GAMES AND ECONOMIC BEHAVIOR **23,** 97–105 (1998)
ARTICLE NO. GA970605

NOTE

On the Centipede Game

Robert J. Aumann*

Institute of Mathematics and Center for Rationality and Interactive Decision Theory, The Hebrew University, 91904 Jerusalem, Israel

Received July 10, 1996

In Rosenthal's Centipede Game, if at the start of play it is commonly known that the players will choose rationally at vertices that are actually reached, then the backward induction outcome results; that is, the first player ''goes out'' at the first move. *Journal of Economic Literature* Classification Numbers: C72, C73, D82.

1. INTRODUCTION

In perfect information (PI) games, common knowledge of rationality implies that the backward induction outcome is reached (Aumann, 1995; henceforth [A]). Conceptually, this result depends on the notion of *counterfactual conditional.* It is assumed common knowledge that each player i would act rationally at each of his vertices v, even when i *knows* that v will *not* be reached; in [A], we called this condition *substantive* rationality. Though it is generally acknowledged that counterfactual reasoning is inescapable in game theory, much of the discussion of [A] (e.g., Binmore, 1996) has revolved around the counterfactual nature of substantive rationality.

An alternative condition, called *material* rationality in [A], stipulates that i act rationally at those of his vertices that are actually reached. The counterfactual component of this condition is much smaller than in substantive rationality. Though the players must still consider what actions might be taken at unreached vertices, nothing is *a priori* assumed about

*Dov Samet has become an indispensable sounding board for our work in this area. Once again, we thank him heartily for his generous input. We also gratefully acknowledge research support from the Game Theory Program at SUNY–Stony Brook.

0899-8256/98 $18.00

these actions; in particular, they are not required to be rational. Unfortunately, common knowledge of material rationality does *not* in general lead to the backward induction outcome; see, e.g., Game 1 in [A].

In recent years, Rosenthal's (1982) centipede game has become a touchstone of the theory of PI games. Almost every paper on the subject mentions it, and in many it is the chief object of analysis.[1] Much of this discussion revolves around the counterfactual aspect of substantive rationality. It is therefore of some interest that in the particular case of the centipede game, common knowledge of material rationality *is* sufficient to ensure backward induction; one need *not* assume rationality at unreached vertices. The stronger, more subtle condition of substantive rationality is not needed in the centipede game.

As in [A], time plays an important role. Here, like there, "common knowledge" refers to the start of play. "Rationality" means that *when choosing*, the chooser does not know of a choice that yields him more; in [A], we called this *ex post rationality*.[2] Thus, what we show here is that *in Rosenthal's centipede game, if at the start of play there is common knowledge of ex post material rationality, then the backward induction outcome results: the first player "goes out" immediately.*

2. FORMAL STATEMENT OF THE RESULT AND OUTLINE OF THE PROOF

We freely use the terminology and notation of [A]. At first, we work with general PI games, specializing to the centipede game only afterward.

Let v be a vertex of Player i. Denote by Ω^v the event "v is reached," i.e., the set of all states ω for which the branch of the game tree determined by the strategy profile $\mathbf{s}(\omega)$ goes through v. Recall that the partition $\mathcal{H}_i$ of Ω represents i's information at the start of play. Let $\mathcal{H}_i^v$ be the join (coarsest common refinement) of $\mathcal{H}_i$ and the partition $\{\Omega^v, \sim \Omega^v\}$. Conceptually, $\mathcal{H}_i^v$ represents i's information when he learns whether v is reached; that is, the information he had at the start of play, updated by the information that v is (or isn't) reached.[3] If E is an event, $K_i^v E$ denotes the union of those elements of the partition $\mathcal{H}_i^v$ that are included in E. In words, $K_i^v E$ is the event that *i will know E when learning whether v is reached*; or for short, *i knows E at v*. The operators K_i

[1] Most recently, Binmore has used it as a prime witness in his (1996) critique of [A].

[2] Please see Sections 4b and 5 for further discussion of these concepts.

[3] If i gets other information between starting play and reaching v, then his information "at v" is finer than $\mathcal{H}_i^v$. It may be seen that our result remains true in that case (cf. [A, Sect. 5e]).

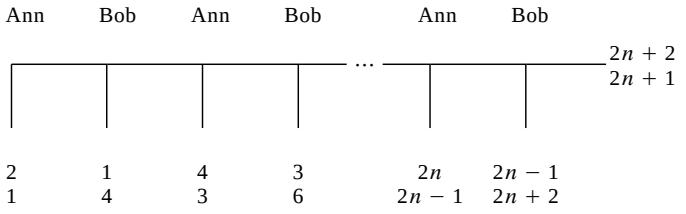

FIG. 1. The game Γ_{2n}.

continue to refer to the start of play; thus $K_i E$ is the event that i knows E at the start of play, and CKE is the event that at the start of play, E is commonly known.

In a given state ω, call *i ex post rational at v* if there is no strategy that i knows at v would have yielded him a conditional payoff at v larger than that which in fact he gets; call *i ex post materially rational* if he is ex post rational at each of his reached vertices. Denote by R_i^v the event[4] "i is ex post rational at v;" then the event "i is ex post materially rational," denoted R_i^M, is given by

$$R_i^M = \bigcap_{v \in V_i} (\sim \Omega^v \cup R_i^v), \tag{1}$$

where V_i is the set of i's vertices (for each of i's vertices v, either v is not reached, or i is ex post rational at v). The event "all players are ex post materially rational," denoted R^M, is the intersection of all the R_i^M.

Denote by Γ_{2n} the game in Fig. 1; this is one of Rosenthal's (1982) "centipede" games. Another centipede game, denoted Γ_{2n-1}, is obtained by denying Bob the option of choosing *across* at his last move.

THEOREM. *$CKR^M \subset I$ in the game Γ_r.*

In words: In Rosenthal's centipede game, if it is commonly known at the start of play that all players choose rationally at all *reached* vertices, then the first player "goes out" at the first move. Needless to say, this result holds also for any game ordinally equivalent to Γ_r.

The idea of the proof is as follows: Let m be the last vertex that is reached at any state in CKR^M; thus in some state ω in R^M, the vertex m is reached, and it is commonly known that no vertex beyond m is reached. If m is the first move in the game, we are finished. Otherwise, suppose w.l.o.g. that m belongs to Ann. Then when the vertex just before m is reached in ω, Bob knows that he can improve his payoff by going *down* rather than *across*; this contradicts his ex post material rationality.

A formal proof is given in the next section.

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

3. PROOF

We begin with some preliminaries. Recall that $h_i^v(s)$ denotes i's conditional payoff at the vertex v for the strategy profile s. Denoting the initial vertex of the game tree by v^1, define $h_i := h_i^{v^1}$; that is, $h_i(s)$ is i's actual payoff if the strategy profile s is played. Recall that $\mathbf{s}_i(\omega)$ is i's strategy in state ω, and $\mathbf{s}(\omega) := (\mathbf{s}_1(\omega), \ldots, \mathbf{s}_n(\omega))$ is the profile of all the players' strategies in that state. Functions defined on Ω (like $\mathbf{s}$ or $\mathbf{s}_i$), which appear in boldface, 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 $\mathbf{s}_i(\omega) = s_i$ holds); $[h_i^v(\mathbf{s}; t_i) > h_i^v(\mathbf{s})]$ is the event that i's conditional payoff at v would have been higher if he had chosen the strategy t_i rather than what he did choose; and $[h_i(\mathbf{s}) = 3]$ is the event that i's actual payoff is 3. As in [A], we assume that if i chooses the strategy s_i, then he knows that he chooses it; in symbols,

$$[\mathbf{s}_i = s_i] \subset K_i[\mathbf{s}_i = s_i] \qquad \text{for all strategies } s_i \text{ of } i. \tag{2}$$

The event "i is ex post rational at v" is given by

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

where S_i is the set of i's strategies (for each of his strategies t_i, it is not the case that i knows at v that t_i would yield him a higher conditional payoff at v than the strategy he chose).

The following four lemmas from [A] will be needed:

LEMMA 4. *$CKE = K_i CKE$.*

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

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

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

These lemmas are standard in knowledge theory, and also easily proved directly. They yield:

LEMMA 8. *If a^v is an action of i at v, then $[\mathbf{s}^v = a^v] \subset K_i[\mathbf{s}^v = a^v]$ (if i chooses a^v at v, then he knows that he does).*

Proof. Let A_i^v be the set of those strategies s_i of i for which $s_i^v = a^v$. Then by (2) and Lemma 5,

$$[\mathbf{s}_i^v = a^v] = \bigcup_{s_i \in A_i^v} [\mathbf{s}_i = s_i] \subset \bigcup_{s_i \in A_i^v} K_i[\mathbf{s}_i = s_i]$$
$$\subset \bigcup_{s_i \in A_i^v} K_i[\mathbf{s}_i^v = a^v] = K_i[\mathbf{s}_i^v = a^v]. \quad \text{☺}$$

Finally, note that if v is a vertex of i, then

$$K_i(E) \cap \Omega^v \subset K_i^v(E \cap \Omega^v). \tag{9}$$

Formally, this follows without difficulty from the definitions. Intuitively, if v is reached, then at v, Player i knows this, as well as what he knew at the beginning of play.

So much for the preliminaries, which apply to all PI games. We come now to the proof itself, which, of course, refers to the centipede game only. Denote the vertices $1, 2, \ldots, r$, and let m be the last vertex that is reached in any state in CKR^M; thus

$$CKR^M \subset \sim \Omega^{m+1}, \tag{10}$$

and there is a state ω with

$$\omega \in \Omega^m \tag{11}$$

and

$$\omega \in CKR^M. \tag{12}$$

If $m = 1$, we are finished. Assume therefore that $m > 1$. Let m belong to Ann, say $m = 2k - 1 \geq 3$. From (12), Lemma 4, (10), and Lemma 5 we get

$$\omega \in CKR^M = K_B CKR^M \subset K_B \sim \Omega^{m+1}, \tag{13}$$

where B stands for Bob. By the definition of Ω^j, we have $\Omega^m \subset \Omega^{m-1}$, so by (11),

$$\omega \in \Omega^{m-1}. \tag{14}$$

By (11) and Lemma 8, $\omega \in [\mathbf{s}_B^{m-1} = across] = K_B[\mathbf{s}_B^{m-1} = across]$; from this, (13), and Lemma 6, we get $\omega \in K_B([\mathbf{s}_B^{m-1} = across] \setminus \Omega^{m+1})$. So by

(14), (9), and the rules of the game,

$$\omega \in K_B^{m-1}\big(\big[\mathbf{s}_B^{m-1} = \textit{across}\big] \cap \big(\Omega^{m-1} \setminus \Omega^{m+1}\big)\big) = K_B^{m-1}\big(\Omega^m \setminus \Omega^{m+1}\big)$$
$$= K_B^{m-1}\big[h_B(\mathbf{s}) = 2k-1\big]. \tag{15}$$

In particular, applying Lemma 5 to K_B^{m-1},

$$\omega \in K_B^{m-1}\big[\mathbf{s}_A^j = \textit{across} \text{ when } j = 1, 3, \ldots, m-2\big], \tag{16}$$

where A stands for Ann. Now let t_B be a strategy of Bob that chooses *across* at all vertices of Bob before $m-1$, and chooses *down* at $m-1$. Then

$$\big[\mathbf{s}_A^j = \textit{across} \text{ for } j = 1, 3, \ldots, m-2\big] \subset \big[h_B(\mathbf{s}; t_B) = 2k\big];$$

so from (16), $\omega \in K_B^{m-1}[h_B(\mathbf{s}; t_B) = 2k]$; so from (15) and Lemmas 6 and 5 applied to K_B^{m-1},

$$\omega \in K_B^{m-1}\big(\big[h_B(\mathbf{s}; t_B) = 2k\big] \cap \big[h_B(\mathbf{s}) = 2k-1\big]\big)$$
$$\subset K_B^{m-1}\big[h_B(\mathbf{s}; t_B) > h_B(\mathbf{s})\big]. \tag{17}$$

By (12), Lemma 7, (1), $K_i^v \Omega^v = \Omega^v$, (3), and Lemma 6,

$$\omega \in CKR^M \subset R^M \subset R_B^M \subset \sim \Omega^{m-1} \cup R_B^{m-1}$$
$$\subset \sim K_B^{m-1}\Omega^{m-1} \cup \sim K_B^{m-1}\big[h_B^{m-1}(\mathbf{s}; t_B) > h_B^{m-1}(\mathbf{s})\big]$$
$$= \sim K_B^{m-1}\big(\Omega^{m-1} \cap \big[h_B^{m-1}(\mathbf{s}; t_B) > h_B^{m-1}(\mathbf{s})\big]\big). \tag{18}$$

Since $m-1$ is reached in each state in Ω^{m-1}, it follows from the definition of h_B that

$$h_B^{m-1}(\mathbf{s}(\nu); t_B) = h_B\big(\mathbf{s}(\nu); t_B\big) \quad \text{and} \quad h_B^{m-1}(\mathbf{s}(\nu)) = h_B(\mathbf{s}(\nu))$$

for each ν in Ω^{m-1}. Therefore

$$\Omega^{m-1} \cap \big[h_B^{m-1}(\mathbf{s}; t_B) > h_B^{m-1}(\mathbf{s})\big] = \Omega^{m-1} \cap \big[h_B(\mathbf{s}; t_B) > h_B(\mathbf{s})\big];$$

so (18), Lemma 6, and $K_i^v \Omega^v = \Omega^v$ yield

$$\omega \in \sim K_B^{m-1}\big(\Omega^{m-1} \cap \big[h_B(\mathbf{s}; t_B) > h_B(\mathbf{s})\big]\big)$$
$$= \sim K_B^{m-1}(\Omega^{m-1}) \cup \sim K_B^{m-1}\big[h_B(\mathbf{s}; t_B) > h_B(\mathbf{s})\big]$$
$$= \sim \Omega^{m-1} \cup \sim K_B^{m-1}\big[h_B(\mathbf{s}; t_B) > h_B(\mathbf{s})\big],$$

which is inconsistent with (14) and (17). This completes the proof when m belongs to Ann; when it belongs to Bob, the proof is similar. ☺

4. DISCUSSION

a. Subjunctives

Much of the discussion of the centipede game has centered around the following argument: On the one hand, we are told that under common knowledge of rationality (*CKR*), Ann *must* go out at her first move. On the other hand, the backward induction argument for this is based on what the players *would* do if Ann stayed in. But, if she did stay in, then *CKR* is violated, so the argument that she will go out no longer has a basis.

We have argued elsewhere ([A]; Aumann, 1996) that the above argument is unsound. Be that as it may, this note shows that the whole issue may be circumvented. It is not necessary to use the subjunctive mood; the proof in this note refers only to rationality at vertices that are actually reached, not to whether players ''would'' play rationally if their vertices ''were'' reached. Rather than reasoning from what happens after a given vertex, it reasons from what happens *before*: If some vertex is the last that can possibly be reached, then already the one before it should have been the last. This is quite different from the backward induction proof in [A].

b. Time

The formal contrast between the ex post notion of rationality and the ex ante notion of common knowledge may disturb some readers, but closer examination reveals no awkwardness. Rationality is *inherently* ex post: Whether or not a given choice is rational necessarily depends on the information available at the time of the choice. This has nothing to do with the *knowledge*—or common knowledge—of such rationality, which may well refer to an entirely different time. It makes perfect sense to speak of common knowledge at the *start of play* that the players will choose rationally if and when one of their vertices is actually reached. That, precisely, is the hypothesis of this note.

c. Probability

Our theorem carries over without change to probabilistic models, in which play is defined as rational if and only if it maximizes expected utility (caution: this applies to rationality only; *knowledge*—and common knowl-

edge—must still be defined in terms of absolute certainty, not probability 1 belief). As above, the expectations must be ex post. Compare Aumann (1996), Sections 1 and 9.

d. The Proof

The reader may wonder why we adduce such a lengthy formal proof for an argument that while not immediate, seems simple enough once one has found it. The reason is that this area is *very* tricky, and unless one is extremely careful and formal, it is easy to go astray—as we have found, to our dismay, on more than one occasion. (The same remark applies to [A].)

5. PREVIOUS WORK

This section, which discusses technical relations between the current note and [A], may be omitted without affecting the understanding of the rest of the note.

''Common knowledge'' (*CK*) here is precisely the same as in [A]. ''Ex post rationality'' here is the same as ''rationality'' in [A], except that the ex post knowledge operators K_i^v appearing in the definition (3) of ex post rationality replace the ex ante knowledge operators K_i appearing in the definition [A, (3)] of rationality. Thus the terminology here is fully coordinated with that of [A]. To avoid confusion, we henceforth call the rationality concept of [A] by the name ''ex ante rationality.''

Ex post rationality is stronger than (i.e., implies) ex ante rationality [A, Sect. 5e], so *CK* of ex post rationality implies *CK* of ex ante rationality. [A] shows that *CK* of ex ante substantive rationality implies backward induction; therefore, *CK* of ex post substantive rationality also does. Thus, though [A] formally uses the less intuitive ex ante version of rationality, the same result with the more intuitive ex post version follows easily.

The reason that [A] uses the less intuitive ex ante version is that it is simpler than the ex post version—that it requires less baggage, both formal and conceptual. Specifically, the ex ante version works directly with the same partitions $\mathcal{H}_i$ that are used to define *CK*, rather than with the ex post partitions $\mathcal{H}_i^v$. Since the ex ante result is logically stronger, nothing is lost.

Why, then, don't we use the simpler ex ante version here too? The reason is that here, unlike in [A], the ex ante version won't do; *CK* of ex ante material rationality does *not* yield backward induction in the centipede game. To show this, we now adduce an example of a centipede game in which there is common knowledge of material rationality—in the

ex ante sense of [A], but not the ex post sense used here—and the first player does *not* ''go out'' at the first vertex.

Recall that a player is *ex ante materially rational* if he is ex ante rational at each of his reached vertices. Denote by R^{AM} the event ''all players are ex ante materially rational;'' in symbols,

$$R^{AM} = \bigcap_i \left(\bigcap_{v \in V_i} \left(\sim \Omega^v \cup \bigcap_{t_i \in S_i} \left(\sim K_i \left[h_i^v(\mathbf{s}; t_i) > h_i^v(\mathbf{s}) \right] \right) \right) \right).$$

Consider now the game Γ_3, with a state space Ω consisting of three states, α, β, γ. Ann's partition at the beginning of play is $(\alpha, \beta\gamma)$, whereas Bob's is $(\alpha\beta, \gamma)$. Ann's strategy is (*across*, *down*) in state α, (*down*, *across*) in states β and γ. Bob's strategy is *across* in states α and β, *down* in state γ. It may be seen that both players are (ex ante) materially rational at all three states, so all states are in CKR^{AM}, but in state α, the players play *across* at the first two moves.

Bob's behavior in state α may seem strange, since he knows that if his vertex is reached, Ann will go *down* at her next vertex, so it would be advantageous for him to go *down* rather than *across*. But, at the start of play, Bob does not know whether the state is α or β. If it is α, he can improve his conditional payoff at his vertex by switching to *down*; this translates to a gain in actual payoff, because in state α, his vertex is actually reached. If the state is β, his conditional payoff is larger if he goes *across* than if he goes *down*, so he ''loses'' by switching. But, this ''loss'' is only in the conditional payoff, not the actual payoff, because in state β, his vertex is not reached. And he knows this, already at the start of play! Nevertheless, the fact remains that at the start of play he is not sure that he can improve his conditional payoff at his vertex, so by our definition, he is rational. This quirk in the definition of ex ante material rationality provides additional motivation for the ex post notion used here.

REFERENCES

Aumann, R. J. (1995). ''Backward Induction and Common Knowledge of Rationality,'' *Games Econ. Behav.* **8**, 6–19 **([A] in the text)**.

Aumann, R. J. (1996). ''Reply to Binmore,'' *Games Econ. Behav.* **17**, 138–146.

Binmore, K. (1996). ''A Note on Backward Induction,'' *Games Econ. Behav.* **17**, 135–137.

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