Rationality in the Centipede[1]

Ken Binmore
Economics Department
University College London
Gower Street
London WC1E 6BT, UK

Abstract

The literature on refinements of Nash equilibirum is now generally acknowledged to have failed in its task of providing a viable equilibrium selection theory for games. Part of the reason for its failure lies in its shaky foundations as a theory of rational behaviour. Backward induction is a particularly doubtful rationality principle. This paper uses the Centipede Game as a setting within which to explain why Aumann's recent attempt to rehabilitate backward induction is unsuccessful.

1 Introduction

Rational belief and rational action are usually studied axiomatically. Axioms are propounded and the properties of rational individuals are then deduced mathematically. If the necessary mathematics is sufficiently challenging, attention then concentrates on whether an author's theorems are true rather than the more fundamental question of whether the axioms are successful in formalizing the concepts they are intended to capture. In spite its widespread acceptance, I therefore believe that much orthodox work on the foundations of game theory is at best doubtful. In this paper, I plan to use the notion of backward induction as a case study to illustrate this contention.

Backward induction is widely regarded as one of the cornerstones of game theory. Zermelo used it to prove one of the first theorems of game

[1]The support of National Science Foundation grant SES 9122176 is gratefully acknowledged.

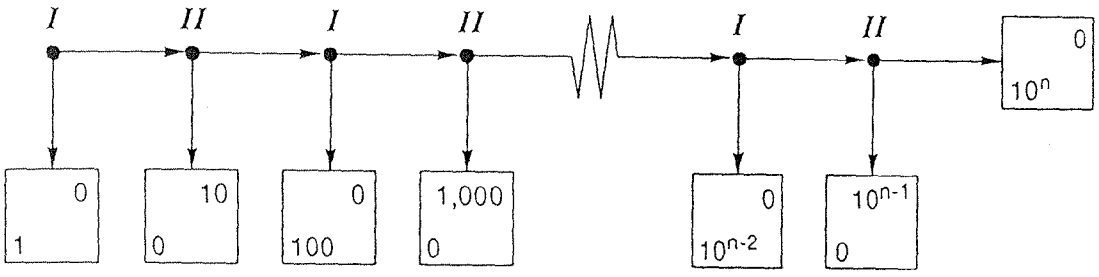

Figure 1: The Centipede Game

theory—namely that Chess has a solution. Selten [14] made it the basis of his notion of subgame-perfect equilibrium, since which time it has been incorporated into most of the many refinements of Nash equilibrium that have been proposed. Indeed, Kohlberg and Mertens [9] make it a *sine qua non* for a refinement concept. It is therefore surprising that only recently have attempts been made to deduce the validity of backward induction from more primitive rationality principles. The difficulties encountered in such attempts have led Binmore [5], Reny [11], Bicchieri [3,4] and others to question whether backward induction is indeed compelling as a rationality principle. However, Aumann [1] has recently sprung to its defense by offering an argument that claims to deduce the backward induction solution in finite games of perfect information from the hypothesis that the rationality of the players is common knowledge before the game begins.

Notice that the use of backward induction in the Centipede Game of Figure 1 requires that the opening action be *down*. Similar counter-intuitive conclusions are obtained by applying backward induction in the finitely repeated Prisoners' Dilemma or in Selten's [13] chain-store paradox.

This paper explains why I believe that Aumann is mistaken in arguing that prior common knowledge of rationality implies backward induction. I believe that the object lesson is that intuition is inadequate as a source of inspiration when rationality axioms are being invented. Where then should

we look for inspiration?

I think that the answer lies in taking the evolutionary approach to game theory seriously. It is true that attacking the problem in this way will at best lead to the conclusion that some types of bounded rationality are more succesful than others. But, as I have argued elsewhere (Binmore [5]), this is perhaps the best for which one can reasonably hope. However, if one does adopt an evolutionary approach, the outlook for backward induction is bleak indeed.

Evolution and backward induction As Binmore, Gale and Samuelson [6] show, one cannot count on evolution taking a population to the subgame-perfect equilibrium, even in Selten's one-shot Chain-Store Game—which is the standard example employed in textbooks when justifying the concept. Instead, the evolutionary dynamics studied often take the system to a Nash equilibrium of the game which is not subgame-perfect. At the subgame-perfect equilibrium, a potential competitor to an established chain-store sets up a rival store because he predicts that the chain-store will not respond by initiating a mutually damaging price war. At the alternative Nash equilibrium sometimes selected by evolution, the potential entrant stays out. But if he were to enter, he would find that sometimes the chain-store would respond by fighting and sometimes it would acquiesce in splitting the market.

Computer simulations lead us to similar results in the Ultimatum Game. In this simple bargaining game, player I and player II have a sum of mnoney to divide if they can agree on how it should be split. The rules of the game specify that player I makes a take-it-or-leave-it offer to player II who can then accept or refuse. If she refuses, both get nothing. Backward induction says that player I should propose a split in which player II gets at most one penny, and that she should accept. However, the much replicated experiment of Güth, Schmittberger and Schwarze [8] demonstrates that real people do not use this subgame-perfect equilibrium. Instead, player I tends to offer player II something between $\frac{1}{2}$ and $\frac{1}{3}$ of what is available. If he offers less than $\frac{1}{3}$, he has a probability of about $\frac{1}{2}$ of being turned down. Some authors deduce from this conclusion that game theory is irrelevant to bargaining because agents are motivated by fairness norms rather than strategic considerations. Orthodox game theorists prefer to argue that the

subjects in such experiments have exotic utility functions which incorporate factors other than money. However, the evolutionary simulations of Binmore, Gale and Samuelson [6] show that, amongst the many Nash equilibria of the game which that are not subgame-perfect, there is a particular equilibrium with a huge basin of attraction relative to the dynamics studied. At this equilibrium, player I offers player II about $\frac{1}{4}$ of the money available and player II accepts with a probability that is just high enough to make the offer worthwhile.

Space precludes further discussion of such evolutionary results, although I hope that the lecture for which this short paper has been written will be more forthcoming. The results are mentioned only to make the point that backward induction is a very dubious principle for applied purposes—even in the simplest of games. On this point, Aumann [2] and I have no differences. As he frequently insists, there are few real-world situations for which it makes sense to postulate commmon knowledge of rationality. However, I do not believe that backward induction is justified even when it does make sense to postulate that there is prior common knowledge of rationality.

2 Prior Common Knowledge of Rationality

Aumann's [1] paper has been through several drafts. Binmore and Samuelson [7] criticized the original draft, which was presented at the recent Nobel Symposium in Björkborn, Sweden. Although Aumann's paper has now changed, I do not want to bring our critique up to date because I have found nobody who thinks it reasonable to follow Aumann in regarding it as reasonable to model a player as a collection of independently acting agents—one for each node at which the player makes a decision. However, later versions of Aumann's paper contain a new argument which does not use this hypothesis but still claims that prior common knowledge of rationality implies that the opening move in the Centipede Game will necessarily be *down*.

I agree with Aumann that his argument does indeed demonstrate that a 'rational player' will begin the Centipede Game by playing down under the conditions he specifies. But I shall then ask whether it is really rational to be a 'rational player'. The issues depend to a substantial degree on

how the subjunctive conditionals which appear in the arguments should be interpreted. Binmore and Samuelson [7] go on at some length on this subject, but here space will preclude all but a bare-bones account.

It will be taken for granted that there is always common knowledge of the game being played, so that attention can be concentrated on what players know or do not know about their opponents. A player's knowledge of the characteristics of his opponents (including what what the opponents know or do not know about him) is not usually accorded a formal role in game-theoretic analyses. Usually, the assumptions being made about what players know about each other are implicit in the equilibrium concept that an analyst chooses to consider. However, in what follows, I shall assume that part of our enterprise is to label each node x in the Centipede Game with a pair (S,T) of sets. The interpretation is that, if node x were to be reached, then it would be common knowledge that player I lies in set S and player II lies in set T. With such a convention, Aumann's assumption of common knowledge of rationality can be expressed by labeling the first node with a pair (R_1,R_2), where both R_1 and R_2 contain only 'rational players'.

During the course of a game, the actions that a player takes will enrich the information about his characteristics available to his opponent. Suppose that node x is labeled with the pair (S,T). Suppose also that, if player I were to take action a at node x, then the next node would be y. Finally, suppose that there is at least one player in the set S who sometimes would play a if node x were reached. Then it will be assumed[2] that y is labeled with the pair (S',T), where $S' \subseteq S$. A similar assumption is made if it is player II who moves at node x. The perennial problem of refinement theory arises when *no* player in S would ever take action a if node x were reached. How then should node y be labeled? This problem will not go away, but it will be put to one side for the moment.

If the first node of the n-legged Centipede Game of Figure 1 is labeled (S,T), then it will be denoted by $G_n(S,T)$. An elaborate definition of a 'rational player' is not necessary if the aim is only to show that the 'rational' opening move is *down* in $G_n(R_1,R_2)$, where R_1 and R_2 will always denote

[2]The assumption implies that actions taken by one player are not informative about the other. I make this assumption only to keep things simple.

sets of 'rational players'. A 'rational player' need only be taken to be someone who would maximize his payoff when making the first move in all games $G_n(R_1, R_2)$, provided that his knowledge were adequate to determine the maximizing action.

For all R_1 and R_2, it follows that *down* would always be played in $G_1(R_1, R_2)$. As an induction hypothesis, suppose that the first move of $G_{n-1}(R_1, R_2)$ would be *down* for all R_1 and R_2. Now consider the first move of $G_n(R_1, R_2)$. If the play of *across* were a possible opening move of $G_n(R_1, R_2)$, then the second node $G_n(R_1, R_2)$ would need to be labeled (R_1', R_2), where $R_1' \subseteq R_1$. But we know that *down* would be the opening move of $G_{n-1}(R_1', R_2)$ if this were reached. Moreover, the common knowledge assumption implies that the player making the opening move of $G_n(R_1, R_2)$ knows this also. It follows that a 'rational player' in the set R_1 would be making a suboptimal move by playing *across*, since his payoff from playing *down* would be greater. From this contradiction, we deduce that no 'rational player' in the set R_1 would ever open $G_n(R_1, R_2)$ by playing *across*. When there is prior common knowledge that everybody is a 'rational player', it follows that the Centipede Game will open with the play of "down".

Is it rational to be a 'rational player'? To address this point, consider the labeling of the second node of $G_3(R_1, R_2)$. Since this node cannot be reached, we have no rule to assist in its labeling. Nevertheless, to assess the rationality of a 'rational player' who plays *down* at the opening move, we need to ask what payoff he *would* get if he *were* to play *acrosss*.

I have italicized the subjunctives in the preceding sentence to emphasize that we have a subjunctive conditional to consider. Aumann's agrees with Binmore and Samuelson [7] that subjunctive conditionals are important in his arguments and later versions of his paper emphasize their appearance in his reasoning. In presenting my version of his argument, I have therefore tried to be pedantic in my use of the subjunctive mood, even though too much use of the subjunctive is oppressive to the modern ear. However, the new subjunctive conditional that we need to consider has a different character from those that appear in Aumann's argument because its interpretation requires entering a different possible world from those he implicitly considers.

Subjunctive conditionals A brief aside on the subjunctive conditionals will now be offered because there is a great danger of their being confused with the material implications that are always adequate when interpreting 'If P, then Q' sentences in pure mathematics. However, the following simple example may provide some food for thought. Consider the subjunctive conditional, "If my dean were a man, my salary would be tripled." If this sentence is treated as a material implication, then it is true because my dean is actually a woman. But, as used in ordinary conversation, it is clearly false.

Philosophers have written at length on the subject of how subjunctive conditionals are to be interpreted (Sandford [12]). But the orthodox view seems entirely adequate for the purposes of decision theory. I shall therefore follow the usual practice of interpreting a subjunctive conditional using the notion of a *possible world.*

When the antecedent P in the subjunctive conditional "If P were true, then Q would be true" is actually false, one looks to the context in which the sentence is uttered for relevant possible worlds in which P is true. If Q does indeed follow in such possible worlds, then the subjunctive conditional is said to be true. For example, since my dean is actually a woman, someone interpreting the subjunctive conditional "If my dean were a man, my salary would be tripled" needs to consider what possible world I have in mind when seeking to make sense of my statement. In these enlightened times, the relevant possible world is clear enough. It is created by replacing my current female dean by a male dean, leaving everything else the same. However, were Isaac Newton to have said "If my dean were a woman, my salary would be tripled", we would certainly not have thought it appropriate simply to replace his male dean by female dean, leaving everything else the same. For a female dean to be possible in the seventeenth century, all sorts of other changes in society would need to be postulated.

Specifying the context Within what possible world should we interpret a subjunctive conditional that begins, "If a 'rational player' were to play *across* ..."? *Inside* Aumann's argument, the relevant possible world is one in which the 'rational player' who played *across* is still regarded as 'rational' in spite of his choice of action. Aumann [1] insists on this interpretation, and I agree that it is the appropriate interpretation within his argument.

But, nothing says that subjunctive conditionals *outside* his argument that begin, "If a 'rational player' were to play *across* ..." must be interpreted in the manner that is appropriate *inside* his argument. How a subjunctive conditional is interpreted depends on the context in which it arises—just as in the example involving my dean. If the context is uncertain, then it is the duty of the analyst to clarify the context he has in mind by making formal assumptions if necessary.

One way of specifying a context for our troublesome subjunctive conditional is to name a label (S, R_2) for the second node of $G_3(R_1, R_2)$. One can, for example, follow Zermelo[3] and insist that $S \subseteq R_1$. If we do insist that $S \subseteq R_1$, then we can justify backward induction. However, the assumption that $S \subseteq R_1$ seems strange to the layman, who argues that the play of *across* has refuted the hypothesis that the opening player lies in the set R_1. This leads him to propose that $S \subseteq CR_1$. One possibility for the set S is then that its members would always play *across* no matter what. Backward induction would then fail since player II would play *across* if the second node of $G_3(R_1, R_2)$ were reached, because player II would then believe player I would play *across* if the third node were reached. In this situation it would definitely be irrational to be a 'rational player'.

But nothing compels us to adopt either $S \subseteq R_1$ or $S \subseteq CR_1$ as properties of the relevant possible world within which to interpret our troublesome counterfactual.[4] If we wish to justify the rationality of a 'rational player', we therefore need to add something to the assumption of prior common knowledge of 'rationality'—something that tells us what *would* be known if a 'rational player' *were* to play *across*. As stressed in Binmore [5], it follows that the formal definition of a game is not adequate for a full analysis of the game, even if one adds the proviso that there is prior common knowledge of rationality. Often this point is made by asserting that an equilibrium concept needs to incorporate a "theory of mistakes".

---

[3]Note that, *within* Zermelo's proof of the minimax theorem for Chess, this is indeed the appropriate interpretation. In computing security levels, Zermelo always needs to make whatever hypothesis is least favorable about the future course of play for the player currently being considered.

[4]The logic of the layman's argument is flawed because anything follows from a contradiction. But it does not therefore follow that his suggestion that $S \subseteq CR_1$ can be rejected.

The simplest such theory of mistakes is Selten's [14] trembling-hand story. This tells us that 'irrational play' by a 'rational player' is always to be attributed to transient random errors. We then have a rationale for assuming that $S \subseteq R_1$. But nothing says that this story is the only story that can be told. Indeed, adopting such a story would seem to close the door on any hopes that game-theoretic results might be relevant to the play of real people. We all know that bad play by actual people is usually the result of a failure to think things through properly—and people who have reasoned badly in the past are likely to reason badly in the future.

Kreps, Milgrom, Roberts and Wilson's [10] "gang of four" paper tells a different story in which there are 'irrational' types as well as 'rational' types of player. Within such a story, the observation of an action that would be a mistake for a 'rational' player is explained by attributing it to an 'irrational' player—just as our layman would wish. In Selten's terminology, trembles are then correlated and so backward induction cannot be justified. Of course, the analysis of a game with such a theory of mistakes is much harder than with Selten's trembling-hand story. But if we want a theory that is at all relevant to what real people do when they play games, it seems to me that this is the route we must follow.

Rationality as a property of players Sometimes a crude form of error is made when variants of the fallacy of the twins are used in an attempt to prove that cooperation in the one-shot Prisoners' Dilemma is rational. Player I supposedly reasons as follows:

> I am rational. So anything I decide to do will necessarily be rational. Player II is also rational and hence will aways make the same decision as I make when placed in identical circumstances. Therefore, he will do whatever I do in the Prisoners' Dilemma. Hence I should cooperate.

The mistake is to argue that a choice is rational because it is made by a rational person. But this is to put the cart before the horse. We do not say that a choice is rational because it has been chosen by a rational person. We say that a person is rational because he chooses rationally.

I offer this fallacy as a warning before considering a possible objection to my claim that it is not necessarily rational to be a 'rational player'. One can argue that my definition of a 'rational player' is unduly narrow—and

it is certainly true that there is more to rationality than my definition of a 'rational player' allows. However, there is a danger to be avoided if the definition is extended. We must not argue that, if the opening player in $G_3(R_1, R_2)$ plays *down*, then this action is necessarily optimal *because* he is known to be rational. This would force us to conclude that being rational implies knowing that, if *across* were to be played, then player II would necessarily not deduce that player I is someone who always plays *across*. But what would be the source of such knowledge? It seems to me that the causal chain should always flow from knowledge to action rather than the reverse.

3 Conclusion

In brief, I believe that the rationality of a 'rational player' must necessarily remain open as long as we have no idea what would be believed about him if he were to make an 'irrational' move. In consequence, we have no grounds for claiming that we know the 'solution' to games like the Centipede or the finitely repeated Prisoners' Dilemma. For myself, the realisation that one could be a game theorist without having to argue for systematic defection in the finitely repeated Prisoners' Dilemma came as a great relief.

References

[1] R. Aumann. Backwards induction and common knowledge of rationality. 1993. Paper presented at the Nobel Symposium on Game Theory, to appear in *Games and Economic Behavior*.

[2] R. Aumann. Irrationality in game theory (preliminary notes). 1988. Hebrew University of Jerusalem".

[3] C. Bicchieri. Common knowledge and backwards induction: a solution to the paradox. In M. Vardi, editor, *Theoretical Aspects of Reasoning About Knowledge*, Morgan Kaufmann, Los Altos, 1988.

[4] C. Bicchieri. Strategic behavior and counterfactuals. *Erkerntnis*, 30:69–85, 1989.

[5] K. Binmore. Modeling rational players, I and II. *Economics and Philosophy*, 3 and 4:179–214 and 9–55, 1987.

[6] K. Binmore, J. Gale, and L. Samuelson. Learning to be imperfect: the ultimatum game. 1993. University of Wisconsin Discussion Paper.

[7] K. Binmore and L. Samuelson. Rationalizing backward induction? 1993. (Forthcoming in the proceedings of the International Economics Association conference in Turin on "Rationality").

[8] W. Guth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Behavior and Organization*, 3:367–388, 1982.

[9] E. Kohlberg and J. Mertens. On the strategic stability of equilibria. *Econometrica*, 54:1003–1037, 1986.

[10] D. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

[11] P. Reny. Rationality, common knowledge, and the theory of games. 1985. Working Paper, University of Western Ontario.

[12] D. Sandford. *If P, Then Q*. Routledge, London, 1989.

[13] R. Selten. The chain-store paradox. *Theory and Decision*, 9:127–159, 1978.

[14] R. Selten. Reexamination of the perfectness concept for equilibrium points in extensive-games. *International Journal of Game Theory*, 4:25–55, 1975.