each vertex v of i an action at that vertex. Denote the set of i's strategies by S_i. A *knowledge system* (for the given game) consists of

a set Ω (the *states of the world,* or simply *states*),

a function $\mathbf{s}$ from Ω to $\times_i S_i$,

and for each player i,

a partition $\mathcal{K}_i$ of Ω (i's *information partition*).

Intuitively, i can distinguish between two states if and only if they are in different atoms of $\mathcal{K}_i$, and $\mathbf{s}(\omega)$ represents the n-tuple of the players' strategies at the state ω. We assume that

$\mathbf{s}_i$ is measurable with respect to $\mathcal{K}_i$, which means that $\mathbf{s}_i(\omega) = \mathbf{s}_i(\nu)$ whenever ω and ν are in the same element of $\mathcal{H}_i$. Intuitively, measurability of $\mathbf{s}_i$ says that i knows his own strategy.

An *event* is a set of states. If E is an event, K_iE denotes the union of those elements of the information partition $\mathcal{H}_i$ that are included in E; in words, K_iE is the event that *i knows E*. Set $KE := \cap_i K_iE$ and

$$CKE := KE \cap KKE \cap KKKE \cap \cdots.$$

In words, CKE is the event that E is *commonly known.*

Let s be an n-tuple of strategies, one for each player, and let v be a vertex of Player i. Define i's *conditional payoff* $h_i^v(s)$ *for s at v* as his payoff at the leaf to which the players are led if starting at v, they play[3] s. In a given state ω of the world, call *i rational at v* if there is no strategy that i knows would have yielded him a conditional payoff at v larger than that which in fact he gets.[4] Call *i rational* if he is rational at each of his vertices. Denote by R_i the event[5] "i is rational." The event "all players are rational," denoted R, is the intersection of all the R_i.

In each state ω, there is a unique outcome: that determined by $\mathbf{s}(\omega)$. Denote by I the event that this is the inductive outcome.

The first result states that if rationality is commonly known, the inductive outcome results. The formal statement is

THEOREM A. $CKR \subset I$.

The second result is that common knowledge of rationality is indeed possible in every PI game; i.e., for every PI game there is a knowledge system with a state of the world at which it is commonly known that all players are rational. This distinguishes the present model from certain others, in which common knowledge of rationality is in many games unachievable. The formal statement is

THEOREM B. *For every* PI *game, there is a knowledge system with* $\emptyset \neq CKR$.

3. Proofs

It is convenient to develop some notation, and recall some facts from the theory of knowledge, before coming to the proofs themselves.

If s is a strategy profile and v a vertex, denote by s^v the action that s

[3] When v is not reached, i's conditional payoff is different from his actual payoff.

[4] I.e., larger than i's conditional payoff at v for $\mathbf{s}(\omega)$. This is weaker than (i.e., implied by) expected utility maximization; if an agent is maximizing expected utility, surely he cannot have an option that he knows, with certainty, will yield a preferred outcome. Thus our result remains true if we substitute expected utility maximization for the above definition of rationality. See Section 4c.

[5] For a characterization of R_i in symbols, see (3) in Section 3.

prescribes at v; if a^v is another action at v, denote by $(s; a^v)$ the profile obtained from s by replacing s^v by a^v. If t_i is a strategy of i, denote by $(s; t_i)$ the profile obtained from s by replacing s_i by t_i. Thus if b is the strategy profile that assigns the inductive choice to each vertex v, then for each vertex v of each player i,

$$h_i^v(b) \geq h_i^v(b; a^v) \qquad \text{for all actions } a^v \text{ at } v. \tag{1}$$

Functions defined on Ω (like $\mathbf{s}$ or $\mathbf{s}_i$), which appear in bold face, may be viewed like random variables in probability theory. An assertion about such a function corresponds to an event, which we denote by putting square brackets around the assertion. For example, $[\mathbf{s}_i = s_i]$ is the event that i chooses the strategy s_i (i.e., the set of all states ω for which $\mathbf{s}_i(\omega) = s_i$); and $[h_i^v(\mathbf{s}; t_i) > h_i^v(\mathbf{s})]$ is the event that i's conditional payoff would have been higher if he had chosen the strategy t_i rather than what he did choose. With this notation, the measurability requirement for $\mathbf{s}_i$ translates to

$$[\mathbf{s}_i = s_i] \subset K_i[\mathbf{s}_i = s_i] \qquad \text{for all } s_i \in S_i \tag{2}$$

(if i chooses the strategy s_i, then he knows that he chooses it); the event "i is rational" is given by

$$R_i = \bigcap_{v \in V_i} \bigcap_{t_i \in S_i} (\sim K_i[h_i^v(\mathbf{s}; t_i) > h_i^v(\mathbf{s})]), \tag{3}$$

where $\sim$ denotes complementation (for each of his vertices v and strategies t_i, it is not the case that i knows that t_i would yield him a higher conditional payoff at v than the strategy he chooses); and the event "the inductive choice is made at v" is

$$I^v := [\mathbf{s}^v = b^v].$$

The proof uses seven lemmas; Lemmas 4 through 9 are standard in knowledge theory, and also easily proved directly.

LEMMA 4. *$CKE = K_iCKE$.*

LEMMA 5. *If $E \subset F$, then $K_iE \subset K_iF$.*

LEMMA 6. *$K_iE \cap K_iF = K_i(E \cap F)$.*

LEMMA 7. *$CKE \subset E$.*

LEMMA 8. $K_i \sim K_iE = {\sim}K_iE$.

LEMMA 9. $K_iE \subset E$.

LEMMA 10. *$I^v \subset K_iI^v$ for each vertex v of Player i (if i chooses b^v at v, then he knows this).*

Proof. Let B_i^v denote the set of those strategies s_i of i for which $s_i^v = b^v$. Then by (2) and Lemma 5, $I^v = [\mathbf{s}_i^v = b^v] = \bigcup_{s_i \in B_i^v}[\mathbf{s}_i = s_i] \subset \bigcup_{s_i \in B_i^v} K_i[\mathbf{s}_i = s_i] \subset \bigcup_{s_i \in B_i^v} K_i[\mathbf{s}_i^v = b^v] \subset \bigcup_{s_i \in B_i^v} K_iI^v = K_iI^v$. ☺

This completes the preliminaries, and brings us to the

Proof of Theorem A. It suffices to show that

$$CKR \subset I^v$$

for each vertex v. Proceeding inductively, assume

$$CKR \subset I^w$$

for all $w > v$. Let v be a vertex of Player i. By Lemmas 4 and 5, $CKR = K_iCKR \subset K_iI^w$ for all such w, so by Lemma 6,

$$CKR \subset \bigcap_{w>v} K_iI^w = K_i \bigcap_{w>v} [\mathbf{s}^w = b^w] = K_i[\mathbf{s}^{>v} = b^{>v}], \tag{11}$$

where $s^{>v}$ denotes the profile of actions prescribed by s at vertices after v. Using Lemma 7, and setting $t_i = b_i$ in (3), we get

$$CKR \subset R \subset R_i \subset {\sim}K_i[h_i^v(\mathbf{s}; b_i) > h_i^v(\mathbf{s})]. \tag{12}$$

Now since $h_i^v(s)$ depends only on s^v and $s^{>v}$, Lemma 6 yields

$$\begin{aligned} &K_i[\mathbf{s}^{>v} = b^{>v}] \cap K_i[h_i^v(\mathbf{s}; b_i) > h_i^v(\mathbf{s})] = K_i[\mathbf{s}^{>v} = b^{>v} \wedge h_i^v(\mathbf{s}; b_i) > h_i^v(\mathbf{s})] \\ &= K_i[\mathbf{s}^{>v} = b^{>v} \wedge h_i^v(b) > h_i^v(b; \mathbf{s}^v)] \\ &= K_i[\mathbf{s}^{>v} = b^{>v}] \cap K_i[h_i^v(b) > h_i^v(b; \mathbf{s}^v)], \end{aligned}$$

so by complementation w.r.t. $K_i[\mathbf{s}^{>v} = b^{>v}]$,

$$\begin{aligned} &K_i[\mathbf{s}^{>v} = b^{>v}] \cap {\sim}K_i[h_i^v(\mathbf{s}; b_i) > h_i^v(\mathbf{s})] \\ &= K_i[\mathbf{s}^{>v} = b^{>v}] \cap {\sim}K_i[h_i^v(b) > h_i^v(b; \mathbf{s}^v)] \subset {\sim}K_i[h_i^v(b) > h_i^v(b; \mathbf{s}^v)]. \end{aligned} \tag{13}$$

Finally, (11–13, 1, 2), and Lemmas 8–10 yield

$$\begin{aligned} &CKR \subset K_i[\mathbf{s}^{>v} = b^{>v}] \cap \sim K_i[h_i^v(\mathbf{s}; b_i) > h_i^v(\mathbf{s})] \subset \sim K_i[h_i^v(b) > h_i^v(b; \mathbf{s}^v)] \\ &= \sim K_i[\mathbf{s}^v \neq b^v] = \sim K_i \sim I^v = \sim K_i \sim K_i I^v = \sim\sim K_i I^v = K_i I^v \subset I^v, \end{aligned}$$

as was to be proved. ☺

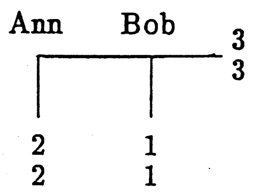
Proof of Theorem B. Define a knowledge system by stipulating that Ω consists of a single state ω, in which each player makes his inductive choice at each of his vertices. Then $\omega \in CKR$. ☺

4. Interpretation

a. Strategies. A player's strategy specifies what he does at each of his vertices, *if* reached. To clarify and fix the ideas, it may be useful to think of the players as attaching automata to their vertices before play starts. They program the automata to take a unique action at each vertex; if play reaches a vertex v, the attached automaton is triggered, and the programmed action is played. Each player knows how he programs his own automata; also, he may—but need not—know something about how the other players' automata are programmed, and about what they know about each other's knowledge. It is this knowledge that is represented by the knowledge system.

b. Rationality. Rationality of a player at a vertex v is defined in terms of what happens at vertices *after*[6] v; his payoff *if* v is reached. The idea is that when programming his automaton at v, the player does so as if v will be reached—even when he knows that it will not be! Each choice must be rational "in its own right"—a player may not rely on what happened previously to get him "off the hook."

This point is crucial. If we demand rationality only at vertices that are actually reached, the theorem fails already in the simplest cases. In Game 1, for example, if it is commonly known that both Ann and Bob go down, then it is commonly known that both are rational at reached vertices,



GAME 1

[6] This should not be misunderstood; see Section 5d.

since Bob's vertex is not reached. Nevertheless, the outcome is not that of backward induction, and indeed seems inconsistent with the plain meaning of the words "Ann is rational and knows that Bob is." The point is that to decide what to do, Ann must ask herself what Bob would do if she went across. The subjunctive mood—what he "would" do, even when not given the opportunity—is of the essence here.

Please see Sections 5b and 5c for further discussion of this point.

c. Probability. Though our model uses no probabilities, one can, if one wishes, introduce them. Thus for each state ω and player i, one may specify a probability distribution over i's information set[7], signifying his probabilities over the state space Ω, given his information. In that case rationality in the traditional sense of utility maximization (i.e., that a player's action maximizes the expectation of his conditional payoff) implies rationality in the sense of Sections 2 and 3; so it follows from our theorem that common knowledge of rationality in the sense of utility maximization implies that the outcome is that of backward induction.

d. Time. Already in the introduction, we touched on the question of the time at which the knowledge obtains. When we say that a player knows something, it is appropriate to ask, when? To what time are we referring? Clearly, a player may learn something—e.g., about actions of other players—during the course of play, so it is important to specify the time of the knowledge.

The formalism, which does not refer explicitly to this matter, is subject to different interpretations. We adopt the most straightforward, that all knowledge refers to the start of play: For a player to "know" something means that he knows it before any actions are taken.

5. Discussion

a. The Literature. Most of the recent literature on the subject of this paper (see Section 1) takes one of two positions (and sometimes both): that common knowledge of rationality (*CKR*) is impossible in most perfect information (PI) games; or, that it is possible, but need not lead to backward induction (BI). We do not take issue with this literature here. Whether or not *CKR* is possible and implies BI depends crucially on the precise model; the models in the cited literature are substantially different from ours, and our results do not formally contradict anything in this literature.

Nevertheless, there is a very clear, compelling intuition that *CKR* both

[7] The element of $\mathcal{H}_i$ containing ω.

implies BI and is possible. The model presented here is designed to capture this intuition. Moreover, we feel that it faithfully reflects the usual meanings of the key concepts: knowledge and rationality.

In another direction, a number of recent papers have been devoted to the program of formulating the fundamental notions of noncooperative game theory directly in terms of one person decision theory—i.e., in terms of the rationality of each player and what he knows about the game and about the other players. These include Aumann (1987), which treats correlated equilibrium in the strategic ("normal") form; Brandenburger and Dekel (1989) and Aumann and Brandenburger (1995), which treat Nash equilibrium in the strategic form; and others (e.g., Tan and Werlang, 1988). The current work extends this literature in two directions: First, it treats PI games—a particular class of extensive games—rather than the strategic form games treated in the previous literature; second, it treats backward induction, which is a refinement of Nash equilibrium. It remains to be seen whether our methods will shed light on the question of refinements in more general extensive games.

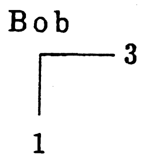
b. Conditionals. A crucial role in this paper is played by conditionals—statements of the form "if p, then q." In mathematics, this means simply "q or not p;" such conditionals are called *material.* But in everyday discourse, a conditional usually expresses a substantive relation between the antecedent p and the consequent q. We call such conditionals *substantive*[8].

A *counterfactual* is a substantive conditional with a false antecedent. Thus a substantive conditional may be viewed as the disjunction of a material conditional and a counterfactual.

Consider, for example, the statement "If White pushes his pawn, Black's queen is trapped." For this to hold in the material sense, it is sufficient that White does not, in fact, push his pawn. For the substantive sense, we ignore White's actual move, and *imagine* that he pushes his pawn. If Black's queen is then trapped, the substantive conditional is true; if not, then not.

If White did not push his pawn, we may still say "If he *had* pushed his pawn, Black's Queen *would* have been trapped." This is a counterfactual. To determine whether it holds, we proceed as above: imagine that the pawn was pushed, and see whether the Queen was trapped.

[8] This term was coined by us. Philosophers have various names for different kinds of substantive conditional, but there is no consensus on a general term. Some use "strict," but apparently this is reserved for a special kind of nonmaterial conditional. Some use "subjunctive"; the subjunctive mood does indeed have the flavor we are trying to express (see Section 4b), but sounds like a grammatical technicality. In plain English, moreover, the subjunctive is widely used with a false antecedent only, which is not what we want here.



GAME 2

Philosophers have long grappled with the task of assigning precise meaning to counterfactuals like "If Hitler had crossed the channel after Dunkirk, he would have won the war." Such statements are indeed problematic, because the hypothesis is not completely specified. If Hitler had crossed the channel, the world would have been different in a myriad of ways. To assign meaning to such a conditional, one must be more specific about the hypothetical world created by the crossing. That is a nontrivial task, even in principle, and there is no consensus among philosophers on how to approach it.

But there is no such problem with the chess example above; the conditional has a clear, precise, natural meaning. Similarly, the substantive conditionals we shall encounter below have unambiguous meanings.

Though they do not occur in pure mathematics, one really cannot discuss rationality, or indeed decision making, without substantive conditionals and counterfactuals. Making a decision means choosing among alternatives. Thus one *must* consider hypothetical situations—what would happen if one did something different from what one actually does. In Game 2 (actually a one-person decision problem), Bob's only rational choice is "across." That means that if he were to go down, he would get less—a counterfactual. In interactive decision making—games—you must consider what other people would do if you did something different from what you actually do.

c. Rationality and Substantive Conditionals. Substantive conditionals are not part of our formal[9] apparatus, but they are important in interpreting four key concepts that were formally defined in Section 2: strategy, conditional payoff, rationality at a vertex, and rationality. Briefly, the interpretations are as follows:

A player's strategy specifies what he does at each of his vertices, *if* reached. His conditional payoff (Section 2) at a vertex v signifies what he gets *if* v is reached. For him to be rational at v means that he cannot knowingly increase his payoff *if* v is reached. And for him to be rational means that for each of his vertices v, he cannot knowingly increase his payoff *if* v is reached. All these "if"s are substantive.

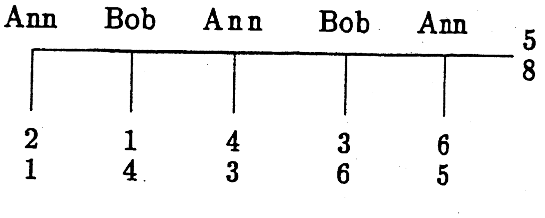
[9] A formal theory of substantive conditionals has recently been devised by Samet (1996).

One may consider substituting a material "if" for the substantive "if" in the interpretation of rationality. Formally, this amounts to demanding rationality at reached vertices only (rather than at all vertices, as in the definition of rationality). Let us call this *material rationality;* it is clearly weaker—less demanding—than rationality as defined[10]. Game 1 (Section 4b) shows that common knowledge of material rationality does *not* imply backward induction.

It may be useful to think of (substantive) rationality as an attribute of *players,* and of material rationality as an attribute of *play*. A rational player is a habitual maximizer. Since he can be depended on always to choose rationally, one may say that if an unreached vertex v *were* reached, he *would* choose rationally there. On the other hand, "play" may reasonably be understood to refer to reached vertices only. For Ann to go down in Game 1 is, as we said above, inconsistent with the plain meaning of the words "Ann is rational and knows that Bob is." But it is not inconsistent with common knowledge of rational play; Bob does not get to play, so if Ann expects him to go down, all actual play is rational.

d. Taking Account of Past Play. In Section 4b, we said that "rationality of a player i at a vertex v is defined in terms of what happens at vertices *after* v." What this means is that when evaluating his conditional payoff at v, the player must assume that v is reached, even when he knows that it is not; he cannot say "Since I know that v is not reached, whatever I do there is rational." It does *not* mean that i should ignore what happened before v in forming his opinions about what will happen afterwards. He can base his opinions as to what happens after v on whatever he wants; it's just that he cannot say "let's ignore v, it's irrelevant."

For example, consider Game 3, which is one of Rosenthal's (1982) "centipede" games. At each vertex, the inductive choice is to go down. If Bob finds himself at his first vertex, then it is clear to him that Ann did not make her inductive choice at her first vertex. It's quite possible, then, that he will consider it likely that she will not make her inductive choice



GAME 3

[10] Which might also be called *substantive* rationality.

at her second vertex either. In that case it is entirely rational for him to go "across" at his first vertex.

To be sure, *common knowledge* of rationality does imply that the strategies of both players call for them to go down at each vertex; that follows from our Theorem A. But rationality itself does not. Rationality permits the players to take account of past play in any way they want when forming estimates as to what will happen in the future. In fact, it places no restrictions at all on what they think others will do.

e. Ex Post Rationality. As noted in Section 4d, it is best to think of knowledge in our model as obtaining at the start of play. Thus for a player to be rational at a vertex v means that he has no choice at v that he knows *already at the start of play* would yield him a conditional payoff larger than that which in fact he gets. To underscore this point, we shall in this subsection refer to what we have called "rationality" as "ex ante rationality."

There is another kind of rationality that might seem more relevant. Call a player "ex post rational" at v if he has no choice at v that he knows *at the time of his move* would yield him a conditional payoff larger than that which in fact he gets. One may ask whether our results remain valid when we substitute ex post for ex ante rationality.

The answer is "yes." When the time comes for a player to move, he certainly knows at least as much as he did when play started. So if he knew when play started that his choice is suboptimal—i.e., that a different choice would yield him a larger conditional payoff—then a fortiori he knows this at the time of his move. So if he does not know at the time of his move that his choice is suboptimal, then a fortiori he did not know this at the start of play. Thus ex post rationality implies ex ante rationality. Therefore common knowledge of ex post rationality implies common knowledge of ex ante rationality. By Theorem A, common knowledge of ex ante rationality implies backward induction (BI); so common knowledge of ex post rationality also implies BI. In the other direction, it may be seen as in the proof of Theorem B that common knowledge of ex post rationality is possible in every PI game.

In brief, ex ante rationality is weaker—less demanding—than ex post rationality, and so more desirable in a hypothesis.

The reader may ask why we do not formalize the definitions of ex post and ex ante and make formal theorems out of the verbal discussion here. The reason is that to do this, we would have had to assign several knowledge operators to each player, one for each of his vertices, and we wanted to keep the formal model as transparent and simple as possible.

f. Off-Path Behavior. The results of this paper say nothing about the behavior of players at vertices that are off the backward induction path and are actually reached. Under common knowledge of rationality (*CKR*), vertices off the backward induction path *cannot* be reached; and when *CKR* does not obtain, the results do not apply.

g. Belief with Probability 1. Our approach views knowledge as absolute certainty. It does not work with probability 1 belief. Probabilistic approaches to backward induction present their own problems[11]—having to do mainly with off-path behavior—which are not easy to sort out. We have nothing to report on this matter here.

h. Backward Induction and Rationality: A Disclaimer. We have shown that common knowledge of rationality (*CKR*) implies backward induction. Does that mean that in perfect information games, only the inductive choices are appropriate or wise? Would we always recommend the inductive choice?

Certainly not. *CKR* is an ideal[12] condition that is rarely met in practice; when it is not met, the inductive choice may be not only unreasonable and unwise, but quite simply irrational. In Rosenthal's (1982) centipede games, for example, even minute departures from *CKR* may make it incumbent on rational players to "stay in" until quite late in the game (Aumann, 1992); the resulting outcome is very far from that of backward induction. What we have shown is that *if* there is *CKR, then* one gets the backward induction outcome; we do not claim that *CKR* obtains or "should" obtain, and we make no recommendations.

References

Aumann, R. J. (1987). "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica* **55**, 1–18 [Chapter 33].

Aumann, R. J. (1992). "Irrationality in Game Theory," in *Economic Analysis of Markets and Games*, Essays in Honor of Frank Hahn (P. Dasgupta, D. Gale, O. Hart, and E. Maskin, Eds.), pp. 214–227. Cambridge: MIT Press [Chapter 35].

Aumann, R. J., and Brandenburger, A. (1995). "Epistemic Conditions for Nash Equilibrium," *Econometrica* **63**, 1161–1180 [Chapter 37].

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

[11] One difficulty is that the reasoning in Section 5e breaks down; ex post rationality need not imply ex ante rationality. If you knew something at the beginning of play, then you certainly know it later; but if you ascribe probability 1 to it at the beginning of play, you need not do so at a later vertex, if that vertex initially had probability 0. One can try to fix this, but then something else unravels. Perhaps it can be done, but we have not succeeded.

[12] This is not a value judgment; "ideal" is meant as in "ideal gas."

BEN PORATH, E. (1997). "Rationality, Nash Equilibrium, and Backward Induction in Perfect Information Games," *Rev. Econ. Studies* **64**, 23–46.

BICCHIERI, C. (1989). "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntniss* **30**, 69–85.

BICCHIERI, C. (1992). "Knowledge-Dependent Games: Backward Induction," in *Knowledge, Belief and Strategic Interaction* (C. Bicchieri and M. L. Dalla Chiara, Eds.), Cambridge: Cambridge Univ. Press.

BICCHIERI, C., AND ANTONELLI, G. A. (1995). "Game-Theoretic Axioms for Local Rationality and Bounded Knowledge," *J. Logic Lang. Info.* **4**, 145–167.

BINMORE, K. (1987). "Modelling Rational Players I," *Econ. Philos.* **3**, 179–214.

BINMORE, K., AND BRANDENBURGER, A. (1990). "Common Knowledge and Game Theory," in *Essays on the Foundations of Game Theory*, pp. 105–150. Oxford: Basil Blackwell.

BONANNO, G. (1991). "The Logic of Rational Play in Games with Perfect Information," *Econ. Philos.* **7**, 37–65.

BRANDENBURGER, A., AND DEKEL, E. (1989). "The Role of Common Knowledge Assumptions in Game Theory," in *The Economics of Missing Markets, Information, and Games* (F. Hahn, Ed.), pp. 46–61. Oxford: Oxford Univ. Press.

KUHN, H. W. (1953). "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games II* (H. W. Kuhn and A. W. Tucker, Eds.), Annals of Mathematics Studies, Vol. 28, pp. 193–216, Princeton: Princeton Univ. Press.

PETTIT, P., AND SUGDEN, R. (1989). "The Backwards Induction Paradox," *J. Philos.* **4**, 1–14.

RENY, P. (1992). "Rationality in Extensive Form Games," *J. Econ. Perspect.* **6**, 103–118.

ROSENTHAL, R. (1982). "Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox," *J. Econ. Theory* **25**, 92–100.

SAMET, D. (1996). "Hypothetical Knowledge and Games with Perfect Information," *Games Econ. Behav.* **17**, 230–251.

SAVAGE, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley.

SELTEN, R. (1965). "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit," *Z. Ges. Staatswiss.* **121**, 301–324.

SELTEN, R. (1975). "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *Int. J. Game Theory* **4**, 25–55.

TAN, T., AND WERLANG, S. (1988). "The Bayesian Foundations of Solution Concepts of Games," *J. Econ. Theory* **45**, 370–391.

ZERMELO, E. (1913). "Ueber eine Anwendung der Mengenlehre auf die Theorie des Schachspiels," in *Proceedings of the Fifth International Congress of Mathematicians, Cambridge, 1912,* Vol. II, pp. 501–504. Cambridge: Cambridge Univ. Press.