GAMES AND ECONOMIC BEHAVIOR **17,** 138–146 (1996)
ARTICLE NO. 0099

Reply to Binmore

''Go to Father,'' she said,
When he asked her to wed,
For she knew that he knew that her father was dead,
And she knew that he knew what a life he had led,
So she knew that he knew what she meant when she said,
''Go to Father.''

—*Folk ditty*

Though we found it difficult to follow Binmore's (1996) critique (henceforth [B]) of Aumann (1995) (henceforth [A]), we acknowledge that [A] is not transparent. Here we reply to [B], and also explain the analysis of [A] in verbal, conceptual terms, covering some points not covered in [A].

Section 1 responds to Binmore's criticism of our definition of rationality. Sections 2 and 3 clear up two fundamental misunderstandings in [B]. Sections 4 through 7 discuss the proof of our Theorem A; Section 6 makes a point of general methodological interest; Section 8 discusses the proof of Theorem B. Section 9 reformulates our results in terms of ordinary Bayesian rationality (rather than the weaker form of rationality used in [A]).

1. [B]'s primary point concerns our definition of ''rationality''. Roughly, this specifies that a player i is rational if and only if it is not the case that he knows that he would be able to do better (for the precise definition, please see [A], p. 9).

Binmore questions the ordering of the words in this definition. He suggests two different emendations. In the first, i is rational if and only if he knows that he would not be able to do better. The second calls for a standard Bayesian framework, with utilities and probabilities; this enables us to apply the standard definition, according to which a player is rational if and only if he would be unable to increase his expected utility.

Leaving aside the merits of these two definitions, let us say at once that *with either one*, *both theorems of* [A] *remain true without any change*.

To see why, call the three forms of rationality, respectively, [A]-rationality, [B]-rationality, and Bayes rationality. Of the three, [A]-rationality is the weakest (easiest to achieve) and [B]-rationality the strongest, while Bayes rationality is between the two. That is to say, every [B]-rational choice is Bayes rational, and every Bayes rational choice is [A]-rational. Therefore, common knowledge (CK) of [B]-rationality implies CK of [A]-rationality; and, also, CK of Bayes rationality implies CK of [A]-rationality. But by Theorem A of [A], CK of [A]-rationality implies backward induction.

0899-8256/96 $18.00

Therefore *CK* of either of the other two forms of rationality implies backward induction (BI). In the other direction, one may verify directly, as in [A], that *CK* of [B]-rationality (and a fortiori of Bayes rationality) is possible in every perfect information game.

We agree with Binmore that Bayes rationality is the most natural and straightforward of the three. We nevertheless favored [A]-rationality, for several reasons. One is that it yields a stronger result; that is, the result with [A]-rationality implies that with Bayes rationality, but not the other way around. Another is that this strength is obtained at no extra cost. In fact it is cheaper, because working with Bayes rationality requires the explicit introduction of probabilities, which just complicate the system without serving any essential purpose. Finally, justifying the standard Bayesian framework requires an axiom system such as that of Savage (1954). While we ourselves have no problem with this, there are others who do; so we felt it preferable to avoid implicitly assuming axioms that really have nothing to do with the matter at hand.

In brief, our result says that *CK* of an extremely weak form of rationality already implies BI; a fortiori this is so for stronger forms of rationality, including the standard Bayesian form.

The gist of this discussion appears already[1] in [A, 4c and Footnote 4]. To avoid future misunderstandings, we present a formal treatment in Section 9 below.

2. On a more conceptual level: [B] opens with the sentence, ''It now seems to be generally accepted that rational players would not necessarily use their backward induction strategies if there were to be a deviation from the backward induction path.'' Apparently, Binmore sees some kind of inconsistency or contradiction between this and the results of [A].

In fact, there is no inconsistency. We agree wholeheartedly with [B]'s first sentence. Indeed, we go further: even if there has been *no* deviation from the backward induction path up to some point, a rational player may well deviate at that point. A rational player may even deviate from the backward induction path at the very first move of the game. We have said this repeatedly. Thus [A, p. 18]: ''. . . the inductive choice may be not only unreasonable and unwise, but quite simply irrational;'' and Aumann (1992) shows that in the centipede game adduced in [B], it may be *incumbent* on rational players to ''stay in'' until quite late in the game.

Our results concern a situation with *CK* of rationality, *not* just rationality. Binmore continuously confounds these concepts, using them almost

[1] Footnote 3 of [B] acknowledges this, but dismisses it for reasons that we do not understand.

interchangeably. Here and there he pays lip service to *CK* of rationality (*CKR*); but substantively he argues as if only simple rationality were assumed. Needless to say, our argument will not work in such an environment. We do make essential use of *CKR*.

3. Another error of Binmore is his failure to distinguish between the subjunctive and indicative moods. In [A, 5f], we wrote

> "The results of this paper ([A]) say nothing about the behavior of players at vertices that are off the backward induction path and are actually reached." (*)

[B] takes this to mean that nothing can be said about what players *would* do if off-path vertices *were* to be reached; he concludes that we "deny . . . that the rationality of choosing *down* need involve any knowledge at all of what would happen if *across* were played." But we neither meant nor said that.[2] Indeed [A] heavily stresses the precise opposite: "The subjunctive mood—what (a player) *would* do, even when not given the opportunity—is of the essence" [A, 4b]; or, "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 . . . games, you must consider what other people would do if you did something different from what you actually do" [A, 5b].

These are not just empty words; they lie at the heart not only of [A]'s conceptual approach, but also of its formal treatment. A strategy of a player is defined as a function that assigns an action to *each* of his vertices, reached or not; the rationality of a player is defined in terms of his rationality at each of his vertices; and this, in turn, is defined in terms of what he knows about the others players' strategies. Thus in deciding whether a player i is rational when choosing *down* in the centipede game, we do explicitly take into account what i knows or thinks about what the

[2] Our intention could not have been made clearer. The indicative mood is heavily stressed: "are actually reached." "*Are*"; not "would", not "were to be". For the case that somebody might still misunderstand, the word "*actually*" is thrown in. And if, by some stretch of the imagination, *somebody* might *still* misunderstand, our very next sentence (not cited by Binmore) would surely clear things up: "Under *CKR*, vertices off the backward induction path *cannot* be reached; and when *CKR* does not obtain, the results do not apply."

Binmore calls (*) oxymoronic. That is beyond our comprehension. Why can't vertices off the BI path actually be reached? To be sure, they can't be reached under *CKR*, as we say explicitly in the very next sentence. But why can't they be reached when *CKR* does not obtain? And if Binmore had somehow understood that we were implicitly assuming *CKR* in (*), surely the very next sentence should have disabused him.

other player would have done if i had gone across. Binmore is 180° off the mark.

4. Though Binmore makes an unconvincing case, the malaise that he evinces is not totally groundless. We believe our analysis to be both conceptually and formally sound; but it is not straightforward, and we are grateful for this opportunity to elucidate it. This and the next section discuss Theorem A of [A], according to which *CKR* implies BI; Theorem B is treated in Section 6.

To start with, one really must differentiate very sharply between rationality and *CKR* (see Section 2 above). That a player is rational at a vertex v means that his choices at v and at his subsequent vertices maximize his expected payoff given his beliefs[3] at v. In forming his beliefs at v, the player may take into account whatever he wishes. In particular, he may take into account the actions of other players at previous moves [A, 5d]. Thus in [B]'s centipede game, Player II (P2) may, at his first vertex, play *across* for precisely the reason that Binmore adduces: that noting that Player I (P1) played *across* at *her* first vertex, he estimates that she will play *across* again. P1, in turn, may take this into account when making *her* first move, so that she may well wish to play *across* at *her* first move. In fact, it makes no difference how the players form their beliefs; as long as their actions maximize their expected payoffs, they are rational. As a result, rational players may well play *across* for a very long time in the centipede game. We have said this again and again, both here and in [A], but apparently one cannot say it often enough.

But Theorem A of [A] assumes not only that the players are rational; it assumes also that this is common knowledge. In particular, P1 knows that if P2's last vertex *were* reached, he *would* play down (subjunctive!). She knows this for *sure*, without a shadow of a doubt, because she *knows* that P2 is rational. Therefore, if P1's last vertex were reached, *she* would play down. Now P2 *knows* that P1 is rational, and he knows that she knows that he is rational. So he knows what we just concluded: that if her last vertex were reached, she would play down. He knows this for sure, without a shadow of a doubt. Therefore, if his next-to-last vertex were reached, he, being rational, would play down. And so on, until we conclude that P1 plays down at her first vertex.

The point is that while the beliefs of a rational player *might* motivate him to play *across*, they don't *have* to. The assumption of *CKR* gives

[3] The presentation here is in terms of Bayes rationality, in accordance with Sections 1 above and 9 below.

us—and the players—additional information, information that enables us to conclude that each player would go down at each of his vertices v, if that vertex were reached.

5. "But," the reader may ask, "something still bothers me. You have proved that under *CKR*, P1 must go *down* at the first vertex. You did this by working you way backward from the last vertex, showing that at each vertex v, the player at v would have to go *down* if v were reached. Very good. Having proved this, you know it, I know it, and the players know it. Now let's reexamine the proof. P1 must go *down* at her first vertex. In deciding on this, she must take the alternatives into account. You yourself stressed this, both in [A] and above (Section 3). So P1 must ask herself, what would happen if she went *across*? Well, we know that under *CKR*, she *can't* go *across*; we've proved that. So if she *did* go *across*, she would be demonstrating that *CKR* does not obtain. It would then be illegitimate to use the conclusions of *CKR* also insofar as they apply to the second vertex; that is, it would be illegitimate to conclude that P2 would necessarily go *down* at the second vertex. So, he might go *across*. In that case, P1 would prefer to go *across* at the first vertex. So the proof that *CKR* implies *down* at the first vertex, which looked good at first, breaks down on reexamination. It carries within it the seeds of its own destruction."

6. Before responding to this substantively, we make a methodological point. It is difficult to evaluate the validity of this kind of contorted reasoning using verbal tools only. That is a function of mathematical formalisms. In a formal model the conclusions are derived from definitions and assumptions. Once one is satisfied that the derivation is mathematically correct, it remains only to examine the appropriateness of the definitions and assumptions. But with informal, verbal reasoning as complex as the above, one never knows for sure whether the argument is sound. One can argue until one is blue in the face, without convincing one another, because there is no criterion for deciding the soundness of an informal argument.

So we say, gentle reader, it is indeed possible that the conclusions of [A] are unsound; but if so, that can only be for one of two reasons: either there is a mathematical error in our proof, or one of our definitions or assumptions is conceptually inappropriate. No one has challenged the correctness of our mathematics. [B] does challenge the appropriateness of our definition of rationality; we respond in Sections 1 above and 9 below. We welcome other challenges to our definitions or assumptions. But the kind of verbal argument typified by most of [B], and by Section 5 above, concerns the reasoning process, the process of drawing conclusions from assumptions; and in a matter of this complexity, that is better left to the mathematics.

7. Nevertheless, we now respond to the question in Section 5 within the same informal genre. The error in the argument is that it mixes the conclusion of the proof[4] into the proof itself. That is not legitimate. Suppose we have proved a theorem of the form "p implies q." But our hypothetical reader is skeptical. "The proof sounds right," he says, "but let's look again. Assume p. Perhaps, after all, this could jibe with 'not q.' So suppose that it does—i.e., that q does not obtain. But then, since we have proved that p implies q, it cannot be that p obtains. So we conclude that after all, p doesn't obtain. But if p doesn't obtain, we no longer have grounds for concluding q. So your proof doesn't hold up under examination—you'll have to abandon it."

Clearly, this argument is absurd.

But that is exactly the argument in Section 5. Starting with *CKR*, we prove that P1 goes down at the first vertex. Now, we say, let's try it again. Must P1 *really* go down at the first vertex? Let's suppose not—i.e., that she goes across. But then we have a contradiction to *CKR*, and anything whatever follows from a contradiction! So we must abandon *CKR*. But then there is no longer any reason to go down at the first vertex.

Just as clearly, *this* argument is absurd.

8. We come now to Theorem B, which says that *CKR* is possible in every perfect information game. Here we do *not* assume *CKR*; we want to prove that it is possible, for appropriate choices of moves and information. The proof is simple: just let each player make his inductive choice at each of his vertices v, if v is reached, and stipulate that this be commonly known. It is easily verified that *CKR* then indeed obtains.

In the centipede game, for example, we stipulate that each player play *down* at each vertex, if reached, and that this be commonly known. *CKR* then *follows*.

9. As promised in Section 1, we now provide a formal account of the results of [A] using Bayes rationality rather than the very weak version of rationality used in [A] (called [A]-rationality in Section 1). We freely use the terminology, notation, and results of [A].

Start with a knowledge system $\{\Omega, \mathbf{s}, \{\mathcal{K}_i\}_i\}$ as in [A, 2], where i ranges over the players. For each vertex v and strategy n-tuple s, set $s^{\geq v} := (s^v, s^{>v})$; thus $s^{\geq v}$ is the profile of actions prescribed by s at v and at subsequent vertices. Denote by $\mathcal{F}^v$ the field of events generated by the "random variable[5]" $\mathbf{s}^{\geq v}$; that is, the smallest field with respect to which the function $\mathbf{s}^{\geq v}$ is measurable (note that this field is finite). In words, an

[4] That vertices off the backward induction path cannot be reached under *CKR*.

[5] Recall that $\mathbf{s}(\omega)$ is the n-tuple of the players' strategies in the state ω.

event is in $\mathscr{F}^v$ if and only if it is describable in terms of the actions taken at v and at subsequent vertices; the events in $\mathscr{F}^v$ describe what might happen if v would be reached.

Now define a *knowledge-belief system* to consist of a knowledge system as above, and for each player i, each of i's vertices v, and each state ω, a probability measure $\pi_i^v(\cdot;\omega)$ on $\mathscr{F}^v$; $\pi_i^v(E;\omega)$ signifies i's probability for E at v in the state ω. Set $B_i^v E := \{\omega: \pi_i^v(E;\omega) = 1\}$; in words, $B_i^v E$ is the event that i believes E—that is, ascribes probability 1 to E—at v. Assume that

$$K_i E \subset B_i^v E \tag{1}$$

for all E in $\mathscr{F}^v$ and all vertices v of i; this says that if before the beginning of play, i knows something about the actions that would be taken if v were reached, then he assigns it probability 1 at v. In a given state ω of the world, call i *Bayes rational* if

$$\mathrm{Exp}_{i,v,\omega} h_i^v(\mathbf{s}) \geq \mathrm{Exp}_{i,v,\omega} h_i^v(\mathbf{s}; t_i) \tag{2}$$

for each of i's vertices v and strategies t_i, where $\mathrm{Exp}_{i,v,\omega}$ denotes the expectation with respect to the probability measure $\pi_i^v(\cdot;\omega)$; since h_i^v is defined in terms of what happens starting at v only, $h_i^v(\mathbf{s})$ and $h_i^v(\mathbf{s}; t_i)$ are $\mathscr{F}^v$-measurable, so the expectations are defined.

In words, (2) says that i's strategy at ω maximizes his expected conditional payoff at v. Denote by R^B the event that all players are Bayes rational.[6]

THEOREM A. $CKR^B \subset I$.

THEOREM B. *For every PI game, there is a knowledge-belief system with* $\varnothing \neq CKR^B$.

In words, Theorem A says that if Bayes rationality is commonly known, the inductive outcome results; Theorem B, that common knowledge of Bayes rationality is indeed possible in every PI game.

Denote by R_i^B the event "i is Bayes rational;" thus R^B is the intersection of the R_i^B. Recall that R_i is the event "i is rational" in the sense of [A], and that the intersection of the R_i—i.e., the event that all players are rational—is denoted R.

LEMMA. $R_i^B \subset R_i$.

[6] I.e., the set of all ω such that all players are Bayes rational in the state ω.

In words, if a player is Bayes rational, then he is rational in the sense of [A].

Proof. Let $\omega \in R_i^B$. Then for each vertex v and strategy t_i of i, we have (2); that is, in ω, i's expectation at v of $h_i^v(\mathbf{s}; t_i)$ is not greater than his expectation at v of $h_i^v(\mathbf{s})$. So it cannot be that in ω, Player i assigns probability 1 to the event $[h_i^v(\mathbf{s}; t_i) > h_i^v(\mathbf{s})]$; in symbols, $\omega \in \sim B_i^v[h_i^v(\mathbf{s}; t_i) > h_i^v(\mathbf{s})]$. So by (1), $\omega \in \sim K_i[h_i^v(\mathbf{s}; t_i) > h_i^v(\mathbf{s})]$. Since this holds for all v and t_i, it follows that $\omega \in \bigcap_{v \in V_i} \bigcap_{t_i \in S_i} \sim K_i[h_i^v(\mathbf{s}; t_i) > h_i^v(\mathbf{s})] = R_i$, by (3) of [A]. ☺

Proof of Theorem A. From the lemma and the definitions of R and R^B, we get $R^B \subset R$. Now it is known that for any events E and F, if $E \subset F$, then $CKE \subset CKF$. So $CKR^B \subset CKR$. But $CKR \subset I$, by Theorem A of [A]. So $CKR^B \subset I$. ☺

Proof of Theorem B. Define a knowledge-belief system by letting Ω consist of a single state ω, in which each player makes his inductive choice at each of his vertices. Then $\omega \in CKR^B$. ☺

A final note[7]: Bayes rationality is here defined in ''ex-post'' terms—what i would think[8] if v were reached, rather than in ''ex ante'' terms—what he thinks at the beginning of the game. The formal development in [A] uses the ex ante definition, because in the context of [A], it is weaker,[9] and so yields a stronger version[10] of Theorem A; see [A, 5e]. But in the current Bayesian context, ex ante rationality is neither weaker nor stronger than (nor equivalent to) ex post rationality. As explained in [A, 5e], ex post rationality seems more relevant than ex ante rationality. Therefore, since in the current context, the ex post result does *not* follow from the ex ante one, use of the ex post definition is indicated here. We stress, though, that this applies to rationality as such only; ''knowledge'' and ''common knowledge'' remain ex ante, i.e., refer to the beginning of the game.

[7] This is a technical note, which may be ignored without affecting the understanding of the rest of the paper.

[8] The term ''think'' is meant to encompass both ''know'' and ''attribute probability.''

[9] That is, ex post [A]-rationality implies ex ante [A]-rationality.

[10] A weaker hypothesis means a stronger result.

REFERENCES

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

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, MA: MIT Press.

Binmore, K. (1996). "A Note on Backward Induction," *Games and Econ. Behav.* **17**, 135–137 **([B] in the text).**

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

*Robert J. Aumann**

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

* Research support from the Game Theory Program at SUNY-Stony Brook is gratefully acknowledged. Also, we thank Dov Samet for many helpful conversations about this work. The folk ditty was communicated to us by Miriam Aumann (1892–1961).

Statement of ownership, management, and circulation required by the Act of October 23, 1962, Section 4369, Title 39, United States Code: of

GAMES AND ECONOMIC BEHAVIOR

Published monthly by Academic Press, Inc., 6277 Sea Harbor Drive, Orlando, FL 32887-4900. Number of issues published annually: 12. Editor: Dr. Ehud Kalai, MEDS Dept., J. L. Kellogg Graduate School of Management, Northwestern University, Leverone Hall, 2001 Sheridan Road, Evanston, IL 60208.
Owned by Academic Press, Inc., 525 B Street, Suite 1900, San Diego, CA 92101-4495. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages, and other securities: None.
Paragraphs 2 and 3 include, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting, also the statements in the two paragraphs show the affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner. Names and addresses of individuals who are stockholders of a corporation which itself is a stockholder or holder of bonds, mortgages, or other securities of the publishing corporation have been included in paragraphs 2 and 3 when the interests of such individuals are equivalent to 1 percent or more of the total amount of the stock or securities of the publishing corporation.
Total no. copies printed: average no. copies each issue during preceding 12 months: 1048; single issue nearest to filing date: 1050. Paid circulation (a) to term subscribers by mail, carrier delivery, or by other means: average no. copies each issue during preceding 12 months: 228; single issue nearest to filing date: 228. (b) Sales through agents, news dealers, or otherwise: average no. copies each issue during preceding 12 months: 301; single issue nearest to filing date: 313. Free distribution (a) by mail: average no. copies each issue during preceding 12 months: 64; single issue nearest to filing date: 64. (b) Outside the mail: average no. copies each issue during preceding 12 months: 24; single issue nearest to filing date: 24. Total no. of copies distributed: average no. copies each issue during preceding 12 months: 617; single issue nearest to filing date: 629. Percent paid and/or requested circulation: average percent each issue during preceding 12 months: 86%; single issue nearest to filing date: 86%.

(Signed) Janice M. Peterson, Director, Fulfillment and Special Projects