1 Introduction

In 1987, I wrote a paper (Binmore [14]) that questioned the rationality of the backward induction principle in finite games of perfect information. Since that time, a small literature has grown up in which Antonelli and Bicchieri [1], Ben-Porath [9], Bicchieri [10,11], Bonanno [21,22], Pettit and Sugden [28], Reny [30], Samet [32], Stalnaker [37,36] and numerous others have attempted with varying success to treat the issues formally.

I believe my claim that rational players would not necessarily use their backward-induction strategies if there were to be a deviation from the backward-induction path is now generally accepted. But Aumann’s [3] has recently offered a formal defense of the proposition that prior common knowledge of the players’ rationality implies that play will nevertheless necessarily follow the backward-induction path. He argues that the conclusion is counter-intuitive in certain games, but attributes our discomfort with the result to a failure to appreciate how strong his assumptions are. However, although Aumann’s deep and thought-provoking contributions to the foundations of game theory provide the chief inspiration for this note, my purpose is not to comment specifically on his recent article. Its purpose is to question the significance of this and other results of the formalist genre.

I have a text, it always is the same,  
And always has been, since I learnt the game.  

Chaucer, The Pardoner’s Tale
Without intending any disrespect to the authors, I believe that there is little of genuine significance to be learned from any of the literature that applies various formal methods to backward induction problems—even when the authors find their way to conclusions that I believe to be correct. It seems to me that all the analytical issues relating to backward induction lie entirely on the surface. Inventing fancy formalisms serves only to confuse matters. The related literature on the Surprise Test Paradox provides a particularly blatant example. The paradox has a trivial resolution (Quine [29], Binmore [16]), but the various exotic logics that have been brought to bear on the problem never come near exposing the piece of legerdemain by means of which we are deceived when the problem is posed.

Formalists will object, saying that an argument is open to serious evaluation only after it has been properly formalized. But this is a disingenuous response. It is true that, if we were in serious doubt about whether an author had succeeded in analyzing his or her model correctly, then it would be foolish not to insist that the argument be given in precise terms. However, the literature on backward induction seldom provokes doubts at this level. The issue is almost never whether a particular model has been analyzed correctly—but whether the correct model has been analyzed.

In brief, I think that the backward induction problem-like much else in the foundations of game theory-poses only a very small challenge to our powers of formal analysis. The real challenge is not to our powers of analysis, but to our ability to find tractable models that successfully incorporate everything that matters. In particular, it seems entirely elementary that, whatever model of a player is used, it must be rich enough to encompass irrational behavior as well as rational behavior (Binmore [14]). What keeps a rational player on the equilibrium path is his evaluation of what would happen if he were to deviate. But, if he were to deviate, he would behave irrationally. Other players would then be foolish if they were not to take this evidence of irrationality into account in planning their responses to the deviation. A formal model that neglects what would happen if a rational player were to deviate from rational play must therefore be missing something important, no matter how elaborately it is analyzed. However, Aumann [3, Section 5c], for example, is insistent that his conclusions say nothing whatever about what players would do if vertices of the game tree off the backward-induction path were to be reached. But, if nothing can be said about what would happen off the backward-induction path, then it seems obvious that nothing can be said about the rationality of remaining on the backward-induction path. How else do we assess the cleverness of taking an action than by considering what would have happened if one of the alternative actions had been taken? But this is precisely what Aumann's [3] definition of

\[ \text{It is always easy to predict which line of research will be fruitful after the event.} \]

\[ \text{Although avoids the subjunctive mood, I take this to be the meaning of the apparently oxymoronic sentence, "The results of this paper say nothing about the behavior of the players at vertices that are off the backward induction path and are actually reached."} \]
rationality fails to do. (See Justification 6 of Section 4.)

Figure 1: The Centipede Game

In Binmore [14], I used Rosenthal's [31] Centipede Game of Figure 1(a) as an example when criticizing the defense of the backward induction principle that was then current. Figure 1(b) shows the strategic form of the special case when \( n = 3 \) (the three-legged Centipede). In this note, I plan to use the same example to elaborate on the criticism just expressed of the tighter defense of the principle that is possible if one follows Aumann [3] in abandoning claims about what would happen off the backward-induction path.

It is easy to verify that the backward induction principle requires that each player always plan to play down in the Centipede. In particular, the unique subgame-perfect equilibrium \( S \) in the three-legged Centipede is \((\text{old, d})\). However, the three-legged Centipede has other Nash equilibria. Part of the reason for writing this note is to argue that such alternative Nash equilibria have been too readily dismissed in the past—a theme pursued at greater length in Binmore et al [17,18]. In particular, the three-legged Centipede has a mixed Nash equilib-
rium N in which player I uses his backward-induction strategy with probability one, but player II mixes between \(a\) and \(d\), using the former with probability \(1/3\). If player I knows that player II will play \textit{across} with this probability, it is false that rationality requires that he play down. In fact, he is indifferent between playing \textit{down} and \textit{across}. Although he plays down with probability one in equilibrium, it is nevertheless equally rational for him to play \textit{across}.

Among other things, this note argues that prior common knowledge of rationality should not lead us to reject the equilibrium N. On the contrary, it is argued that N, rather than S, is the equilibrium of interest for the issues that the Centipede was constructed to explore. It is tempting to wave this point aside by conceding that perhaps prior common knowledge of rationality in the Centipede may lead to the play of N and so does not, after all, necessitate that player I open the Centipede by playing \textit{down}. But who cares if player I only plays \textit{across} with probability zero? But there is more riding on this issue than immediately meets the eye, as I hope will be evident by the end of this note. In particular, I hope that it will become apparent that we need not follow Aumann [6,3] in perceiving a sharp discontinuity between what happens when there is perfect common knowledge of rationality and when this condition is relaxed slightly. In particular, there is no need for game theorists to seek to insulate themselves from the criticism of experimentalists by claiming that their theorems have no relevance to how real people behave.

Section 2 comments briefly on the importance of common knowledge assumptions in general. Section 3 explores one of the reasons for the popularity of the claim that prior common knowledge of rationality implies the backward induction principle. It describes my version of a folk argument that purports to demonstrate that prior common knowledge of rationality in the Centipede Game implies that its opening move will necessarily be \textit{down}. As with Aumann's [3] more complicated theorem, the argument is correct, in the sense that the conclusion does indeed follow from the premises. But something must be wrong at the conceptual level, because the conclusion that player I will begin by playing down is obtained without any reference to his beliefs about what would happen if he were to play \textit{across}. But if the probability that player I assigns to the event that player II would then also play \textit{across} is sufficiently high, it is obviously not...

A The three-legged Centipede has other mixed Nash equilibria. Figure 1(c) shows a prism, the points of which represent all pairs of mixed strategies for the reduced strategic form of the three-legged Centipede (obtained by deleting the row \(da\)). A point in the prism corresponds to a Nash equilibrium if and only if \(\mathcal{J}\) lies on the closed line segment \(NS\). This note discusses possible deductive analyses of the Centipede (Binmore[14]). In an evolutive analysis, it would be of interest to note the trajectories of the replicator dynamics indicated in Figure 1(c). No Nash equilibrium is an asymptotic attractor in these dynamics. Moreover, after an equilibrium \(E\) in the relative interior of \(NS\) has been perturbed by introducing a small fraction of player I's who use \(a\), the system returns to an equilibrium that is nearer \(N\) than \(E\). Drift induced by a tendency on the part of player I to use \(a\) rather than \(ad\) when making an out-of-equilibrium deviation will therefore result in a movement along \(NS\) towards \(N\). A tendency to use \(ad\) rather than \(au\) will result in a drift in the opposite direction towards \(S\)....
optimal for player I to begin by playing down. I believe that this apparent paradox arises partly as a consequence of a failure to appreciate how counterfactual reasoning works. Section 4 therefore seeks to demystify this question. Section 5 attempts to resolve the paradox by retelling the story with a less restrictive background model. However, once a paradox-free model has been adopted, the door is no longer closed on the Nash equilibrium N. Finally, Section 6 tries to say something about what the conclusions mean by taking up a clarion call from one of Aumann’s [7] previous papers, and asking what we are trying to accomplish when we prove theorems in game theory. Personally, I think it is because this question has been so neglected that the foundations of game theory are now in such a mess.

2 Common Knowledge

Before Aumann [2] put the concept of common knowledge on an operational footing, game theorists were very casual about what their players did or did not know. More recently, it became fashionable to insist that everything is to be assumed to be common knowledge, regardless of its relevance to the issue at hand. Now the wheel has turned again, with Aumann and Brandenburger [8] insisting that no common knowledge at all is necessary to justify Nash equilibrium. As they observe, it is obviously true that justifying Nash equilibrium requires no more than that each player knows the game and the strategies used by his opponents. In the hope of averting confusion, this section is devoted to pointing out that this truth, and its elaborations to more complex situations, are largely beside the point.

The purpose of prescriptive game theory is to advise players about the rational course of action in a game. If a prescriptive theory begins with the assumption that the players already know its prescriptions, then it preempts its own reason for existing. To be useful, a prescriptive theory needs to begin with assumptions that do not trivialize the problem. It is important to insist that such assumptions often do require that various things are common knowledge—notably, the rules of the game, the preferences of the players over the outcomes, and the players’ beliefs about chance moves in the game. One may summarize these assumptions as the requirement that the game being played is common knowledge among the players. Examples that demonstrate how slight relaxations of this requirement can sometimes have big effects are a commonplace of the theory. For example, all Nash equilibria in the IV-times repeated Prisoners’ Dilemma result in each player defecting at each stage with probability one. However, this conclusion evaporates if the value of N is not

\[ 5 \] Haranyi’s theory of incomplete information does not provide a counter-example. His theory provides a recipe for adding information to a situation in which information is incomplete until a game has been constructed whose structure is common knowledge (Binmore [3, Chapter 11]).
common knowledge (Binmore [13, Chapter 10]).

Of course, after a game-theoretic analysis has been brought to a successful conclusion and the results published in a book, it may well be reasonable to assume that each player knows that his opponents will play as the book prescribes. But it does not follow that one can then discard the knowledge assumptions that led its author to write what he wrote. The reason that rational players know what their opponents will do after reading the book is because they are able to check that the author's conclusions do indeed follow from his hypotheses. However, within the static formalism that Aumann [4,5,8] now favors, there is no way to compare the state of things before and after an analysis, or even to ask why a player takes one action rather than another. As Aumann [4] puts it, players “just do what they do”. An analyst can only look on from outside the world in which the game is played and comment that if the players happen to make a Bayesian-rational choice then they will be operating a correlated equilibrium. As Aumann [4] insists, such a model is neither prescriptive nor descriptive, but what he calls “analytic”. Personally, I think that such analytic models have their uses, but they will not suffice as a foundation stone for all of game theory. If it sometimes seems otherwise, it is because, as with George Orwell’s newspeak, criticism of the defects of analytic models sometimes cannot even be expressed in the language within which they are formalized.

Although I believe it would be a mistake to follow Aumann and Brandenburger [8] in focusing attention away from the underlying common knowledge assumptions of game theory, this is not because I think one will often be led to wrong conclusions by treating common knowledge considerations in an informal manner. It is at the interpretive level that the importance of common knowledge assumptions needs to be acknowledged. In Harsanyi’s theory of incomplete information, for example, one really does need to assume that the underlying distribution of types is common knowledge. But the theory is nevertheless widely applied to situations in which this assumption is highly implausible. In any case, although no formal definitions are introduced, it is important to emphasize that this paper takes both common knowledge of the game and prior common-knowledge of rationality for granted. If they feel the need, those familiar with Aumann [3] will have no difficulty in formalizing what it means for something to be common knowledge in the simple case covered by Proposition 1. My own view is that we run into conceptual difficulties when considering the implications of prior common knowledge of rationality in finite games of perfect information because we are unsure how rationality should be defined—not because we are unable to handle the technicalities of common knowledge.

Finally, while on the subject of rationality, I am anxious to clear the air by stipulating that the general difficulties I have raised elsewhere (Binmore [14]) about the coherence of the notion of “perfect rationality” evaporate in the context of a finite game of perfect information. In such games, the issue of whether rational players are so perfect that they can decide the undecidable does not arise.
3 ‘Proving’ Backward Induction

Knowledge of a finite game of perfect information can be summarized as a list of conditional sentences of the type “If both players were always to play across, then player I would get a payoff of $n$ and player II would get $n + 1$”. A player’s knowledge of the characteristics of his opponents (including what the opponents know or do not know about him) is not always 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 seek to label each node $x$ in the Centipede Game with a pair $(S, T)$ of finite 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, the assumption of prior 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. (When the word ‘rational’ appears in quotes, there will be reason later to ask whether the framework within which it is being used is capable of bearing the load.)

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 that $y$ may be labeled with a 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(a) is labeled $(S, T)$, then it will be denoted by $G_n(S, T)$. An elaborate definition of a ‘rational’ player is not needed if the aim is only to show that the ‘rational’ opening move is necessarily $\text{down}$ in $G_n(R_1, R_2)$, where $R_1$ and $R_2$ will always denote 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 $G_n(R_1, R_2)$.

**Proposition 1.** When there is prior common knowledge that everybody is a ‘rational’ player, the Centipede Game necessarily begins with the play of $\text{down}$. 

Proof. For all $R_1$ and $R_2$, it follows immediately from the definition of a ‘rational’ player that $\text{down}$ is always played in $G_1(R_1, R_2)$. As an induction hypothesis, suppose that the first move of $G_{n-1}(R_1, R_2)$ is $\text{down}$ for all $R$ and

---

6 The assumption implies that actions taken by one player are not informative about the other. I make this assumption only to keep things simple. Binmore [15] uses Selten’s Horse Game as an example when exploring the alternative hypothesis.
Now consider the first move of $G_n(R_1, R_2)$. If the play of across is a possible opening move of $G_n(R_1, R_2)$, then the second node $G_{n-1}(R_1, R_2)$ should be labeled $(R'_1, R_2)$, where $R'_1 \subseteq R_1$. But we know that down is the opening move of $G_{n-1}(R'_1, R_2)$ if this is 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 $R_1$ makes a suboptimal move by playing across, since he knows that his payoff from playing down would be greater. From this contradiction, we deduce that no 'rational player' in the set $R_1$ ever opens $G_n(R_1, R_2)$ by playing across.

The proposition shows that, with prior common knowledge of 'rationality', the opening move of the Centipede Game cannot be across. A formalist might be satisfied to stop at this point, but I think it important to continue by asking the seemingly stupid question:

Is it rational to be 'rational'?

To address this question, consider the labeling of the second node of $G_3(R_1, R_2)$. Since it has been shown that this node cannot be reached, we have no rule to assist in its labeling. We therefore lack the information we need to predict what would happen if the second node were reached. 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 across.

The subjunctives in the preceding sentence have been italicized to emphasize that we have a counterfactual conditional to evaluate. Aumann's [3] discussion of such counterfactuals, which actually play no role whatsoever in his formal analysis, obscures the important point that the interpretation of a counterfactual depends on the context in which it arises. An aside on this issue is therefore necessary.

4 Conditionals

Sanford’s [33] book, *If* $P$, *then* $Q$, reviews the many attempts that have been made to provide an adequate account of conditional reasoning. The debate began in ancient times and continues unabated into the present. There seems to be no consensus even on definitions. However, I shall adopt the terminology of Flew’s [24] widely quoted Dictionary of Philosophy. Flew distinguishes material conditionals, subjunctive conditionals and counterfactual conditionals.

A material conditional is usually called a material implication by mathematicians and written as $P \Rightarrow Q$. Such conditionals are all that is needed in pure mathematics and so mathematicians are often reluctant to concede that...
other types of conditional may sometimes be useful. However, the *subjunctive conditional*, “If my dean were a man, then my salary would be astronomical” is false as used in ordinary conversation—even though my dean is actually a woman and so the conditional has a false antecedent (which would make it true if interpreted in the material sense). A subjunctive conditional with such a false antecedent is said to be a *counterfactual conditional*.

How are subjunctive conditionals to be interpreted? The simplest approach uses the notion of a *possible world*. For example, since my dean is actually a woman, someone interpreting the counterfactual conditional, “If my dean were a man, then my salary would be astronomical” needs to consider what possible world I have in mind when seeking to make sense of the sentence. 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 astronomical”, we would certainly not have thought it appropriate simply to replace his male dean by a 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. How a subjunctive conditional is to be interpreted therefore depends on the context in which it arises. If the the context is uncertain, it is the duty of the analyst to clarify the context that he has in mind by making formal assumptions if necessary.

In Section 3 of this paper, all three types of conditional appear. Subjunctive conditionals appear in the description of the game. When we are told, “If both players were always to play across, then player I would get a payoff of n”, we are accepting a constraint on the possible worlds we are allowed to postulate during an analysis of the game. The definition of a ‘rational’ player makes a similar use of subjunctive conditionals. We are told something about what such a player would do if he were to play the game $G_n(R_1, R_2)$. (I find it helpful to think of a ‘rational’ player as a computer program that would produce certain outputs if it were offered certain inputs. However, only some of the many potential inputs it might receive will actually be realized.)

Section 3 continues with the proof of Proposition 1. Wherever the indicative mood has been used in this proof, the conditionals are intended as material conditionals—as in a regular mathematical proof. In such arguments by contradiction, it is of some importance to be clear on this point. As Sanford [33] documents, it is easy to go astray when attempting to argue by contradiction when the conditionals are subjunctive, because nothing says that all subjunctive conditionals that appear in an argument must be interpreted within the same possible world.

Section 3 ends by questioning the rationality of a ‘rational’ player who opens $G_n(R_1, R_2)$ with the play of down, by asking what such a player would get if he were to play across. Here we are definitely faced with a counterfactual conditional. As stressed in Binmore [14], the standard definition of a game tells us nothing whatever about the nature of the possible world or worlds within
which such a counterfactual conditional is to be interpreted. It is true that, when arguing by contradiction in Proposition 1, we maintained the hypothesis that common knowledge of ‘rationality’ was still in place when contemplating the possibility that player I might start by playing across. But after Proposition 1 has been proved, we have to live with the fact that common knowledge of ‘rationality’ would be refuted if player I’s opening move were across. The meaning of the counterfactual conditional, “If a ‘rational’ player were to open \( G_n(R_1, R_2) \) by playing across, he would get a smaller payoff than if he played down,” therefore remains open for debate unless further information is supplied to establish a context for its interpretation.

One way of specifying a context for our troublesome counterfactual conditional is to name a label \((S, R_2)\) for the second node of \( G_3(R_1, R_2) \). To make the traditional backward-induction argument work, we need that \( S \subseteq R_1 \). 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 player I lies in the set \( R_1 \). This leads him to propose that \( S \not\subseteq S \).

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. If a precondition of being rational is to be ‘rational’, we are therefore led to the odd conclusion that a rational player cannot know that if an opponent ever were to play across in the Centipede, then he would always play across. But, when we assumed prior common knowledge of the players’ rationality, did we really intend to restrict what the players might or might not believe in the counterfactual event that someone were to play irrationally?

But nothing compels us to adopt either \( S \subseteq R_1 \) or \( S \not\subseteq R_1 \) as properties of the relevant possible world within which to interpret our troublesome counterfactual. 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. I can think of a number of ways in which someone defending backward induction might seek to justify or evade the requirement that \( S \subseteq R_1 \).

Justification 1: One could simply add the assumption \( S \subseteq R_1 \) to the other assumptions being made. For example, in the spirit of Selten’s [34] agent-normal form of a game, it is not unknown for authors to propose that a player be modeled as a collection of independently acting agents, one for each decision node at which the player might have to make a decision. As Binmore and Samuelson [19] observe, it then seems relatively innocent to propose that, if
prior common knowledge of the rationality of one of the players were to be refuted during the course of the game by some display of irrational behavior on the part of one or more of his agents, then rationality should still be attributed to those of his agents who have yet to play. One might even be forgiven for regarding the assumption as being so natural that it need not be explicitly stated when assigning a meaning to the counterfactual conditionals that arise when irrational play needs to be contemplated. However, since a rational player is clearly not simply a collection of independently acting agents, there seems little point in deducing the backward induction principle from the assumption that he is.

Justification 2: One could follow Selten [34] in his defense of perfect equilibrium by attributing any ‘irrationalities’ that may arise during the game to transient random errors that have no significance for a player’s future play.

Justification 3: One might follow Zermelo [38] in his study of Chess by proceeding on the assumption that each player needs always to take the least favorable view of his opponent when assessing possible futures. This is certainly the correct procedure when computing a player’s security level in a game—and security strategies are what we care about in two-person, zero-sum games like Chess. Moreover, in such games, the least favorable assumption is always that the opponent will behave rationally in the future no matter how irrationally he may have behaved in the past. But the Centipede Game is not zero-sum. Zermelo is therefore irrelevant as an authority in this context.

Justification 4: One might adopt a strict revealed-preference line as in Binmore [16]. If we take the choices that would be made at nodes as fundamental, then the reason, for example, that down is assigned a larger payoff at the final node of the Centipede than across is because it is part of the data of the problem that if player I were to reach the final node and make whatever deductions about his situation that he then thought fit, he would definitely choose down and not across.

But the literature on backward induction does not adopt a strict revealed-preference approach because it then becomes a tautology that backward induction holds. Whatever behavior we might observe in a finite game of perfect information is compatible with backward induction if we are allowed to fill in the payoffs after the event. Binmore [16] makes this point to refute philosophers who argue that prior common knowledge of rationality in games like the Centipede makes backward induction necessarily irrational-as opposed to Aumann’s [3] claim to the contrary.

Justification 5: Alternatively, one might argue as follows. Because player I is known to be rational, if he plays down at the first node of the Centipede Game, then he must know that he would get a worse payoff from playing across. That is to say, his rationality includes his knowing that if he were to play across, then player II would not conclude that he is the type of person who always plays across. But what would be the source of this knowledge? Surely the
causal chain should run from knowledge to action rather than from action to knowledge. However, if rationality implies 'rationality', we have already seen that the assumption of prior common knowledge of rationality commits us to precisely this difficulty.

Justification 6: Aumann [3] seeks to evade the requirement that $S \subseteq R_1$ by denying that the rationality of choosing down need involve any knowledge at all of what would happen if across were played. Instead of saying that it is rational for a player to choose down if he knows or believes that choosing across would not result in a higher payoff, the strategem is to say that it is rational for a player to choose down if he does not know that choosing across would result in a higher payoff. But does this mean that we are to assume that he is so ignorant that the standard Bayesian assumption that he can quantify his ignorance with a probability distribution is to be denied? If not, then he must assign a probability $p$ to the event that the result of his playing across would be a payoff of 4 rather than the payoff of 2 he gets by playing down. If only down is Bayesian-rational, then $p < \frac{1}{2}$. But how come $p < \frac{1}{2}$? We seem to be stuck with the implication that this fact is somehow built into the assumption that there is prior common knowledge of rationality. That is to say, we have essentially the same problem that we faced in Justification 5.

Of the preceding attempts to justify or evade the assumption that $S \subseteq \text{ill}$, it is the second that I feel has most to be said for it. It is an up-front attempt to explain how a 'rational' player might come to act 'irrationally'. Often, the need for such an explanation is expressed by asserting that each equilibrium concept needs to incorporate a "theory of mistakes". However, nothing says that Selten's [34] trembling-hand story is the only story of mistakes 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 [26] "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. Indeed, within such a framework,
Fudenberg, Kreps and Levine [25] have shown that no refinements of Nash
equilibrium can be justified at all. Of course, the analysis of a game with a
realistic 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.

5 Paradox of Rationality?

The previous section argues that it may not be rational to be ‘rational’. But, whatever full rationality may be, surely it includes being ‘rational’?

In my view, the appearance of a paradox arises only because we have been
working with an inadequate background model. To explore this viewpoint, let us
first consider the case when it is not true that there is prior common knowledge
that all the players are rational. Instead, there is prior common knowledge that
a player is rational with probability 1 - \( \epsilon > 0 \) and irrational with probability
\( \epsilon > 0 \). Only the extreme case in which irrational players always play \text{across} will
be considered. This set-up was analyzed in Binmore [14] in much the same way
that the gang of four [26] analyzed the finitely repeated Prisoners’ Dilemma.
The analysis shows that equilibrium play in the Centipede requires the rational
players to mix between \text{across} and \text{down} at each stage of the game. If the
Centipede has sufficiently many legs, they play \text{across} with probability one in
the early stages of the game. At the final stage, they necessarily play \text{across}
with probability zero. At intermediate stages, they play \text{across} with a
probability that declines over time from one to zero.

For an equilibrium in the three-legged Centipede, the initial phase in which
rational players choose \text{across} with probability one is absent. Equilibrium play
requires that rational players use both \text{across} and \text{down} with positive proba-
bilities at nodes 1 and 2. The probabilities with which a rational player mixes
at one of these nodes are chosen to make a rational player at the other node
indifferent between playing \text{down} or \text{across}. As \( \epsilon \to 0 \), the probability that a
rational opening player chooses \text{across} declines to zero. However, he remains
indifferent between choosing \text{across} or \text{down} all the way up to the limit, where
the players’ behavior is summarized by the Nash equilibrium \( \mathcal{N} \) introduced in
Section 1.

To capture this phenomenon while actually working at the limit, it is nec-
ecessary to abandon the assumption of Section 3 that the sets of players to be
considered are finite, in favor of a model in which a set may be of measure zero
without being empty. Knowledge must also be reinterpreted as being “belief
with probability one”. The existence of irrational players in $G_3(R_1, R_2)$ is then not ruled out altogether. They may exist with probability zero. One can then contemplate a mixed equilibrium in which a null set of rational players open $G_3(R_1, R_2)$ by playing across. (The backward induction argument of Section 3 shows only that the set of ‘rational’ players who play across cannot have positive measure.) It is then not true that a rational player must play down at the opening move of $G_3(R_1, R_2)$. He will be indifferent between his two choices. This is particularly important when the general question of backward induction in the Centipede is at issue. Although only a null set of rational players would begin $G_n(R_1, R_2)$ by playing across, the existence of this set nevertheless has an enormous impact on what would happen if later nodes were to be reached.

The general procedure followed in this section seems usable whenever a puzzling counterfactual arises. Instead of ignoring the difficulty or inventing exotic methods of analysis to deal with the problem of what would happen if impossible events were to occur, one enlarges the model so that the impossible events cease to be impossible (Selten and Leopold [35], Binmore [14]). One then returns to the idealized world in which the problem first arose by allowing appropriate parameters to approach their extremal values. In this section, for example, we studied a world parametrized by $c > 0$, and then considered the limit as $c + 0$. The idealized world obtained in this way then retains the essential properties of the less idealized worlds in which the analysis was conducted—albeit sometimes in vestigial form, as with the null set of rational players who may begin the Centipede by playing across according to the analysis of this section.

In my opinion, we neglect such vestiges of realism at our peril, especially if we hope that the theories we propound will have some relevance to applied work. We know from the work of McKelvey and Palfrey [27] that real people are not inclined to open the Centipede by playing down. Such behavior is simply irreconcilable with Aumann’s [3] idealization of a player. But with the approach outlined in this section, one is offered a clue about which idealizing assumptions need to be relaxed to accommodate the data.

6 What Are We Trying to Accomplish?

Early game theorists seem to have taken for granted that their role was prescriptive—to advise players on how to optimize given their information, on the often tacit assumption that it is common knowledge that the other players are receiving similar advice. On the standard assumption of neoclassical economics that all agents behave as though in receipt of such advice, one would then have a descriptive perspective.
tive model—one that allows predictions to be made about the world. Confusion at the conceptual level is therefore possible because the same theorem may be useful in both prescriptive and descriptive game theory. Such confusion may be compounded when theorems are proved whose interesting interpretations lie in what Aumann [4] calls analytic game theory.

I have commented on Aumann’s [4] careful distinction between an analytic model and models constructed for other purposes elsewhere (Binmore [12]). The issues are simpler in the case of the Centipede Game. In brief, Proposition 1 makes perfectly good sense when interpreted in terms of an analytic model. In Aumann’s [4] expressive catchphrase, players in an analytic model ‘just do what they do’. We do not ask how it comes about that they behave like they do, we simply write down some conditions that are assumed to constrain their behavior and explore their implications. In the case of the Centipede Game, common knowledge of ‘rationality’ turns out to imply that player I begins by playing down. But it also turns out to imply that player I acts as though he knows or believes that if he were to play across, then the probability that player II would then play across is less than \( \frac{1}{2} \).

But suppose that we try to use Proposition 1 for prescriptive purposes. Imagine that it is common knowledge that players I and II have hired Von Neumann and Morgenstern respectively to give them advice. Von Neumann applies Proposition 1 and advises player I to choose down at the first node. Morgenstern tells player II that she will not require his services since the second node will not be reached. Player I now asks Von Neumann why he should choose down at the first node, and the reply is that this conclusion follows from Proposition 1. But player I very reasonably finds this answer inadequate and persists by asking what advice Von Neumann predicts that player II will receive from Morgenstern. On receiving the reply that Morgenstern will offer her no advice because he will believe that the second node will not be reached, player I then tells Von Neumann that he is sure that player II would play across if node 2 were reached, and she had to act without the benefit of Morgenstern’s advice. Von Neumann objects that this is not part of the data of the problem as proposed to him, and player I agrees that Von Neumann was only asked to offer advice on the assumption that player II would act as advised by Morgenstern. However, player I explains that he has private information about player II’s past history of play when she acts without advice, that he thought irrelevant when he learned that player II had retained Morgenstern’s services.

Can Von Neumann now plausibly reply that it follows from Proposition 1 that player I cannot have this information? If he dared, player I would simply respond that Von Neumann should adjust his model to the data rather than trying to adjust the data to his model.

As for possible alternative models, Von Neumann can reason as in Section 5 and so reconcile player I’s claim to have private information about player II with a version of Proposition 1 in which the players are drawn from an infinite population. He will then be led to the mixed Nash equilibrium \( \hat{N} \) of the
Centipede Game introduced in Section 2. When player I now asks what advice Von Neumann predicts that Morgenstern will give player II, Von Neumann answers that Morgenstern will randomize, and sometimes advise her to play across and sometimes down, with the probabilities chosen so that it doesn’t matter to player I whether he takes Von Neumann advice or not. If player I now asks Von Neumann why he should therefore take Von Neumann advice, does Von Neumann now reply that he must do so because Morgenstern predicted that it was almost certain that he would? Player I would just respond that he doesn’t care whether Morgenstern’s prediction is verified or not.

Notice that the problem has now ceased to have backward induction per se as its focus. It now pivots around the old chestnut of why a player should mix between strategies that all yield the same payoff. But the standard defenses of mixed Nash equilibria all require introducing trembles of some sort and then taking a limit. However, if Von Neumann appeals to one of these defenses, he must abandon the postulates of Proposition 1 and argue instead that the world is such that players do not always follow the advice of their tame game theorists with probability one. Indeed, we already had to contemplate this possibility when finding a way for Von Neumann to accommodate player I’s private information about player II in the first place.

However hard we struggle, it is therefore necessary to face the fact that the story we are telling is only fully coherent in a world to which the postulates of Proposition 1 are only an idealizing approximation. The source of the paradox in the prescriptive case is that the story cannot be told while actually at the limit without fudging one issue or another.

What of the possibility of applying Proposition 1 to a neoclassical descriptive model? As a referee comments, in practice N and the subgame-perfect equilibrium S predict the same thing in such a model: namely that player I will play down with probability one. If the trembles are sufficiently small, what difference does it then make whether we predict N or S? But, as the difficulties we have encountered in seeking to interpret Proposition 1 should have warned us, it turns out that it may matter very much if one is interpreting an equilibrium as the end-product of some equilibrating process. One then needs to worry about the stability of equilibria. The case of the four-legged Centipede is particularly striking in this regard, since Cressman and Schlag [23] have recently shown that, although the subgame-perfect equilibrium S is locally stable with respect to the standard replicator dynamics, the same is not true of the equilibrium N at the other end of the component of Nash equilibria.13 In fact, trajectories lead away from N far into the interior of the phase space, where they wander all over the place before approaching the component of Nash equilibria again. The equilibrium N is therefore unsafe as a prediction of the end-product of an

---

13 Their work was partly inspired by the very informative simulations of perturbed and unperturbed replicator dynamics carried out by Giovanni Ponti at University College London. Actually, his simulations show that it is enough to study a slightly perturbed version of the replicator dynamics in the three-legged Centipede to make a sufficiently similar point.
equilibrating process like the replicator dynamics.\textsuperscript{14}

I know that these remarks on interpretation just seem like waffle to formalists. However, I hope that some readers at least will agree that it is a mistake to invent definitions of rationality that make it look as though theorems that are strictly applicable only in analytic models can be applied without careful appraisal to prescriptive or descriptive purposes.

References


\textsuperscript{14} For more general comments on when it is unsafe to use an equilibrium of an idealized model as a prediction of behavior in an unidealized model which it approximates, see Binmore and Samuelson [18].


