The Power of Paradox:

Some Recent Developments in Interactive Epistemology

Adam Brandenburger*
Stern School of Business
New York University
44 West Fourth Street
New York, NY 10012
abranden@stern.nyu.edu
www.stern.nyu.edu/∼abranden

First Version 06/09/01
Current Version 12/04/02

Bohr: It was a fascinating paradox.

Heisenberg: You actually loved the paradoxes, that's your problem. You revelled in the contradictions.

(*Copenhagen*, by Michael Frayn, Methuen, 1998)

*This survey owes much to joint work and many conversations with Robert Aumann, Amanda Friedenberg, Jerry Keisler, and Harborne Stuart. Scott Ashworth, Carliss Baldwin, Ken Corts, Joe Halpern, Rena Henderson, Martin Meier, and participants in TARK 2001 (University of Siena, July 2001), the SITE Workshop on Reasoning in Games (Stanford University, August 2001), and the Finance and Business Economics Seminar at the College of Business, University of Notre Dame, provided important input. Financial support from Harvard Business School and the Stern School of Business is gratefully acknowledged. paradox-12-04-02

Abstract

This survey describes a central paradox of game theory, viz. the Paradox of Backward Induction (BI). The paradox is that the BI outcome is often said to follow from basic game-theoretic principles–specifically, from the assumption that the players are rational. Yet, for many games, the BI prediction is both intuitively unsatisfying and experimentally invalid. We describe recent work on resolving this paradox. We suggest that the BI Paradox has proved to be very fruitful, in that its resolution has furthered the development of several new conceptual frameworks in game theory. These new frameworks are: (i) Belief Systems, which expand the traditional description of a game to include the players' beliefs, beliefs about beliefs ... about the game; (ii) Conditional and Lexicographic Probability Systems, which extend the usual Kolmogorov theory of probability to take account of probability-zero events; (iii) Complete Belief Systems, which are systems that contain every possible belief of each player; and (iv) Formal Languages, which are models of how the players reason about a game. The survey explains the role of each of these concepts in resolving the BI Paradox. We end with another paradox, akin to Russell's Paradox from set theory, that leads to an impossibility result on complete belief systems. This result points to an open area in game theory.

1 Introduction

Discoveries of paradoxes have often played a very useful role in the development of science. Among the best-known examples are Zeno's Paradox of Achilles and the tortoise, which stimulated understanding of the infinite; Russell's Paradox in set theory, which spurred the development of modern mathematical logic; the paradoxical findings of the Michelson-Morley experiment on the speed of light, which led to relativity theory; and the paradox of the wave-particle duality of light, from which came quantum mechanics.

The role played by paradox has been described as follows: "Whenever, in any discipline, we discover a problem that cannot be solved within the conceptual framework that supposedly should apply, we experience an intellectual shock. The shock may compel us to discard the old framework and adopt a new one. It is to this process of intellectual molting that we owe the birth of many of the major ideas in mathematics and science" (Rapaport 1967).[1]

Naturally, not everything that seems paradoxical turns out to be a true paradox in this sense. Turning to game theory, we can see the distinction clearly. Many findings of the subject feel paradoxical, at least when first encountered. Perhaps most famous is the Prisoner's Dilemma, with its conflict between the individually rational and mutually optimal courses of action. There is the idea of 'strategic inflexibility,' in which a player may be able to increase his payoff by deliberately discarding some of his available strategies. And there are many other examples, of course.

However, after a period of assimilation, findings such as these have become essential and 'positive' parts of game theory. If they appear counter-intuitive, then that means that game theory is working. It is giving us non-obvious insights into how strategic interactions operate, which is a sign of success of the field. Moreover, to understand the Prisoner's Dilemma or strategic inflexibility, to take our two examples above, only standard game-theoretic tools are needed; nothing new is required.

But there are other paradoxical-seeming ideas in game theory that remain

[1] The above list of paradoxes is also from Rapaport, op. cit.. See Barrow (1998) for a thorough discussion of these matters.

disturbing and troublesome, even after time has passed,[2] and that turn out to prompt the development of new frameworks. We will look at one of these–the paradox of Backward Induction–and we survey recent work aimed at resolving it. We will see that the Backward-Induction paradox is indeed a true one, in that its resolution requires us to rethink some of the standard tools of game theory and to develop some new ones.

This paper is not, of course, a substitute for the technical papers of the literature. Nor, we should emphasize, is it a comprehensive survey.[3] It tells the 'story' of Backward Induction in a way that definitely reflects the author's biases. (However, we do refer in footnotes to various parts of the literature that we don't cover properly here.)

2 Paradox Found

Rosenthal (1981) introduced a game, now commonly referred to as the Centipede Game, that has become very important in both theoretical and experimental work. A version of the game is depicted in Figure 1 below.[4]

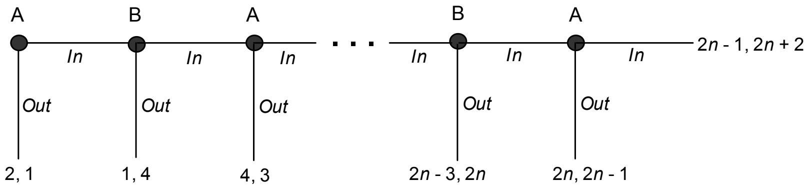


Figure 1

[2] For the record, we should note that Rapaport (1967), whom we quoted above, puts the Prisoner's Dilemma in this category. We do not. Aumann (1989, pp.21-22) emphasizes that the Prisoner's Dilemma involves no real paradox, but simply, and importantly, exhibits a very fascinating tension.

[3] Dekel and Gul (1997) is a very useful broad survey of foundational matters in game theory.

[4] The first number at each terminal node is the payoff at that node to Ann, and the second number the payoff to Bob.

A story that goes with the game[5] is that two players, Ann and Bob, are sitting at a table. In front of Ann are two stacks of coins, one totalling \$2 and the other \$1. Ann can either take the larger stack, leaving the smaller one for Bob, or pass both stacks to Bob. In the first case, the game is over; in the second case, an umpire adds \$2 to the larger stack. Bob then faces a similar decision. He can end the game by taking the larger stack (totalling \$4), so that Ann then gets the smaller stack (totalling \$1). Or he can pass both stacks to Ann. If he opts for the latter, the umpire now adds \$2 to the smaller stack, and it is again Ann's turn to decide. She can end the game, or she can give the stacks back to Bob, in which case the umpire adds \$2 to the larger stack. The game continues in this way, with the umpire alternating between adding \$2 to the larger stack and adding \$2 to the smaller stack. There is a pre-specified number n such that the game ends either: (i) when one of the players takes the larger stack; or (ii) when the stacks have ended up in front of Ann for the nth time, and she passes them back to Bob, at which point the umpire adds another \$2 to the larger stack, and Bob must take this stack.[6]

Notice that the payoffs to the players are designed so that the total pie grows each time one player gives the move to the other–i.e. makes the choice labelled *In* in Figure 1. But, at any point, a player would be better off ending the game–i.e. making the choice labelled *Out*–rather than continuing, if the other player should respond by ending the game. Intuitively, then, the game involves an interesting tension between obtaining a safe payoff by ending the game and trying to get more–which also involves a risk of ending up with less–by continuing.

Application of the Backward-Induction (BI) algorithm to the game resolves this tension in a stark way. Starting at the last decision node, which belongs to Ann, we select Ann's payoff-maximizing choice there, viz. *Out*. (She takes the larger stack of money.) Turning to the second-to-last node, which belongs to Bob, we select Bob's payoff-maximizing choice, taking Ann's subsequent choice as just determined–i.e. we select *Out* for Bob. (He, too, takes the larger stack.) Proceeding in this fashion, the BI algorithm concludes that, at her first node, Ann will choose *Out*, thereby ending the game

[5] Similar to one given by Sigmund and Nowak (2000).

[6] Thus, different values of n determine different Centipede games, with different numbers of 'legs.' For example, Figure 2 in the text is the three-legged Centipede.

immediately. Apparently, continuing the game cannot be to a player's benefit; the prospect of a potential gain from doing so turns out to be illusory.

This prediction–that Ann will choose *Out* immediately–is falsified when Centipede is actually played in laboratory experiments (see e.g. McKelvey and Palfrey 1992). Intuitively, too, one would expect Ann and Bob both to choose *In*, at least for the first few rounds, until one of them 'loses nerve' and decides to end the game.

Where is the paradox? The paradox is that, while the BI analysis is both intuitively unsatisfying and experimentally invalid in a game such as Centipede, use of the algorithm has been thought to follow inescapably from very basic game-theoretic principles. Indeed, one often sees statements to the effect that the BI path in Centipede must result if each of the players is *rational.* If this is really right, then to explain observed behavior in Centipede, we would have to assume that at least one of the players is acting irrationally. This is a possible assumption,[7] but is it a good one? Must Ann really be irrational to choose *In* at her first node? Wouldn't the choice of *In* be optimal, in fact, for Ann if she believed that Bob would then himself choose *In*?

Further thought along these lines suggests that the appropriate hypothesis might be not that the players are rational, but rather that they are rational and there is *common belief of rationality.* The latter means that each player believes that the other player is rational, each player believes that the other player believes this, etc. Henceforth, we shall use the abbreviation CBR to refer to the joint hypotheses of rationality and common belief thereof. The argument now goes like this: Assume CBR. Then, since Ann is rational, she will certainly choose *Out* at her last node. Turning to Bob at his last node, since he is rational and believes that Ann is rational, he, too, will choose *Out.* Continuing in this way leads to the conclusion that Ann will choose *Out* right away, just as the BI algorithm implies.

Assuming that this argument is really sound, then a resolution of the paradox of Backward Induction is at hand. After all, while the hypothesis that the players are rational is reasonable,[8] the hypothesis of CBR is far

[7] We shall see below how recent game models are able to formalize both rationality and irrationality in games.

[8] Though certainly not an inevitable one, as just noted.

more stringent.[9] There would seem to be nothing unrealistic about a situation where, say, Ann is rational, but isn't entirely sure that Bob is acting rationally. Or, one might imagine that Ann is rational, believes that Bob, too, is rational, but isn't entirely sure that Bob believes that she, Ann, is rational (though, in fact, she is). The point is that departures such as these from the hypothesis of CBR presumably allow for departures from the BI path of play in Centipede–i.e. situations in which Ann chooses *In*, perhaps Bob does too, and the game continues in this way, at least for a while. Theoretical and experimental investigation of Centipede would be brought back in line with each other.

There is just one more thing. We need to show that what we have said holds up in a precise mathematical treatment. We have to produce a formal set-up in which, say, both players are rational and believe the other to be rational, and yet they do not play the BI path. More generally, we want to be able to understand, in a precise formalism, the effect on the play of the game of various assumptions we might make about the players' rationality and beliefs. In doing this mathematics, we will discover that the theory beneath Backward Induction is a lot deeper than it looks from what we have said so far. We will find another paradox along the way and end up at a current frontier of game theory. The "one more thing" turns out to be a big thing.[10]

3 Belief Systems

To begin a formal analysis, we first have to say, a bit more precisely, what we mean by the rationality of a player. A *rational* player will choose a strategy that maximizes his expected payoff, where this expectation is calculated using the player's own (i.e. subjective) probability distribution on the possible strategies chosen by the other players. So, to talk about the rationality of the players, we need to talk about their beliefs (i.e. probability distributions) about strategies. But we also want to talk about the players' beliefs about

[9] Aumann (1995) calls an assumption such as this one "an ideal condition that is rarely met in practice.... This is not a value judgment; 'ideal' is meant as in 'ideal gas'."

[10] Since Rosenthal (1981), several other papers have pointed out various difficulties raised by the use of BI; see, inter alia, Basu (1990), Bicchieri (1989, 1992), Binmore (1987), Bonanno (1991), and Reny (1992). Ben Porath (1997) is an early and influential formal investigation of BI. The more recent papers described below build on these earlier ones.

one another's rationality. A little thought indicates that to do this, we need to talk about what the players believe about one another's beliefs. And so on.

Evidently, for our purposes the traditional game tree is only a partial description of the situation. It tells us the rules of play, and also gives the payoffs that the players assign to the different possible plays of the game. But there is a lot that the tree does not tell us. To go back to Centipede, the tree does not tell us what Ann believes about the strategy Bob actually decides to adopt, what Bob believes about Ann's strategy, what Ann believes Bob believes about her own strategy (which may not, of course, be what Bob actually believes), etc.

An *epistemic* analysis of Centipede starts with a richer description of the game, that includes the players' beliefs about the game. Figure 3 below is an example of such a richer description, for 'three-legged' centipede as depicted in Figure 2:

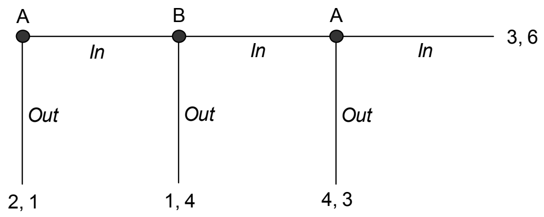


Figure 2

Let us explain the various ingredients of Figure 3, an object which is usually called a *belief system*:

(i) There is one possible *type* of Ann, labelled t^a.[11]

[11] The types encode the possible hierarchies of beliefs (about the game, about the other player's beliefs, etc.) that each player can hold. We will see right away how this works. The device of types goes back to Harsanyi (1967-68), who introduced it to treat uncertainty the players might have about the structure of the game (such as the payoffs). But the same device can equally be used to treat uncertainty about the play of the game (as we do here) or even to treat joint uncertainty about both the structure and the play of the game.

t^a	*Out*	*In*
t^b	0	0
u^b	0	1

t^b	*Out*	*In-Out*	*In-In*
t^a	0	1	0

u^b	*Out*	*In-Out*	*In-In*
t^a	0	1/2	1/2

Figure 3

(ii) There are two possible *types* of Bob, labelled t^b and u^b.

(iii) Associated with Ann's type t^a is a belief about what Ann is uncertain about, viz. what strategy (*Out* or *In*) Bob chooses and what type (t^b or u^b) Bob is. This belief–i.e. probability distribution–is depicted in the first matrix. Thus, we see that t^a assigns probability one to Bob's choosing *In* and being type u^b, and probability 0 to the other three strategy-type possibilities.

(iv) In similar fashion, associated with each of Bob's types is a belief about what strategy (*Out*, *In-Out*, or *In-In*) Ann chooses and what type Ann is. (In fact, since there is only one possible type of Ann, namely t^a, both types of Bob must assign probability one to this type.) The second matrix gives type t^b's belief, and we see that this type of Bob assigns probability one to Ann's choosing *In-Out*. The third matrix shows that type u^b assigns probability $\frac{1}{2}$ to Ann's choosing *In-Out* and probability $\frac{1}{2}$ to Ann's choosing *In-In*.

A *state of the world* specifies each player's strategy and type. Suppose, in fact, that the state is (*In-Out*, t^a, *Out*, t^b). Here, Ann chooses *In* and, if she gets a second move, then chooses *Out*. She assigns probability one to Bob's playing *In* and being type u^b. Since u^b assigns probability $\frac{1}{2}$ to Ann's playing *In-Out* and probability $\frac{1}{2}$ to Ann's playing *In-In*, we can also say that Ann assigns probability one to Bob's assigning probability $\frac{1}{2}$ to Ann's playing *In-Out* and probability $\frac{1}{2}$ to Ann's playing *In-In*. Turning to Bob, he chooses *Out*, and assigns probability one to Ann's playing *In-Out* and being type t^a (Ann's only type). Thus, we also have that Bob assigns probability one to Ann's assigning probability one to his (Bob's) playing *In*. Continuing in this same way, one can read off from Figure 3 all of the players' higher-order beliefs about beliefs about ... strategies. This richer description of the game, beyond the tree itself, is exactly what we were after.

Two observations on belief systems: First, note that in the belief system just given, Ann happens to be wrong about Bob's strategy and type. She assigns probability one to the strategy-type pair (*In*, u^b), not the actual pair (*Out*, t^b). This kind of situation is fully allowed for in the epistemic approach to game theory. There is no presumption that players have correct beliefs about one another. Mistaken beliefs, misperceptions etc. can be part of the picture. Second, a belief system is just a tool for describing what might be happening in a game.[12] A belief system (together with a state of the system) is not a prediction about how the game must be played or about what beliefs the players must hold. It is just a description of what happens to be happening. Thus, in our present example, there is nothing inevitable about the situation described by Figure 3. This is just one possible state of affairs in the underlying Centipede game. There is no reason why the players couldn't make different choices or hold different beliefs. If they did, we would simply describe that state of affairs instead.

This said, let us stay with Figure 3, and ask: Are the players rational at the state (*In-Out*, t^a, *Out*, t^b)? Start with Ann. Since she assigns probability one to Bob's playing *In*, the strategy *In-Out* is clearly optimal for her. So,

[12] Aumann and Brandenburger (1995).

she is rational. Since Bob assigns probability one to Ann's playing *In-Out*, he is rational in playing *Out*. Next, ask: Does Ann believe that Bob is rational? Ann's type t^a assigns probability one to Bob's playing *In* and being type u^b, so we need to see whether the strategy *In* is optimal for u^b. Since u^b assigns probability $\frac{1}{2}$ to Ann's playing *In-Out* and probability $\frac{1}{2}$ to Ann's playing *In-In*, the answer is that *In* is indeed optimal. (It yields Bob an expected payoff of $\frac{1}{2} \times 3 + \frac{1}{2} \times 6 = 4\frac{1}{2}$, greater than the payoff of 4 he gets from *Out*.) We conclude that Ann does believe Bob to be rational. It is also the case that Bob believes Ann to be rational. This is because Bob's type t^b assigns probability one to Ann's playing *In-Out* and being type t^a, i.e. Bob assigns probability one to Ann's actual strategy-type pair, and we have already checked that Ann is rational. Next: Does Ann believe that Bob believes that she is rational? Now the answer is no. Ann assigns probability one to Bob's being type u^b. And type u^b assigns positive probability (in fact, a probability of $\frac{1}{2}$) to Ann's playing *In-In* and being type t^a. But *In-In* yields this type of Ann–who assigns probability one to Bob's playing *In*–an expected payoff of 3 versus an expected payoff of 4 from playing *In-Out*. Thus u^b assigns positive probability to Ann's being irrational. We see that Ann does not believe that Bob believes she is rational; in fact, she believes that he considers it possible that she is irrational.

Summing up, we have now given a precise formulation of a situation in which Ann and Bob are rational, each believes the other to be rational, and Ann chooses *In*, thereby departing from the BI path. This is exactly the kind of scenario that we asked for at the end of Section 2.[13] True, Ann believes that Bob considers it possible that she is irrational, not rational, but we argued in Section 2 that a departure from CBR such as this is not unreasonable.[14] Indeed, the scenario is quite an intuitive one. There is a clear verbal story associated with it, involving the idea of a 'bluff,' as follows. Ann plays *In-Out*, anticipating a payoff of 4. She believes that Bob will choose *In*, anticipating a payoff of $4\frac{1}{2}$, because she believes that Bob puts probability $\frac{1}{2}$ on her playing *In* (and not *Out*) at her second node. In a sense, then, Ann is

[13] And, in longer versions of Centipede (with more legs), we could use a similar construction to get bigger deviations from the BI path, where the game continues for several rounds before one of the players chooses *Out*.

[14] Sorin (1998) gives a method for making quantitative statements about how far from CBR a particular situation is. Here, we make only qualitative statements of the kind in the text.

trying to bluff Bob into believing that she might play *In* at her second node when, in fact, she plans to play *Out*. (As it turns out, Bob plays *Out*, not *In*, so Ann's bluff fails!)

We appear to have made good progress in bringing theoretical analysis of Centipede in line with the experimental evidence. We have a formal analysis that sounds intuitive and that permits the players to depart from the BI path. There is just one more thing we have to check. Our idea was that while CBR yielded the BI path, sensible-sounding departures from this assumption would allow departures from the BI path. We have checked the second half of this argument, but not the first half. Is it, in fact, true that CBR yields the BI path? If yes, then everything hangs together. But if no, then we have a problem. Our intuition (back in Section 2) was that CBR certainly should yield the BI path. If the mathematical set-up of this section does not deliver this result, then either our original intuition was wrong or the formalism is somehow the wrong one. In either case, our understanding of Centipede would be incomplete.

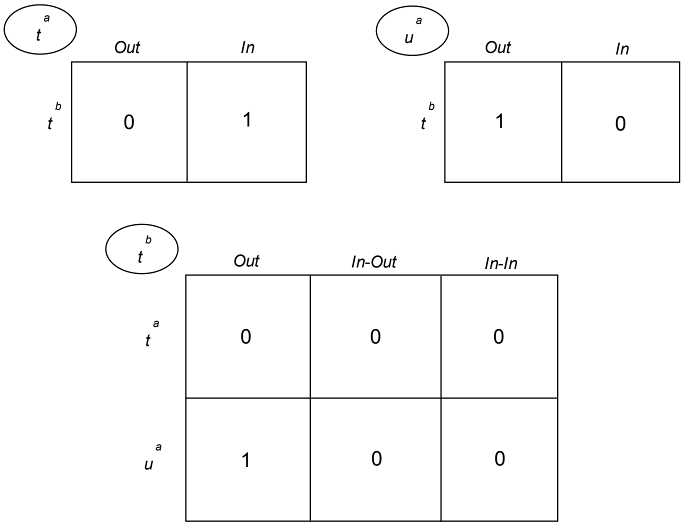
4 Probability Zero

It turns out that the assumption of CBR does not yield the BI path in Centipede–at least it doesn't if CBR is formalized using the tools of the previous section. We now give an example to show how this happens. The example won't lead us to question our original intuition. Rather, it will pinpoint a hole in our mathematical treatment thus far. Of course, we shall then have to look at how the hole can be filled.

Consider Figure 4 below, which is another belief system (different from that in Figure 3) for three-legged Centipede.

Here, there are two possible types of Ann, labelled t^a and u^a, and one possible type of Bob, labelled t^b. Suppose the true state of the world is (*In-Out*, t^a, *Out*, t^b). We assert that CBR holds at this state–and this is so even though Ann plays *In* at her first node, and not *Out* as BI dictates. Certainly, Ann is rational, since type t^a assigns probability one to Bob's playing *In*, making *In-Out* optimal for Ann. As for Bob, his type t^b assigns probability one to Ann's playing *Out*. It follows that Bob is indifferent between playing *In* or

Out himself; he expects a payoff of 1 in either case. In particular, then, he is rational in playing *Out*. For the same reason, Ann believes that Bob is rational since she assigns probability one to Bob's being the type he is (i.e. type t^b) and playing *In*. Bob believes Ann is rational since he assigns probability one to Ann's being type u^a and playing *Out*, and u^a assigns probability one to Bob's playing *Out*, which makes *Out* optimal for Ann. Does Ann believe that Bob believes that she is rational? Yes. Ann assigns probability one to Bob's unique type t^b (as she must), and we already saw that Bob believes that Ann is rational. Note that this is different from the situation in our earlier example (Figure 3), where the answer to this last question was no. A little more thought along these lines indicates that, indeed, CBR holds at the state (*In*-*Out*, t^a, *Out*, t^b). The essential observation is that all three types assign probability one to a rational strategy-type pair of the other player, so that we never 'run into' any irrationality.[15]



t^a	*Out*	*In*
t^b	0	1

u^a	*Out*	*In*
t^b	1	0

t^b	*Out*	*In-Out*	*In-In*
t^a	0	0	0
u^a	1	0	0

Figure 4

[15] This is different from the previous example, where the type u^b of Bob assigned positive probability to an irrational strategy-type pair of Ann. The formal proof that CBR holds at the state (*In*-*Out*, t^a, *Out*, t^b) in the belief system of Figure 4 is an easy induction argument.

But there is something odd about the situation in Figure 4, as the reader may well already have sensed. Ann plays *In* at her first node because she believes that Bob will play *In*. This belief is, in turn, justified by her belief that Bob believes that she plays *Out* (so that Bob then ‘doesn’t mind’ choosing the strategy *In*). But Ann knows that when she goes ahead and actually plays *In*, Bob will see this. So, Ann, by her own action, will falsify the belief (that she plays *Out*) that she is attributing to Bob. Ann’s reasoning looks suspect, if not downright contradictory.

Let us make the same point a bit more formally. Bob’s choice of strategy affects his payoff only in the zero-probability event that Ann plays *In*. His choice doesn’t affect his expected payoff, then, and in this sense he can safely ignore the zero-probability event. This, of course, is just as things are in probability theory, where statements of the form “such-and-such is true with probability one” are commonly made. But the present example shows that in game theory, things are fundamentally different. Here, the key feature is that Ann believes that Bob assigns something probability zero (that she plays *In*), but she can bring that probability-zero event about, and in fact does! So, Ann should realize that Bob will not be able to ignore this event after all. Ann needs to have some model of how Bob treats the unexpected.

Note that this is an intrinsically game-theoretic–i.e. multi-player–effect. It could not arise in a one-player game. Moreover, it is only with recent investigations into the foundations of Backward Induction, and related concepts, that this effect has been clearly isolated. In the next section, we shall review recent work that develops and applies a different probability theory from the usual one, in order to deal with the effect. But before that, let us suggest that we now have good justification for the claim we made back in the Introduction that the BI paradox is a real one. To try to resolve the paradox, we have indeed had to develop new theoretical tools–exactly what Rapaport (1967) says is the hallmark of paradox. The first new tool was that of a belief system (Section 3 above), which we needed because the traditional game tree turned out to be an inadequate description of an interaction. Now, we find that ordinary probability theory is inadequate for game theory, and we are going to need a modified probability theory.

5 Extended Probabilities

Two modifications of probability theory have arisen recently to deal appropriately with zero-probability events. One involves the concept of a *conditional probability system* (CPS), the other the concept of a *lexicographic probability system* (LPS). Here, we will give just the very basic idea of each approach and refer the reader to the relevant technical papers for complete mathematical treatments.[16]

A CPS consists of a number of "If ... then ..." statements, where each statement is of the form "If the player should observe some event E, then he would have beliefs given by an associated probability distribution p_E." In short, a CPS specifies what the player believes, given what he knows. Contrast this with ordinary probability theory, where we specify only one (prior) probability distribution q, which gives the player's beliefs before he learns anything. If the player then comes to know the event E, we calculate the conditional probability distribution $q(\cdot|E)$ in the usual fashion. The key difference arises when $q(E) = 0$–i.e. the player does not expect the event E to occur–for then the conditional probability $q(\cdot|E)$ is undefined. But a CPS does tell us what the player would then believe since the distribution p_E is actually specified as part of the definition of the system.

An LPS deals with the 'unexpected' by simply ruling it out, as follows. An LPS specifies a sequence of probability distributions, with the property that every state receives positive probability under one and exactly one distribution.[17] The interpretation is that the states that get positive probability under the first distribution make up the player's primary 'hypothesis' about what is the true state. But the player recognizes that his primary hypothesis might be mistaken, and so he also forms a secondary hypothesis, consisting of

[16] Battigalli and Siniscalchi (1999, 2001) develop the theory of CPS's and apply it to games. Myerson (1991) provides an axiomatic derivation of CPS's. The CPS concept was introduced by Rényi (1955), who proposed it as an alternative to the usual Kolmogorov theory of probability.

Blume, Brandenburger, and Dekel (1991a) introduce LPS's and give them an axiomatic foundation. Asheim (1999), B-B-D (1991b), Brandenburger and Keisler (2000), and Stalnaker (1996, 1998), inter alia, apply LPS's to games.

Formally, both the CPS and the LPS theories extend the Kolmogorov theory–hence the title of this section.

[17] Technical note: The property as stated applies in the case of a finite state space, but can easily be modified for infinite spaces.

the states that get positive probability under the second distribution. And so on. We can say that the primary states are considered infinitely more likely than the secondary states, which, in turn, are considered infinitely more likely than the tertiary states, etc. But no state, and hence no event E, is considered entirely impossible; nothing is entirely unexpected.

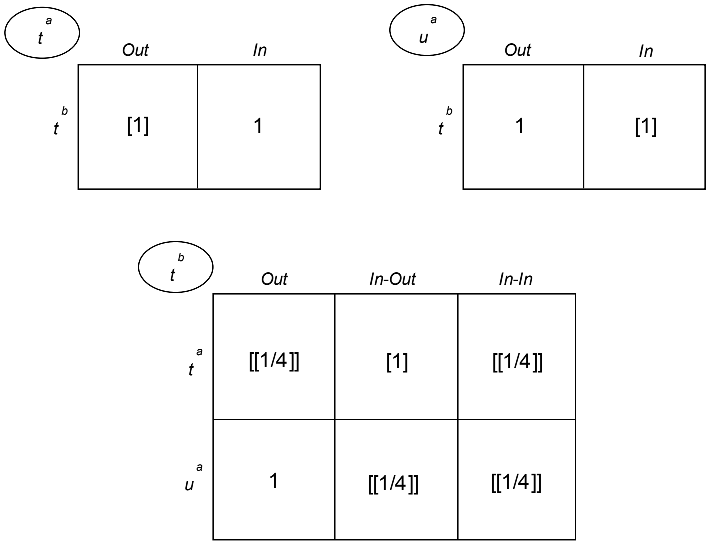
It should be clear that both CPS's and LPS's are well-suited to fixing the problem we found with the belief system of Figure 4. Here, the relevant event E is the event that Ann chooses *In* rather than *Out*. Bob gave this event probability zero, and, with ordinary probabilities, that was all that could be said. But, as we saw, this led to an unsatisfactory analysis, and what was really needed was a model of how Bob would react to this unexpected event. We will now show how LPS's provide such a model. But the treatment with CPS's is equally important, and the reader is urged to refer to the relevant papers.[18]

Figure 5 below is another belief system for three-legged Centipede (Figure 2). It is very similar to the previous belief system (Figure 4), but uses LPS's rather than ordinary probabilities. Start with the (unique) type t^b of Bob. As before, this type assigns probability one to Ann's playing *Out* and being type u^a. But now, this is actually Bob's primary hypothesis about Ann. Bob has also to form a secondary hypothesis, and so on. His secondary hypothesis is depicted by the (point) distribution in square parentheses– i.e. Bob assigns second-order probability one to Ann's playing *In-Out* and being type t^a. Bob's tertiary hypothesis is depicted by the distribution in double square parentheses, i.e. Bob assigns third-order probability of $\frac{1}{4}$ to each of the remaining four strategy-type pairs. Turning to Ann, we see that, just as in Figure 4, her type t^a assigns probability one to Bob's playing *In*.

[18] Technical note: CPS's are appropriate when epistemic analysis is done directly on the game tree; LPS's are used when the analysis is done on the strategic form of the tree. This is clear from the definitions of the two concepts. CPS's involve conditioning on observations, which is, of course, what happens along the game tree. LPS's involve what might be called 'full consideration of possibilities' (Harborne Stuart suggested this term), which fits with choosing a strategy at the outset of the game, without knowing what strategies the other players are choosing.

Here, we use LPS's and so, formally, we are doing analysis on the strategic-form of the tree. But we certainly aren't ignoring the tree. (Our whole interest is in Backward Induction.) Kohlberg and Mertens (1986) explain the philosophy behind taking a strategic-form approach. All this said, the direct analysis on the tree is obviously very important. This is done in Battigalli and Siniscalchi (2001), which we come back to shortly.

But this is now t^a's primary hypotheses. Type t^a's secondary hypothesis is, as it must be, to give probability one to the complementary event that Bob plays *Out*. The situation with Ann's other type u^a is similar.



t^a	*Out*	*In*
t^b	[1]	1

u^a	*Out*	*In*
t^b	1	[1]

t^b	*Out*	*In-Out*	*In-In*
t^a	[[1/4]]	[1]	[[1/4]]
u^a	1	[[1/4]]	[[1/4]]

Figure 5

Continuing to parallel what we did with the belief system of Figure 4, we take the true state to be (*In-Out*, t^a, *Out*, t^b). Before, we had that Bob's type t^b was rational in playing *In* (since he was indifferent between *In* and *Out*). It followed that Ann's type t^a, in assigning probability one to t^b's playing *In*, thereby believed Bob was rational. What is true now, given that we are working with LPS's and not ordinary probabilities? Now, Bob's type t^b is irrational in playing *In*. The reason is that while both *In* and *Out* yield the same expected payoff under Bob's first-order probability distribution (the payoff of 1 that results when Ann plays *Out*), Bob's choice of *Out* yields a higher expected payoff under his second-order distribution than does the choice *In* (a payoff of 4 versus 3). With LPS's, rationality is defined exactly in this lexicographic fashion: One strategy is better than another if the sequence of expected payoffs associated with the first is lexicographically greater than the sequence of expected payoffs associated with the second.[19] The upshot is that (*Out*, t^b) is a rational strategy-type pair of Bob, and (*In*, t^b) is not.

Now, Ann assigns first-order probability zero to (*Out*, t^b). Presumably, then, Ann does not believe that Bob is rational, contrary to the situation in Figure 4. This is indeed so, but to be quite precise, we have first to define the term "belief" in the lexicographic context. (We defined the other ingredient–rationality–above.) With ordinary probabilities, we said that a player believes an event E if he assigns probability one to E. With LPS's the appropriate definition is: A player believes E if he considers states not in E to be infinitely less likely than states in E. That is, the player recognizes that E may not happen, but he is prepared to 'count on' E versus not-E.[20] With this definition in hand, we see that the intuitive thing is indeed true: Ann does not believe that Bob is rational. In fact, since she considers Bob's irrational strategy-type pair (*In*, t^b) to be infinitely more likely than the rational pair (*Out*, t^b), she believes that Bob is irrational.

If Ann does not believe that Bob is rational, certainly we do not have CBR (common belief of rationality), again unlike the situation in Figure 4. But perhaps we caused difficulties for ourselves by the choice of LPS for Bob's type t^b. After all, if Bob assigned second-order probability one to Ann's playing *In-In* rather than *In-Out*, then *In* would yield him a higher expected payoff under his second-order distribution than would the choice *Out* (a payoff of 6 versus 3). Thus, the rational choice for t^b would be *In*, and now Ann would, in fact, believe that Bob is rational. (Recall that her type t^a considers (*In*, t^b) infinitely more likely than (*Out*, t^b).) However, now something else unravels. With this change, Bob would no longer believe that Ann is rational. The reason is that the rational strategy-type pairs of Ann are (*In-Out*, t^a) and (*Out*, u^a). With his LPS as shown in Figure 5, Bob indeed believes that Ann is rational since the other four strategy-type pairs of Ann are each considered infinitely less likely than the two rational pairs. But with the modified LPS, there would now be an irrational strategy-type pair (involving the strategy *In-In*) that Bob considers infinitely more likely

[19] If the first sequence is $(x_1, \ldots, x_n)$ and the second $(y_1, \ldots, y_n)$, then the requirement is that there is a $j = 1, \ldots, n$ such that $x_j > y_j$ and $x_i = y_i$ for all $i < j$.

[20] See Brandenburger-Keisler (2000) for a full discussion of this definition. B-K argue that in the lexicographic case the terminology "the player assumes E" works better than "the player believes E." But we shall stick to talking about belief rather than assumption here.

than the rational pair (*In*-*Out*, t^a).[21] It would no longer be true that Bob believes Ann to be rational.

It turns out that this unraveling of CBR is unavoidable. When rationality and belief are formalized using LPS's, as in this section, we have the following theorem: Fix n-legged Centipede, for some n. Fix also an associated belief system with LPS's, and a state of the system at which there is CBR. Then, Ann will play *Out*.[22]

With this result, we have found what we said at the end of Section 3 was missing. We wanted CBR to yield the BI path in Centipede. This would be a kind of baseline, we said, given which we could then see how sensible-sounding departures from CBR allow departures from the BI path. Section 4 showed that this baseline result was false if we formalized the ingredients (rationality and belief) using ordinary probabilities. Now we have an alternative formalism, using an extended notion of probability, in which the baseline result is true. Moreover, departures from the BI path are as easy to arrange in the LPS set-up as in the usual set-up. (In Section 3, we gave an example of a non-BI outcome using ordinary probabilities. The same example can easily be modified to work with LPS's, or with CPS's for that matter.) So, at this point, we have a pretty complete, and also intuitively satisfying, picture of Centipede.

But it turns out that Centipede is a special game. There are other game trees in which CBR, even when formulated with LPS's, does not yield the BI path. We give an example next, and go on to describe the new condition needed to get the BI outcome. Examination of this new condition will also lead us to another paradox, different from the BI Paradox with which we started, and to a current area of research.

6 Adding Types

Consider the game depicted in Figure 6, which is one of pure coordination. Ann and Bob rank the outcomes the same way; in particular, there is a

[21] This latter pair would have to get positive probability under one of Bob's higher-order distributions (beyond the second order), since every strategy-type pair gets positive probability under some distribution.

[22] Brandenburger and Friedenberg (2002).

common best outcome, which arises when Ann plays *In-In* and Bob plays *In*. This is also the BI outcome, as can easily be checked. Even so, we now give a belief system with LPS's for this game, such that CBR holds and yet Ann plays *Out*. This will establish the claim made above that the conditions that give the BI outcome in the case of Centipede do not always give the BI outcome.

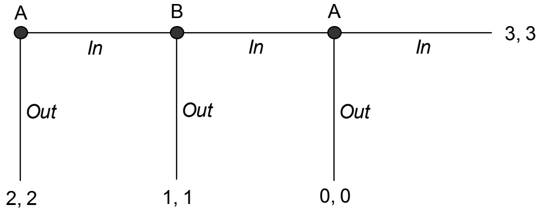


Figure 6

The belief system is depicted in Figure 7. (Note that there happens to be just one type of either player.) We suppose that the true state is (*Out*, t^a, *Out*, t^b). Then, Ann assigns first-order probability one to Bob's playing *Out*, and so is rational in playing *Out* herself. Bob assigns first-order probability one to Ann's playing *Out*, and then second-order probability one to her playing *In-Out*. This makes *Out* rational for Bob. Moreover, Ann believes Bob to be rational since she considers (*Out*, t^b), which we just saw is a rational strategy-type pair of Bob, to be infinitely more likely than the irrational pair (*In*, t^b). Turning back to Bob, does he believe Ann to be rational? Yes. Of the three strategy-type pairs of Ann, only (*Out*, t^a) is rational; both (*In-In*, t^a) and (*In-Out*, t^a) are irrational. Since Bob considers the second and third pairs infinitely less likely than the first pair, he does indeed believe that Ann is rational. Some further thought leads to the conclusion that, in fact, CBR holds at the state (*Out*, t^a, *Out*, t^b).[23]

What is going on here? After all, the $(3, 3)$ outcome, quite apart from being the BI outcome, seems very salient, almost inevitable. It is the best outcome for both players. How then, could rationality–let alone CBR–allow the players to get the $(2, 2)$ outcome instead? In particular, isn't there something 'funny' about Bob's beliefs in the system of Figure 7? Okay, he assigns

[23] The formal proof of this assertion is a straightforward induction.

first-order probability one to Ann's choosing *Out*. But, after that, shouldn't he consider *In*-*In* a more probable choice by Ann than *In*-*Out*? Surely, if Ann forgoes the payoff of 2 she can get by playing *Out* at her first node, she must be planning to play *In* at her second node and thereby get 3. Playing *Out* at her second node gets her 0, less than the payoff of 2 she gave up at her first node. And, if Bob does consider *In*-*In* more probable than *In*-*Out*, then he himself will rationally play *In* rather than *Out*. But then it seems that Ann would rationally play *In*-*In*, and not *Out* as we had earlier. We end up at the BI outcome $(3, 3)$, and not the $(2, 2)$ outcome as in Figure 7.

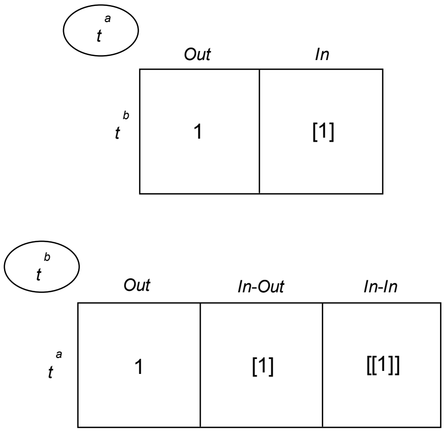


Figure 7

This is a very interesting line of argument, and we are going to examine it shortly. But, before that, we emphasize that there is nothing formally incorrect about our scenario where CBR holds and the $(2, 2)$ outcome results. True, we just said that, contrary to that scenario, Bob should consider *In*-*In* a more probable choice by Ann than *In*-*Out*. But must he? The key is to see that Ann believes that Bob will play *Out*. In this case, both *In*-*In* and *In*-*Out* are irrational choices by Ann; neither is rational! The assumption that Bob believes Ann to be rational therefore has much less bite than one might think. Bob must give first-order probability one to the rational choice for Ann, namely *Out*, but that is all. He is 'free' to consider *In*-*Out* more probable than *In*-*In*, as indeed he does.

We can say that the players find themselves, quite literally, trapped in the wrong belief system. Their particular beliefs prevent them from getting the $(3, 3)$ outcome that they would both prefer. The situation described by the belief system of Figure 7 may be a surprising one, at least at first blush, but there does not appear to be anything pathological about it. Indeed, perhaps it points to an interesting way in which the players in a game find themselves 'trapped by their beliefs.' Still, our question remains: What would the belief system have to look like for CBR to lead to $(3, 3)$?

To proceed, note that the 'problem' in the belief system of Figure 7 is that it does not include the 'right' kinds of beliefs. Specifically, there is no type of Ann that believes that Bob will play *In*. So, suppose now that there were such a type. That is, we add a second type of Ann (labelled u^a, say) that has the 'opposite' belief about Bob–i.e. that gives first-order probability one to Bob's playing *In*. With this change, there would now be two rational strategy-type pairs of Ann, viz., the pair (*Out*, t^a) as before, but also the pair (*In-In*, u^a). As a result, the assumption that Bob believes Ann to be rational would now require Bob to consider *In-In* infinitely more likely than *In-Out*. But then, as we said above, Bob himself would rationally play *In* rather than *Out*. And then, presumably, Ann would play *In-In*. If so, the $(3, 3)$ outcome would result, just as we want.

Intuitively, what we are doing here is letting Bob 'do his best' to rationalize Ann's moving *In*. That is, by adding more types to the system, we are letting Bob try to find a belief to attribute to Ann that explains her behavior.[24] (In the present example, that belief is that Bob will play *In*, a belief that was absent from the system of Figure 7.) And it seems that by

[24] Note on the literature: This is the notion of *forward induction*, introduced by Kohlberg and Mertens (1986). A player looks back (!) up the tree, to take account of moves that another player could have made but didn't make, to try to infer something about what that player must believe to have made the choice he did. In a sense then, the player tries to 'induce forwards' from other players' past behavior to their future behavior. By contrast, the term *backward induction* obviously reflects the fact that the backward-induction algorithm starts at the end of the tree and then works backwards. In a sense, past behavior is inferred from future behavior.

In fact, what we now see is that one really can't talk separately about backwards and forwards reasoning in the tree. In trying to understand what is underneath backward induction, one has to bring in so-called forward induction. Ex post, this is perhaps no surprise. The tree has to be analyzed as a whole, and cannot be understood by going only backwards or only forwards.

doing this, we may be able to make the assumption of CBR yield the BI path after all.

7 Paradox Lost

The idea of using a large belief system that contains many different types of each player is a key step forward in the theoretical understanding of BI. It is due to the previously cited Stalnaker (1998) and Battigalli and Siniscalchi (2001). The latter paper, in fact, uses what is called a *complete* belief system, which is one that contains, in a certain sense, every possible type of each player.[25] This is an elegant formulation that solves at one stroke the problem we had above of having to add more types to the belief system. Now, there are no missing types; they are all automatically present. Moreover, Battigalli-Siniscalchi show that the completeness assumption is, indeed, the final missing ingredient in our quest for the conditions that yield BI. They prove that if CBR is formulated in a complete belief system, then, yes, the play of the game will always be along the BI path.[26]

Just to be clear, let us also note that the terms "backward induction" and "forward induction" aren't parallel. The first refers to a precise algorithm, while the second refers to an informal notion. Of course, we now see that an understanding of the first concept involves formalizing the second concept, anyway.

[25] Thus, the belief system of Figure 7 is certainly not complete. For one thing, as we noted above, there is no type of Ann that believes that Bob will play *In*.

We give a more precise definition of completeness in Section 9 below. For the formal definition, see Brandenburger and Keisler (1999). The concept is closely related to *universality* (Armbruster and Boge 1979, Boge and Eisele 1979, Mertens and Zamir 1985, Brandenburger and Dekel 1993, et al.).

[26] This is a loose statement of their theorem. Battigalli-Siniscalchi use CPS's, i.e. the conditional probability systems we described at the beginning of Section 5. So, they have to give formal definitions of "rationality" and "belief" in the CPS context, just as we had to do above in the LPS context. B-S show that their conditions imply that the players choose so-called extensive-form rationalizable strategies (Pearce 1984), which are known to yield the BI path in a perfect-information (PI) tree.

Brandenburger-Keisler (2000) use LPS's (as in this survey) together with completeness. They get that the players choose iteratively admissible strategies, which also yield the BI path in a PI tree.

Note on the literature: As we warned in the Introduction, this survey is not comprehensive. There are other papers giving conditions for BI that we do not discuss here. See Aumann (1995, 1998), Samet (1996), and Halpern (1998, 1999), inter alia; and also the

We said back in Section 2 that if it is really the case that it is CBR–and not mere rationality of the players–that yields the BI path, then this would provide a nice resolution of the BI paradox.[27] The reason we gave was that the hypothesis of CBR is a very stringent one, and so departures from it are only to be expected. Now we see that there are games, such as that in Figure 6, where even CBR (appropriately formulated with LPS's) does not yield the BI path. In addition to CBR, we have to assume that the players operate in a complete belief system. This gives us still another reason, then, why players might not actually play the BI path; they simply may not be in a complete belief system.

Summing up, the BI path is often not played in practice. But, contrary to what used to be thought, the BI path is also not an inevitable game-theoretic prediction. It rests on a number of specific assumptions about the players' rationality and beliefs, as we have described, and there is nothing inevitable about these assumptions. They may or may not hold. And, of course, if the BI path is not played in a particular case, then we simply conclude that at least one of these assumptions did not hold.

Paradox definitely lost? Not quite.

8 Paradox Regained

Let us go back to what we have just seen is a crucial condition for BI, that the players are in a complete belief system–i.e. a system that contains every possible type of each player. Now, whether this is a meaningful idea is not immediately obvious. Does such a system actually exist? There is good reason to ask the question. After all, a system of all types sounds rather like the kinds of "sets of everything" that are well known to cause difficulties in

exchange between Binmore (1996) and Aumann (1996). Formally, these papers are rather different from those we are surveying. (For one thing, they use *knowledge systems* rather than belief systems.) But they do appear to involve similar issues at the conceptual level; see the excellent presentation in Halpern (1998).

[27] By now, we have certainly seen that the rationality of the players alone does not suffice for BI. Witness the example of Figure 3, which, as we noted earlier, could be modified to work with LPS's (or CPS's) as well as with ordinary probabilities.

mathematics. Indeed, we already mentioned the most famous of these difficulties, viz. Russell's Paradox, in the Introduction.[28] So, somewhat strangely, we have now come full circle. We began by noting that paradoxes–Russell's Paradox among them–have played important roles in several disciplines, and said that we were going to make the case that the BI Paradox has worked similarly in game theory. Now, we find that there may be a real connection, not just a parallel, between the BI Paradox and Russell's Paradox in particular.

Here is the argument that a complete belief system does not exist. First, recall that the notion of self-reference is at the heart of the impossibility results of mathematical logic–Gödel, Russell, Tarski, Turing, etc.[29] Now, self-reference looks to be pretty much built in to the notion of a game, as follows. A type of Ann has a belief about what type Bob is. But each type of Bob has a belief about what type Ann is. Thus, a type of Ann ends up referring to itself, in some sense. To get the actual impossibility, we have to find the right–i.e. contradictory–self-referential statement involving the players' beliefs. This turns out to be:

Ann believes that Bob believes that Ann believes that
Bob has a false belief about Ann

To see the contradiction, ask: Does Ann believe that Bob has a false belief (about Ann)? If so, then the statement that Ann believes Bob believes ("Ann believes that Bob has a false belief about Ann") is true. But then Ann does not believe that Bob has a false belief, and we get a contradiction. So, suppose, instead, that Ann does not believe that Bob has a false belief. But now the statement that Ann believes Bob believes is false, and so Ann does believe that Bob has a false belief, and we again we get a contradiction. Thus, the above configuration of beliefs is impossible. But a complete belief system would, presumably, contain this configuration of beliefs (among many other configurations of beliefs in it). The conclusion is that such a system cannot exist.

[28] To remind the reader, Russell's Paradox concerns "The collection of all sets which are not members of themselves." The contradiction arises if this collection is a set, since then it is a member of itself if and only if it is not a member of itself. A good reference on paradoxes in logic is Barwise and Etchemendy (1987).

[29] Marek and Mycielski (2001) is a nice recent survey.

Of course, we now have to explain how this impossibility result fits with our mention of complete belief systems in the previous section! Do such systems exist or not? The quick answer is that this depends on exactly how complete we require the system to be. To get to a more helpful answer, we must first note that the impossibility argument we just gave is only a verbal one, and a bit slippery at that. But it can be made mathematically precise,[30] as we shall sketch below. Making the argument precise will allow us to pinpoint what determines whether we get a 'positive' (existence) result or a 'negative' (non-existence) result.

9 Language

A brief recap may be helpful. Back in Section 3, we took the step of making the players' beliefs about the game a part of the game. This took us from the traditional game tree to the belief-system concept–and what more generally is called the epistemic approach to game theory. We are now going to add another new ingredient to the analysis of games. This is the idea of specifying a precise mathematical *language* that a player uses to formulate his beliefs about the game.[31] Alternatively put, the epistemic approach says that a player thinks or reasons about the game. He forms a view about the other players–their strategies, beliefs, rationality, etc.–which affects his own choice of strategy. Now we bring in an explicit model of that thinking or reasoning process.

We'll say in a moment just what language a player might use to think about the game. But first let us note that once we have the idea of such a language, we can give a more precise definition of completeness than we have so far. A belief system will be *complete* if every belief of a player that can be stated in the given language is, in fact, present in the system; otherwise, it is *incomplete*. Thus, we first say how a player thinks–i.e. what language he uses. Then we ask whether or not a belief system contains everything he can think of.

[30] Brandenburger-Keisler (1999).

[31] Papers on this idea include Aumann (1999); Fagin, Geanakoplos, Halpern, and Vardi (1999); Heifetz (1999); and other papers in the Special Issue on Interactive Epistemology, *International Journal of Game Theory*, 28, 1999. An important precursor is Samet (1990).

Now, back to the matter of what particular language a player might use. There is an obvious 'baseline' assumption, which is that the belief system itself is his language. (This might sound a little circular, but it isn't. We construct exactly this situation in the next paragraph.) Why look at this case? Because the whole idea of a belief system is that it is a mathematical structure that we, the game theorists, use to think about the game.[32] Unless we want to accord the theorist a 'privileged' position that is somehow denied to the players, it is only natural to ask what happens if a player can think about the game the same way.[33]

So, we want to make the belief system into a formal mathematical language with which a player thinks about the game. This is where the tools of mathematical logic come in, to set up the language properly. We won't give the details here,[34] but will simply assume that this has been done. Once this is done, we can give a precise statement of the impossibility theorem that we described informally in the previous section. We begin with a belief system. We then set up the formal language that is implied by the system.

[32] Recall our discussion of this point in Section 3.

[33] It is important to be very careful about what we are, and what we are not, saying here. We are saying that the player himself is aware of, and uses, the belief system. We are not saying that he is aware of the true state of the system. To assume the latter would be quite contrary to the spirit of the epistemic approach to game theory. It says that if we, the analysts, make a certain prediction about the play of the game–i.e. specify a certain state as the true state–that prediction must be available to the players themselves. So, if we predict that Ann will play a certain one of her strategies, and that Bob will play a certain one of his strategies, then Ann and Bob will know this. But then, provided Ann is maximizing, her strategy must be optimal against Bob's strategy; and vice versa with Ann and Bob interchanged. In short, the pair of strategies must constitute a Nash equilibrium.

This is an old line of argument in game theory–saying that rationality of the players alone, with no other assumptions, implies Nash equilibrium. The 'trick' here is that in assuming that the players know the true state, we rule out the possibility that a player is ever mistaken about another's strategy choice. By contrast, the epistemic approach definitely does not require the players to know the true state. It can therefore model situations where players have incorrect beliefs, as we have already mentioned, in addition to the special case where beliefs are correct. (And Nash equilibrium does not then follow from rationality alone; further assumptions are needed. See Aumann-Brandenburger 1995.)

[34] See Brandenburger-Keisler (1999). The language B-K set up is a first-order logic, with symbols for elements of the belief system. This seems like the natural choice, given that first-order logic is widely considered to be the basic language of mathematics (see e.g. Barwise 1977). A game is a mathematical structure, so a player uses first-order logic to reason about it.

And we then ask whether the system is complete relative to this language–i.e. whether every belief that can be stated in this language is present in the system. This is the precise question, to which the impossibility theorem gives a negative answer: no system contains all these beliefs.[35]

What, then, of our use back in Section 7 of completeness as a condition for Backward Induction? The situation is now clear. Completeness of a belief system is defined relative to a language. (We have to say how the players think, before we can say whether everything they can think of is present.) If we choose a different language from the one we just looked at, perhaps we can get a belief system that is complete relative to that language.

Presumably, the theorems on BI that we described in Section 7 must work this way. They must, implicitly at least, use the 'right' language. That is, the language must be rich enough that a belief system that is complete relative to this language contains enough types to get us BI.[36] At the same time the language cannot be as rich as the one we just looked at since, then, we know we cannot get completeness.[37] But there aren't actually any explicit formal languages in these treatments since they are carried out using probability theory and not mathematical logic.[38]

So, here we arrive at an open research problem: give conditions for BI where the various ingredients are defined in explicitly logical terms. Part of doing this, as we have been discussing, will be getting just the right language for the players. Finding this language will tell us exactly how the players must be reasoning about the game, if the BI outcome is to result.

[35] The proof is essentially the one we gave in the text in Section 8. The key is to show that all the beliefs in the italicized configuration of beliefs there are definable in our formal language. They are, so a complete system would have to contain this configuration. But we saw that this configuration is impossible. Q.E.D.

[36] Recall the example of Section 6, which showed that if CBR is formulated in a belief system with only certain types present, then the BI path need not result.

[37] In particular, the language must be such that the italicized configuration of beliefs at the beginning of Section 8 cannot arise in that language. The contradiction we got won't then arise.

[38] Just as we used only probability theory in this survey–until Section 8 when we started mentioning mathematical logic.

10 Conclusion

This new problem may sound rather like the problem we began the paper with–viz. to give conditions for BI. So, have we simply gone around in a circle? No. The theoretical basis of BI is now a lot better understood than it used to be, as this survey has tried to show. We now have precise conditions on the players' rationality and beliefs that lead to BI, and we can see that these conditions are far from inevitable. Game trees and BI are no longer synonymous. The BI Paradox is resolved–at least at a certain level. But, not surprisingly, there are further levels of analysis to be done. Having a precise logic of BI will be another level of understanding.

We don't have this logic yet,[39] so saying too much more here would be premature. But it is tempting to hope that when we do have this logic, we will learn something about reasoning in games in general, beyond the specifics of BI. If so, this will be another way in which the BI Paradox has spurred the development of new conceptual frameworks in game theory.

[39] Though Meier (2001) is a very promising step in this direction.

References

[1] Armbruster, W., and W. Boge, "Bayesian Game Theory," in O. Moeschlin and D. Pallaschke (eds.), *Game Theory and Related Topics*, North-Holland, Amsterdam, 1979.

[2] Asheim, G., "Common Knowledge of Proper Consistency," unpublished, Department of Economics, University of Oslo, 1999.

[3] Aumann, R., "Agreeing to Disagree," *Ann. Stat.*, 4, 1976, 1236-1239.

[4] Aumann, R., "Game Theory," in Eatwell, J., M. Milgate, and P. Newman (eds.), *The New Palgrave: Game Theory*, Norton, 1989.

[5] Aumann, R., "Backward Induction and Common Knowledge of Rationality," *Games Econ. Behav.*, 8, 1995, 6-19.

[6] Aumann, R., "Reply to Binmore," *Games Econ. Behav.*, 17, 1996, 138-146.

[7] Aumann, R., "On the Centipede Game," *Games Econ. Behav.*, 23, 1998, 97-105.

[8] Aumann, R., "Interactive Epistemology, I and II," *Int. J. Game Theory*, 28, 1999, 263-300 and 301-314.

[9] Aumann, R., and A. Brandenburger, "Epistemic Conditions for Nash Equilibrium," *Econometrica*, 63, 1995, 1161-1180.

[10] Barrow, J., *Impossibility*, Oxford University Press, 1998.

[11] Barwise, J., "An Introduction to First-Order Logic," in J. Barwise (ed.), *Handbook of Mathematical Logic*, Elsevier Science 1977.

[12] Barwise, J., and J. Etchemendy, *The Liar: An Essay on Truth and Circularity*, Oxford University Press, 1987.

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

[14] Battigalli, P., and M. Siniscalchi, "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games," *J. Econ. Theory*, 88, 1999, 188-230.

[15] Battigalli, P., and M. Siniscalchi, "Strong Belief and Forward Induction Reasoning," 2001. Forthcoming in *J. Econ. Theory.*

[16] Ben Porath, E., "Rationality, Nash Equilibrium, and Backward Induction in Perfect Information Games," *Rev. Econ. Studies*, 64, 1997, 23-46.

[17] Bicchieri, C., "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntniss*, 30, 1989, 69-85.

[18] Bicchieri, C., "Knowledge-Dependent Games: Backward Induction," in Bicchieri, C., and M.L. Della Chiara (eds.), *Knowledge, Belief and Strategic Interaction*, Cambridge University Press, 1992.

[19] Binmore, K., "Modelling Rational Players I," *Econ. Phil.*, 3, 1987, 179-214.

[20] Binmore, K., "A Note on Backward Induction," *Games Econ. Behav.*, 17, 1996, 135-137.

[21] Blume, L., A. Brandenburger, and E. Dekel, "Lexicographic Probabilities and Choice under Uncertainty," *Econometrica*, 59, 1991a, 61-79.

[22] Blume, L., A. Brandenburger, and E. Dekel, "Lexicographic Probabilities and Equilibrium Refinements," *Econometrica*, 59, 1991b, 81-98.

[23] Boge, W., and T. Eisele, "On Solutions of Bayesian Games," *Int. J. Game Theory*, 8, 1979, 193-215.

[24] Bonanno, G., "The Logic of Rational Play in Games with Perfect Information," *Econ. Phil.*, 7, 1991, 37-65.

[25] Brandenburger, A., and E. Dekel, "Hierarchies of Beliefs and Common Knowledge," *J. Econ. Theory*, 59, 1993, 189-198.

[26] Brandenburger, A., and A. Friedenberg, "Common Assumption of Rationality in Games," 2002. Available at www.stern.nyu.edu/∼abranden.

[27] Brandenburger, A., and H.J. Keisler, "An Impossibility Theorem on Beliefs in Games," unpublished, 1999.

[28] Brandenburger, A., and H.J. Keisler, "Epistemic Conditions for Iterated Admissibility," unpublished, 2000.

[29] Dekel, E., and F. Gul, "Rationality and Knowledge in Game Theory," in Kreps, D., and K. Wallis (eds.), *Advances in Economics and Econometrics*, Vol. 1, Cambridge University Press, 1997, 87-172.

[30] Fagin, R., J. Geanakoplos, J. Halpern, and M. Vardi, "The Hierarchical Approach to Modeling Knowledge and Common Knowledge," *Int. J. Game Theory*, 28, 1999, 331-365.

[31] Halpern, J., "Substantive Rationality and Backward Induction," 1998, forthcoming in *Games Econ. Behav.*.

[32] Halpern, J., "Hypothetical Knowledge and Counterfactual Reasoning," *Int. J. Game Theory*, 28, 1999, 315-330.

[33] Harsanyi, J., "Games with Incomplete Information Played by 'Bayesian' Players, I-III," *Man. Sci.*, 14, 1967-68, 159-182, 320-334, 486-502.

[34] Heifetz, A., "How Canonical is the Canonical Model? A Comment on Aumann's Interactive Epistemology," *Int. J. Game Theory*, 28, 1999, 435-442.

[35] Kohlberg, E., and J.-F. Mertens, "On the Strategic Stability of Equilibria," *Econometrica*, 1986, 54, 1003-1037.

[36] Marek, V.W., and J. Mycielski, "Foundations of Mathematics in the Twentieth Century," *Am. Math. Monthly*, 108, 2001, 449-468.

[37] McKelvey, R., and T. Palfrey, "An Experimental Study of the Centipede Game," *Econometrica*, 60, 1992, 803-36.

[38] Meier, M., "Non-Probabilistic Conditional Belief Structures," unpublished, University of Bielefeld, 2001.

[39] Mertens, J-F., and S. Zamir, "Formulation of Bayesian Analysis for Games with Incomplete Information," *Int. J. Game Theory*, 14, 1985, 1-29.

[40] Myerson, R., *Game Theory*, Harvard University Press, 1991.

[41] Pearce, D., "Rational Strategic Behavior and the Problem of Perfection," *Econometrica*, 52, 1984, 1029-1050.

[42] Rapaport, A., "Escape from Paradox," *Scientific American*, July 1967, pp.50-56; quoted in Barrow (1998, pp.12-13).

[43] Reny, P., "Rationality in Extensive Form Games," *J. Econ. Perspectives*, 6, 1992, 103-118.

[44] Rényi, A., "On a New Axiomatic Theory of Probability," *Acta Math. Acad. Sci. Hung.*, 6, 1955, 285-335.

[45] Rosenthal, R., "Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox," *J. Econ. Theory*, 25, 1981, 92-100.

[46] Samet, D., "Ignoring Ignorance and Agreeing to Disagree," *J. Econ. Theory*, 52, 1990, 190-207.

[47] Samet, D., "Hypothetical Knowledge and Games with Perfect Information," *Games Econ. Behav.*, 17, 1996, 230-251.

[48] Sigmund, K., and M. Nowak, "A Tale of Two Selves," *Science*, 290, 3 November 2000, 949-950.

[49] Sorin, S. "On the Impact of an Event," *Int. J. Game Theory*, 27, 1998, 315-330.

[50] Stalnaker, R., "Knowledge, Belief and Counterfactual Reasoning in Games," *Econ. Phil.*, 12, 1996, 133-163.

[51] Stalnaker, R., "Belief Revision in Games: Forward and Backward Induction," *Math. Soc. Sci.*, 36, 1998, 31-56.