Int J Game Theory (2007) 35:465–492
DOI 10.1007/s00182-006-0061-2

ORIGINAL PAPER

The power of paradox: some recent developments in interactive epistemology

Adam Brandenburger

Revised: 5 November 2006 / Accepted: 6 November 2006 / Published online: 20 January 2007

Abstract Paradoxes of game-theoretic reasoning have played an important role in spurring developments in interactive epistemology, the area in game theory that studies the role of the players' beliefs, knowledge, etc. This paper describes two such paradoxes – one concerning backward induction, the other iterated weak dominance. We start with the basic epistemic condition of "rationality and common belief of rationality" in a game, describe various 'refinements' of this condition that have been proposed, and explain how these refinements resolve the two paradoxes. We will see that a unified epistemic picture of game theory emerges. We end with some new foundational questions uncovered by the epistemic program.

1 Introduction

The word "paradox" means, literally, "beyond belief." So it seems fitting to use the word to describe some problems that have stimulated much recent work in

This survey owes a great deal to joint work and many conversations with Robert Aumann, Amanda Friedenberg, Jerry Keisler, and Harborne Stuart. Scott Ashworth, John Asker, Carliss Baldwin, Heski Bar-Isaac, Pierpaolo Battigalli, Ken Corts, Konrad Grabiszewski, Joe Halpern, Rena Henderson, Martin Meier, Martin Rechenauer, and participants in various seminars provided important input. The associate editor and referees made very helpful suggestions and observations. Financial support from Harvard Business School and the Stern School of Business is gratefully acknowledged.

A. Brandenburger (✉)
Stern School of Business, New York University,
44 West Fourth Street, New York, 10012 NY, USA
e-mail: adam.brandenburger@stern.nyu.edu
URL: http://www.stern.nyu.edu/∼abranden

the epistemology of games, which is the study of the role of the players' beliefs, knowledge, etc. in games.

Rapaport (1967, p. 50) writes about the productive role paradoxes can play: "Whenever, in any discipline, we discover a problem that cannot be solved within the conceptual framework that supposedly should apply, we experience shock. The shock may compel us to discard the old framework and adopt a new one."[1] The goal of this survey is to suggest that, as an example of this effect, game-theoretic paradoxes have helped prompt the development of new ideas in the foundations of game theory.

2 Two paradoxes

We will look at two paradoxes in game theory—one in the tree and one in the matrix.

The paradox in the tree concerns backward induction (BI). The reasoning behind BI seems clear. If Ann, the last player to move, is rational, she will make the BI choice. If Bob, the second-to-last player to move, is rational and thinks Ann is rational, he will make the choice that is maximal given that Ann makes the BI choice – i.e., he too will make the BI choice. And so on back in the tree. But as many people have pointed out, this reasoning is flawed. For example, BI applied to Centipede (Rosenthal 1981) says the first player will end the game immediately. In their textbook Mas-Colell et al. (1995, p. 282), explain the problem with this conclusion:

> Consider player 1's initial decision to say "stop." For this to be rational, player 1 must be pretty sure that if instead she says "continue," player 2 will say "stop" at her first turn. Indeed, "continue" would be better for player 1 as long as she could be sure that player 2 would say "continue" at her next move. Why might player 2 respond to player 1 saying "continue" by also saying "continue"?... [Because] once she sees that player 1 has chosen "continue"–an event that should never happen...–she might entertain the possibility that player 1 is not rational.... If, as a result, she thinks that player 1 would say "continue" at her next move if given the chance, then player 2 would want to say "continue" herself.

Just what argument does lead to BI? Equally, if the BI path isn't played, what assumptions don't then hold? This has been a big puzzle in game theory.

The second puzzle is in the matrix and concerns weak dominance – specifically, iterated weak dominance (iterated admissibility, or IA). IA is an old concept in game theory, going back at least to Gale (1953). Like BI, it is a powerful solution concept, delivering sharp answers in many games. Also like BI, the reasoning behind IA seems clear. Suppose Ann is rational in the sense that she avoids any inadmissible strategies. If Ann thinks Bob is rational in the same way, she can eliminate from consideration any of Bob's strategies that are inadmissible. So, if Ann is rational and thinks Bob is rational, she should choose only a strategy that is admissible in the submatrix that results after

[1] Both the literal meaning of "paradox" above and this quote are in Barrow (1998, p. 12).

deleting Bob's inadmissible strategies. And so on until reaching the IA set. But this reasoning is flawed, too. Mas-Colell et al. (1995, p. 240) state the problem:

> [T]he argument for deletion of a weakly dominated strategy for player i is that he contemplates the possibility that every strategy combination of his rivals occurs with positive probability. However, this hypothesis clashes with the logic of iterated deletion, which assumes, precisely, that eliminated strategies are not expected to occur.

Can a sound argument be made for IA? This is a second big puzzle in game theory.

These two puzzles – or paradoxes – are really both about the fundamental problem of what it means to say that the players in a game are rational, each thinks the other players are rational, etc. The meaning of this has to be understood both in the matrix and in the tree. That is why there are two paradoxes.

3 Overview

Influential early papers on the BI paradox include Binmore (1987), Bicchieri (1988; 1989), Basu (1990), Bonanno (1991), and Reny (1992). Samuelson (1992) and Börgers and Samuelson (1992) are important papers pointing out the difficulties with IA. In this survey, we will focus on some of the recent epistemic literature on these topics.

The hallmark of the epistemic approach to game theory is that it adds to the traditional description of a game a mathematical language for talking about the rationality or irrationality of the players, their beliefs or knowledge, and related ideas. As such, the approach sounds tailor-made to address the paradoxes.

We will see, though, that several challenges have to be overcome to get languages that can express the issues well. In the next section, we lay out a very basic epistemic framework that can be used to analyze game matrices and ordinary (strong) dominance. We will examine the problems that arise in trying to extend the framework to deal with the tree, or with weak dominance in the matrix, and look at how to overcome these problems. With this background, we will be ready to follow the stories of tackling BI and IA, respectively. We will find resolutions of the two paradoxes. But there will also be some surprises along the way—some new challenges and even theoretical limits in game theory will emerge.

Of course, this paper is not a substitute for the technical papers in the field. It is a partial survey that tries to pull together some of the recent epistemic work.

4 Epistemic analysis

The first step in the epistemic approach to game theory is to enrich the classical description of a game by adding sets of **types** for each of the players. The apparatus of types goes back to Harsanyi (1967–1968), who introduced it as a way to talk formally about the players' beliefs about the payoffs in a game, their beliefs about other players' beliefs about the payoffs, and so on. But the

technique is equally useful to talk about uncertainty about the actual play of the game, either separate from or in addition to uncertainty about the structure of the game. A feature of the epistemic approach is putting these two sources of uncertainty on an equal footing.[2]

We will give a definition of a type structure as commonly used in the epistemic literature, and an example of its use.

Fix an n-player finite strategic-form game $\langle S^1, \ldots S^n, \pi^1, \ldots, \pi^n \rangle$. Some notation: Given sets $X^1, \ldots, X^n$, let $X = \times_{i=1}^n X^i$ and $X^{-i} = \times_{j \neq i} X^j$. Also, given a compact metrizable space Ω, write $\mathcal{M}(\Omega)$ for the space of all Borel probability measures on Ω , where $\mathcal{M}(\Omega)$ is endowed with the topology of weak convergence (and so is again compact metrizable).

Definition 4.1 *An* $(S^1, \ldots, S^n)$-based type structure *is a structure*

$$\langle S^1, \ldots, S^n; T^1, \ldots, T^n; \lambda^1, \ldots, \lambda^n \rangle,$$

where each T^i *is a compact metrizable space, and each* $\lambda^i : T^i \to \mathcal{M}(S^{-i} \times T^{-i})$ *is continuous. Members of* T^i *are called* types *for player i. Members of* $S \times T$ *are called* states *(*of the world*).*

A particular state $(s^1, t^1, \ldots, s^n, t^n)$ describes the strategy chosen by each player, and also each player's type. Moreover, a type t^i for player i induces a probability measure on the strategies that the players $j \neq i$ can choose. (Go from T^i to $\mathcal{M}(S^{-i} \times T^{-i})$ and marginalize to $\mathcal{M}(S^{-i})$.) Call this player i's first-order belief. Type t^i also induces a probability measure on the strategies and first-order beliefs of the players $j \neq i$. (Go from T^i to $\mathcal{M}(S^{-i} \times T^{-i})$, and then to $\mathcal{M}(S^{-i} \times \times_{j \neq i} \mathcal{M}(S^{-j} \times T^{-j}))$ to $\mathcal{M}(S^{-i} \times \times_{j \neq i} \mathcal{M}(S^{-j}))$ via image measures.) Call this player i's second-order belief. Continuing inductively, we see that a state $(s^1, t^1, \ldots, s^n, t^n)$ describes not just the strategies the players choose, but also each player's entire hierarchy of beliefs about the strategies chosen, about other players' beliefs about this, and so on. This richer description is the starting point of the epistemic approach.[3]

Example 4.1 (A Coordination Game) Consider the coordination game in Fig. 1 (where Ann chooses the row and Bob the column), and the associated type structure in Fig. 2.[4]

There are two types t^a, u^a for Ann, and two types t^b, u^b for Bob. The measure associated with each type is as shown. (For example, Ann's type t^a assigns probability 1/2 to each of Bob's strategy-type pairs (R, t^b) and (R, u^b).) Fix the state (D, t^a, R, t^b) . At this state, Ann plays D and Bob plays R. Ann is 'correct' about Bob's strategy. (Her type t^a assigns probability 1 to Bob's playing R.)

[2] Harsanyi argued that all uncertainty about the structure of the game – whether about payoffs, the strategy sets available to the players, etc. – could be captured via payoff uncertainty. See Hu and Stuart (2001) for a formal treatment of this.

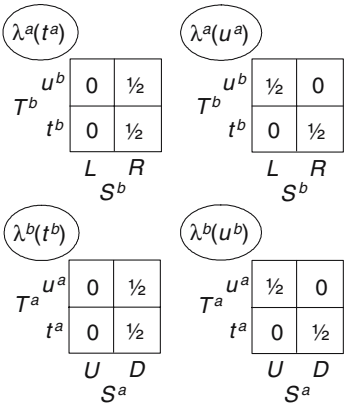
[3] Here we use types to describe uncertainty about the play of the game, not the structure of the game.

[4] Similar to an example in Aumann and Brandenburger (1995, pp. 1166–1167).

Fig. 1

	L	R
U	2, 2	0, 0
D	0, 0	1, 1

Fig. 2



Likewise, Bob is correct about Ann's strategy. Ann, though, thinks it possible Bob is wrong about her strategy. (Her type assigns probability 1/2 to type u^b for Bob, which assigns probability 1/2 to Ann's playing U, not D.) Again, likewise with Bob.

What about the rationality or irrationality of the players? At state (D, t^a, R, t^b), Ann is rational. Her strategy maximizes her expected payoff, given her first-order belief (which assigns probability 1 to R). Likewise, Bob is rational. Ann, though, thinks it possible Bob is irrational. (She assigns probability 1/2 to (R, u^b). With type u^b, Bob gets a higher expected payoff from L than R.) The situation with Bob is again similar.

Summing up, the example is a description of a game situation – a type structure is a descriptive not a predictive tool. Note, too, that the example includes both rationality and irrationality, and also allows for 'incorrect' as well as 'correct' beliefs (e.g., Ann thinks it possible Bob is irrational, though in fact he isn't). These are typical features of the epistemic approach.

A major use of type structures is to identify conditions on the players' rationality, beliefs, etc. that yield various solution concepts. A basic result is on **iteratively undominated** (**IU**) **strategies**. (Delete from the matrix all strongly dominated strategies, then delete all strategies that become strongly dominated in the resulting submatrix, and so on until no further deletion is possible.) Presumably, a rational player i won't play a strongly dominated strategy. Also, if player i assigns probability 1 to the rationality of the other players, then i's marginal on the other players' strategies will assign probability 1 to undominated strategies. So, a player who is rational and believes the other players are rational

won't play a strategy that becomes dominated after the first round of deletions. And so on.

The idea of this argument is very easy. But for all the terms to be formally defined, the type structure apparatus is needed. First, rationality. This is a property of strategy-type pairs. Say (s^i, t^i) is **rational** if s^i maximizes player i's expected payoff under the marginal on S^{-i} of the measure $\lambda^i(t^i)$.

Say type t^i for player i **believes** an event $E \subseteq S^{-i} \times T^{-i}$ if $\lambda^i(t^i)(E) = 1$, and write

$$B^i(E) = \{t^i \in T^i : t^i \text{ believes } E\}.$$

Now, for each player i, let R^i_1 be the set of all rational pairs (s^i, t^i), and for $m > 0$ define R^i_m inductively by

$$R^i_{m+1} = R^i_m \cap [S^i \times B^i(R^{-i}_m)].$$

Definition 4.2 *If* $(s^1, t^1, \ldots, s^n, t^n) \in R_{m+1}$, *say there is* rationality and mth-order belief of rationality *(*RmBR*) at this state. If* $(s^1, t^1, \ldots, s^n, t^n) \in \bigcap_{m=1}^{\infty} R_m$, *say there is* rationality and common belief of rationality *(*RCBR*) at this state.*

With these definitions, one can show: *Fix a type structure and a state* $(s^1, t^1, \ldots, s^n, t^n)$ *at which there is RCBR. Then the strategy profile* $(s^1, \ldots, s^n)$ *is IU. Conversely, fix an IU profile* $(s^1, \ldots, s^n)$. *There is a type structure and a state* $(s^1, t^1, \ldots, s^n, t^n)$ *at which there is RCBR.*

Results like this can be found in the early literature (Brandenburger and Dekel 1987; Tan and Werlang 1988).[5] Again, the idea of it is clear without any formal apparatus. But formalizing epistemic arguments was a crucial step towards solving the harder problems that came later, as we will see.

Other early epistemic results included conditions for correlated equilibrium (Aumann 1987) and Nash equilibrium (Aumann and Brandenburger 1995).

One more comment on type structures: Naturally, we can ask whether Definition 4.1 above is to be taken as primitive or derived. Arguably, hierarchies of beliefs are the primitive, and types are simply a convenient tool for the analyst. Perhaps also, a more primitive way of describing the players' reasoning is via mathematical logic, and a structure such as Definition 4.1 should be derived from such a starting point. See the papers in *Special Issue on Interactive Epistemology* (this journal 1999) and the references there for more on this. Here, we will stay one level above these foundational questions and take type structures as given. (But we will look a bit deeper in Sect. 11.)

[5] Bernheim (1987) and Pearce (1984) introduced the "rationalizable" strategies (which differ from the IU strategies by virtue of an independence requirement), and argued verbally that they will be played under common knowledge of rationality (and independence). See Sect. 13 below on the knowledge vs. belief distinction.

5 Two problems

Two big challenges arise in developing these epistemic tools to a point where we can analyze our starting paradoxes and many other issues in game theory. One is doing epistemics on the tree rather than the matrix. The other is incorporating weak, not just strong, dominance. We start with the tree.

Example 5.1 (A Second Coordination Game) Consider the coordination game in Fig. 3 and the associated type structure in Fig. 4 (where there happens to be one type for each player).

Pick the state (Out, t^a, Out, t^b). Ann plays Out, believing Bob plays Out. Both players get a payoff of 1.

It can be checked that the rational strategy-type pairs are $R_1^a = \{(Out, t^a)\}$ and $R_1^b = \{(Out, t^b), (In, t^b)\}$. Since both types assign probability 1 to rational strategy-type pairs for the other player, we get $R_2^a = R_1^a$ and $R_2^b = R_1^b$, and so $R_m^a = R_1^a$ and $R_m^b = R_1^b$ by induction. In particular then, there is RCBR at the state (Out , t^a, Out, t^b).

This isn't the BI path (on which both players choose In). But as noted earlier, an epistemic model is just a description of a game situation. In the case of a perfect-information (PI) tree, the situation may or may not involve play of the BI path. (Of course, we will be very interested later in looking for conditions under which the BI path is played.)

This said, there is nevertheless a conceptual problem with the scenario. Ann plays Out because she believes Bob plays Out. She believes this because she believes Bob believes she plays Out, and so is indifferent between his choices. (His expected payoff is 1, regardless of his choice.) But Ann knows that if instead she played In, Bob would see this, so she needs to think about how Bob would react. The epistemic model of Sect. 4 doesn't allow us to specify this. In the example, we can calculate Bob's (ex ante) expected payoffs, as above. But we can't calculate his conditional expected payoffs (from In vs. Out), given the event that Ann plays In, since he gives this event probability 0.

Fig. 3

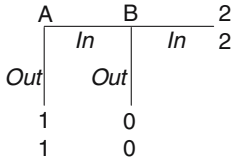


Fig. 4

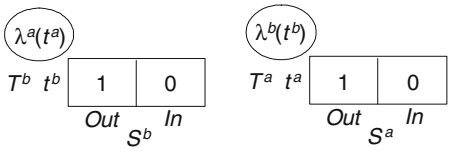


Fig. 5

A \ B	*Out*	*In*
Out	1, 1	1, 1
In	0, 0	2, 2

Now, the second problem we have to solve: incorporating weak as well as strong dominance on the matrix. The results we mentioned in Sect. 4 say that our epistemic set-up yields the IU strategies. Of course, an undominated – even IU – strategy may be inadmissible.

Example 5.2 Figure 5 is the strategic form of the tree in Fig. 3. If we use the same type structure as in Fig. 4, then there is again RCBR at the state (Out, t^a, Out, t^b).

But if we want rationality to mean avoiding inadmissible strategies, then Bob should play *In* not *Out* – even though *Out* isn't strongly dominated. Should Ann then assign probability 0 to Bob's playing *Out*? This leads to the conceptual problem on which we quoted Mas-Colell et al. (1995, p. 240): Wouldn't this conflict with the idea of admissibility, which says that a player considers as possible (even if unlikely) any of the strategies for the other players.

Yet, admissibility seems a very reasonable, even basic, requirement. See Kohlberg and Mertens (1986, Sect. 2.7) for a thorough discussion and defense. Kolhberg and Mertens also point out the connection between admissibility and invariance (Kohlberg and Mertens, 1986, Sect. 2.4) – which we will consider in Sect. 12.

In the next section, we will look at how to modify probability theory to solve both this problem in the matrix and the problem in the tree.

6 Extended probabilities I

Both problems we identified involve the treatment of probability-0 events. Two extensions of ordinary probability theory have been used in the epistemic program, to tackle these problems.

On the tree, an appropriate tool is **conditional probability systems** (**CPS's**), due to Rényi (1955). A CPS specifies a family of conditioning events E and a measure p_E for each such event, together with certain restrictions on these measures. The interpretation is that p_E is what the player believes, after observing E. The key is that even if $p_\Omega(E) = 0$ (where Ω is the entire space), the measure p_E is still specified. That is, even if E is 'unexpected,' the player has a measure if E nevertheless happens. Myerson (1991, Chap. 1) provided a preference-based axiomatization of a class of CPS's. Battigalli and Siniscalchi (1999; 2002) further developed both the pure theory and the game-theoretic application of CPS's, as we will discuss in detail later.

On the matrix, an appropriate tool is **lexicographic probability systems** (**LPS's**), introduced and axiomatized by Blume et al. (1991a;b). An LPS specifies a sequence of probability measures. The interpretation is that the states that get positive probability under the first measure make up the player's primary hypothesis about the true state. But the player recognizes that his primary hypothesis might be mistaken, and so also forms a secondary hypothesis, consisting of the states that get positive probability under the second measure. Then his tertiary hypothesis, and so on. The primary states can be thought of as infinitely more likely than the secondary states, which are infinitely more likely than the tertiary states, etc. Stahl (1995), Stalnaker (1998), Asheim (2001), and Brandenburger et al. (2006), among other papers, use LPS's.

Example 6.1 Let's go back to the game of Fig. 3, to see how CPS's work. Figure 6 is another type structure for this game, different from the one in Fig. 4. The difference is that here the probabilities come from CPS's, as we will explain.

Start with Ann and the first node in the tree. Formally, this is the event $\{In, Out\} \times \{t^b\}$ – i.e., the event that Bob chooses either of his strategies. Ann assigns probability 1 to *Out*, given this event (essentially as before). Next, Bob. At the root of the tree, he assigns probability 1 to Ann's playing *Out* (as before). But now we also have to specify what Bob believes at the second node in the tree. Formally, this is the event $\{(In, t^a)\}$ – i.e., the event that Ann chooses *In*. One of the conditions of a CPS is that $p_E(E) = 1$. (Conceptually, this says that players believe what they observe.) So at the second node, Bob must assign probability 1 to $\{(In, t^a)\}$. This is the measure shown in square brackets.

Example 6.2 Figure 7 is a type structure for the game of Fig. 5, that now specifies LPS's.

Each player has a primary hypothesis that assigns probability 1 to the other player's choosing *Out*. But each player also has a secondary hypothesis that assigns probability 1 to the other player's choosing *In*. These measures are shown in parentheses.

We see how LPS's can solve the conceptual problem with admissibility: All states (i.e., strategy-type pairs) are ruled in, in the sense that every state gets

Fig. 6

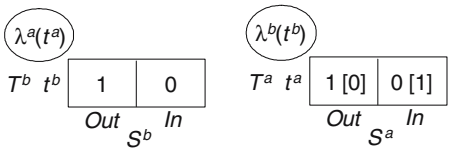
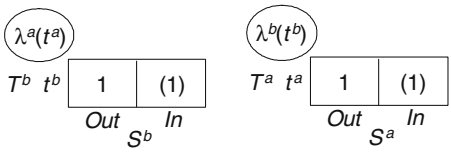


Fig. 7



positive probability under some measure. But states can also be ruled out, in the sense that they can be give infinitely less weight than other states. (We will see later though, in Sect. 9, that there is a further challenge to overcome before LPS's yield an analysis of IA.)

For the general definitions of CPS's and LPS's, and CPS-based and LPS-based type structures (extending Definition 4.1), see the papers referenced above.[6]

7 Extended probabilities II

Let us now check that extended probabilities really do work the way we want. We saw in the type structure of Fig. 4, which used ordinary probabilities, that RCBR holds at the state (Out, t^a, Out, t^b). What epistemic conditions hold at the state (Out, t^a, Out, t^b) in Fig. 6?

To say which strategy-type pairs are rational, we need a definition of rationality with CPS's. Here is the natural definition: Fix a strategy-type pair (s^i, t^i), where t^i is associated with a CPS. Call this pair **rational** (**in the tree**) if the following holds: Fix any information set H for i allowed by s^i, and look at the measure given H (i.e., given the event that the other players' strategies allow H). Require that s^i maximizes i's expected payoff under this measure, among all strategies r^i for i that allow H .

So (Out, t^a) is rational for Ann. At her node, Ann assigns probability 1 to Bob's playing *Out*, making *Out* optimal for her. But (Out, t^b) is irrational for Bob. At his node, he assigns probability 1 to Ann's playing *In* (as he must), and so he gets an expected payoff of 2 from *In*, as opposed to 0 from *Out*. The irrationality of (Out, t^b) is what we want intuitively.

Next, what does Ann think about Bob's rationality? To answer, we need a CPS-analog to belief (as defined in Sect. 4). Ben Porath (1997) proposed the following:[7] Say player i **initially believes** event E if E gets probability 1 given the root of the tree, under i's CPS. (Formally, the conditioning event includes all strategy profiles of the other players.) This implies that E gets probability 1 at any information set H that gets positive probability under the measure given the root. Battigalli and Siniscalchi (2002) strengthened this definition to: Say player i **strongly believes** event E if for every information set H with $E \cap (H \times T^{-i}) \neq \emptyset$, the measure given H assigns probability 1 to E. That is, player i believes E, whenever E is possible given what i observes.

It is immediate in Fig. 6 that Ann's type t^a doesn't initially believe (so certainly doesn't strongly believe) that Bob is rational. This is different from the situation in Example 5.1, where Ann believes Bob is rational – and there is even RCBR. Again, the new answer is the intuitively correct one.

Now, the analysis of Example 6.2. What epistemic conditions hold at the state (Out, t^a, Out, t^b) in Fig. 7?

[6] In particular, definitions on infinite spaces turn out to be crucial; see Sect. 9 below.

[7] We have taken the liberty of changing terminology, for consistency with "strong belief" below.

To answer, we need LPS-analogs to rationality and belief. For rationality, fix strategy-type pairs (s^i, t^i) and (r^i, t^i) for player i, where t^i is now associated with an LPS. Calculate the tuple of expected payoffs to i from s^i, using first the primary measure associated with t^i, then the secondary measure associated with t^i, etc. Calculate the corresponding tuple for r^i. If the first tuple lexicographically exceeds the second, then s^i is preferred to r^i.[8] A strategy-type pair (s^i, t^i) is **rational** (**in the lexicographic sense**) if s^i is maximal under this ranking.

So, as before, (Out, t^a) is rational for Ann. For Bob, both *Out* and *In* give an expected payoff of 1 under his primary measure. But *In* gives him an expected payoff of 2 under his secondary measure, as opposed to an expected payoff of 0 from *Out*. We want (Out, t^b) to be irrational for Bob, since *Out* is inadmissible.

What does each player think about the other's rationality? For this, we need an LPS-analog to belief. An early candidate in the literature was: Say player i **believes** event E **at the** 1**st level** if E gets primary probability 1 under i's LPS (Börgers 1994; Brandenburger 1992).

A stronger concept (but still weaker than belief) is: Say i **assumes** E if all states not in E are infinitely less likely than all states in E, under i's LPS. (See Brandenburger et al. (2006) for the general definition, which covers infinite spaces.) In other words, a player who assumes E recognizes E may not happen, but is prepared to 'count on' E versus not-E.

Clearly, in Fig. 7, Ann's type t^a doesn't 1st-level believe (so certainly doesn't assume) that Bob is rational. (In fact, type t^a assumes Bob is irrational.) Again, this is what we want intuitively.

8 Resolving the paradoxes I

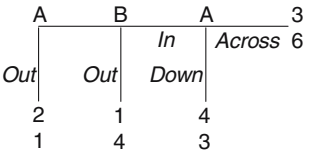
Let us use our language for the tree – involving type structures and CPS's – to go back to the paradox of BI. The problem was whether it is possible to find epistemic conditions that yield BI in a formal and unambiguous manner.

For the simple coordination tree of Example 5.1, the obvious condition of RCBR does not necessarily yield BI, as we saw. But this was because of a deficiency of the language. Does the language with CPS's work better? The answer is yes. Regardless of the type structure, Bob must play *In* if he is rational, because the definition of a CPS requires him, at his information set, to assign probability 1 to Ann's playing *In*. If Ann initially (or strongly) believes Bob is rational, and is rational, she too will play *In*. The BI path results.

This is very straightforward. But will we get a similar answer in more complicated trees? First some definitions. Paralleling Definition 4.2, with CPS's we can define inductively **rationality and** m**th-order initial belief of rationality** (**R**m**IBR**) at a state of a type structure, and **rationality and common initial belief of rationality** (**RCIBR**). (See Ben Porath 1997.) Similarly, we can define **rationality and** m**th-order strong belief of rationality** (**R**m**SBR**), and **rationality and**

[8] If $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, then x lexicographically exceeds y if $y_j > x_j$ implies $x_k > y_k$ for some $k < j$.

Fig. 8



common strong belief of rationality (**RCSBR**). (See Battigalli and Siniscalchi 2002.) The question is then: Does the condition of RCIBR, or perhaps RCSBR, yield BI in a perfect-information (PI) tree?

Example 8.1 (Three-Legged Centipede) Figure 8 is three-legged Centipede (where the top payoffs are Ann's, and the bottom payoffs are Bob's), and Fig. 9 is an associated CPS-based type structure.

Type t^a for Ann has the measure shown in the top-left matrix. This is her measure at the first node in the tree. Since this measure assigns positive probability (in fact, probability 1) to her second node (i.e., to the event that Bob chooses *In*), it determines her measure there. By contrast, type u^a for Ann assigns probability 0 to her second node. The measure there is shown in square brackets (and assigns probability 1 to $\{(In, t^b)\}$). Both of Bob's types initially assign probability 1 to Ann's playing *Out*. At his node, Bob's type t^b assigns probability 1 to $\{(Across, t^a)\}$, while his type u^b assigns probability 1 to $\{(Down, t^a)\}$.

Let us list the rational strategy-type pairs in this example. They are (*Down*, t^a), (*Out*, u^a), (*In*, t^b) , and (*Out*, u^b). We see that both of Ann's types t^a and u^a initially (even strongly) believe Bob is rational. Also, both of Bob's types initially believe that Ann is rational. (But note that t^b doesn't strongly believe that Ann is rational. We come back to this.) Given this, a simple induction shows that at the state (*Down*, t^a, *In*, t^b) for instance, RCIBR holds.

This kind of example is the focus of Ben Porath (1997), a key step forward in the epistemic program. Let's interpret it. Ann plays across at her first node, believing (initially) that Bob will play *In*, so she can get a payoff of 4. Why would Bob play *In*? Because he initially believes that Ann plays *Out*. But in the probability-0 event that Ann plays across at her first node, Bob then assigns

Fig. 9

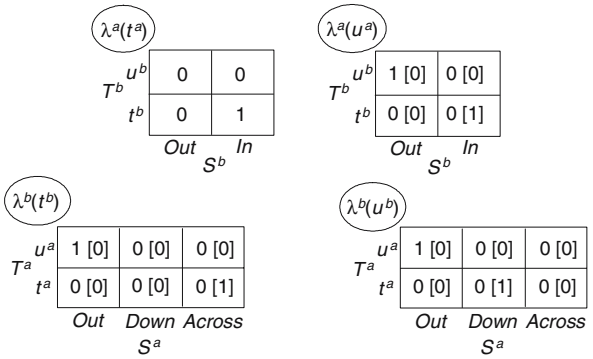
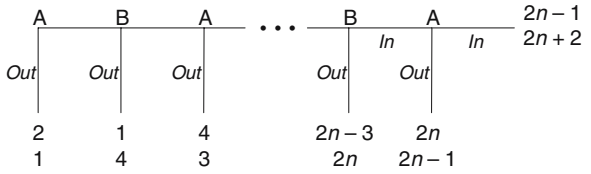


Fig. 10



probability 1 to Ann's playing across at her second node – i.e., to Ann's being irrational. He therefore (rationally) plays *In*. In more everyday language, by playing across at her first node, Ann 'bluffs' Bob into believing she is irrational and will play across at her second node.[9]

Interestingly, this is exactly the line of reasoning from Mas-Colell et al. (1995, p. 282) we quoted earlier. So, in fact, there is no contradiction or impossibility with this reasoning – we have just given a formal set-up in which it holds. The resolution of the BI paradox is, rather, to accept that even under the condition of RCIBR – which a priori might be expected to yield BI – the BI path may not result.

But one can also argue that RCIBR is not the right condition: it is too weak. In the above example, Bob realizes that he might be 'surprised' in the play of the game – that is why he has a CPS, not just an ordinary probability measure. If he realizes he might be surprised, should he abandon his (initial) belief that Ann is rational when he is surprised? Bob's type t^b does so. This brings us back to strong belief (Battigalli and Siniscalchi 2002). The argument says that we want t^b to strongly believe, not just initially believe, that Ann is rational. Type t^b will strongly believe Ann is rational if we move the probability-1 weight (in square brackets) on $(Across, t^a)$ to $(Down, t^a)$. But now (In, t^b) isn't rational for Bob, so Ann doesn't (even initially) believe Bob is rational. It looks as if the example unravels.

So, replacing initial belief with strong belief, the question is: Does RCSBR yield the BI path in Centipede? The answer is yes: *Fix a CPS-based type structure for n-legged Centipede (Fig. 10), and a state at which there is RCSBR. Then Ann plays Out.*

The result follows from Friedenberg (2002). Here is a verbal argument. Suppose to the contrary that there is an RCSBR state at which Ann plays across at the first node. Consider the length of play at each such state (before Ann or Bob plays *Out*), and pick a state (s^a, t^a, s^b, t^b) with the longest play. Suppose it is Bob who ends the game, by playing *Out* at node H. (If it is Ann, a similar argument works.) Then the event "Bob is rational, Bob is rational and strongly believes Ann is rational, ..." (denote this E) and the event "Ann's node $H-1$ is reached" (denote this F) have a nonempty intersection. But Ann's type t^a strongly believes E. (This uses the assumption that t^a strongly believes each of the events "Bob is rational," "Bob is rational and strongly believes Ann

[9] At the state $(Down, t^a, In, t^b)$, the bluff works. But at the state $(Down, t^a, Out, u^b)$, Ann attempts the bluff and it fails.

is rational," ..., and a conjunction property of strong belief.) It follows that at $H-1$, type t^a assigns probability 1 to E. At $H-1$, type t^a also assigns probability 1 to F (by one of the defining properties of a CPS). Therefore at $H-1$, type t^a assigns probability 1 to $E \cap F$. By construction, at $H-1$, type t^a must then assign probability 1 to the event that Bob plays *Out* exactly at H. But then if r^a is the strategy for Ann that plays across until $H-1$ and *Out* at $H-1$, this strategy yields Ann a higher expected payoff at $H-1$ under t^a, than does s^a (which plays across at $H-1$). This contradicts the rationality of (s^a, t^a).[10]

This result gives a second resolution of the BI paradox – at least as far as Centipede is concerned. As above, no contradiction or impossibility in reasoning about the game is found. Moreover, we have found a very natural line of reasoning that actually yields BI, unlike earlier.

Let us reemphasize that an epistemic analysis is not a prediction independent of the specific assumptions made. In Centipede, if RCSBR holds, the BI path results. But RCSBR need not hold. In fact, it seems a stringent assumption[11] and quite plausibly might not hold. For example, we might want to assume only that both players are rational and strongly believe the other is rational – much less than RCSBR.[12] Without RCSBR, the BI path won't necessarily obtain. (We already saw this in Example 8.1 above.)

Later, we will look at whether what we have found for Centipede generalizes to other PI games. First, we want to go back to the matrix and LPS's.

So, again following Definition 4.2, with LPS's we can define inductively **rationality and *m*th-order 1st-level belief of rationality** (**R*m*1BR**) at a state of a type structure, and **rationality and common 1st-level belief of rationality** (**RC1BR**). Likewise, we can define **rationality and *m*th-order assumption of rationality** (**R*m*AR**), and **rationality and common assumption of rationality** (**RCAR**). What do these conditions yield?

Fig. 11

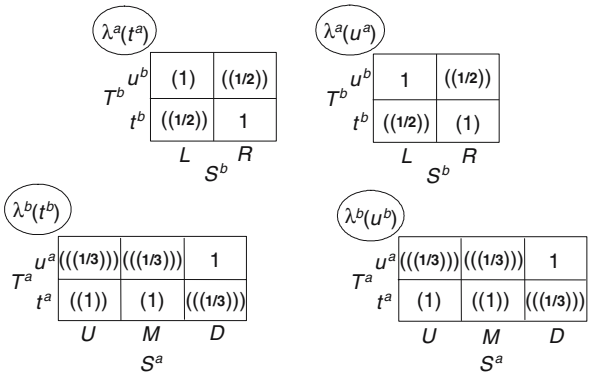
		B	
		L	*R*
A	*U*	0, 1	2, 0
	M	0, 0	0, 1
	D	1, 1	1, 1

[10] Aumann (1998) provides knowledge-based epistemic conditions under which Ann plays *Out* in Centipede (proved via a forward-looking argument). Knowledge-based models are different from the belief-based models we are looking at here; see Sect. 13 below.

[11] 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'."

[12] Note that we are talking only about the plausibility of departures from an epistemic 'baseline.' Sorin (1998) gives a method for quantifying the size of such departures.

Fig. 12



Example 8.2 Figure 12 is an LPS-based type structure for the game of Fig. 11. (The secondary measures are in single parentheses, the third-level measures in double parentheses, the fourth-level measures in triple parentheses.)

The rational strategy-type pairs are (U, t^a), (D, u^a), (R, t^b), and (L, u^b). Also, each type assigns primary probability 1 to rational strategy-type pairs for the other player, so each type believes at the 1st level that the other player is rational. By induction, RC1BR holds at the state (U, t^a, R, t^b), for example.

But notice that while type t^b for Bob believes at the 1st level that Ann is rational, this type doesn't assume Ann is rational. This is because t^b considers the irrational strategy-type pair (M, t^a) for Ann infinitely more likely than the rational pair (U, t^a). Arguably, if Bob is really 'trying to think' that Ann is rational, he should put the rational pair (U, t^a) first. If he does, then he will rationally play L not R. Ann, presumably, will play D. The (unique) IA profile (D, L) results.

If we replace belief at the 1st level with assumption – i.e., consider RCAR in place of RC1BR – in the game of Fig. 11, then the IA outcome $(1, 1)$ will always result. Here is a proof for the finite case. Fix an arbitrary finite LPS-based type structure, and a state at which there is RCAR. Let Ann's type at this state be t^a. Certainly, Ann cannot play M at this state, since this is (even strongly) dominated. Can Ann play U? If so, since she is rational, the primary measure associated with t^a must put positive weight on R. That is, there must be a type v^b for Bob such that (R, v^b) gets positive primary probability. Since Ann's type t^a assumes Bob is rational, (R, v^b) must be rational. Ann's strategy-type pair (U, t^a) gets positive probability under some measure in the LPS associated with v^b. So, for (R, v^b) to be rational, there must be a type v^a for Ann such that (M, v^a) gets positive probability under this same measure, or an earlier measure, in the LPS associated with v^b. But then, v^b does not assume Ann is rational, since it doesn't make the rational strategy-type pair (U, t^a) infinitely more likely than the irrational pair (M, v^a) (recall that M is dominated). This says rationality and 2nd-order assumption of rationality fails – so certainly RCAR does. The conclusion is that under RCAR, Ann will play D, as claimed.

Fig. 13

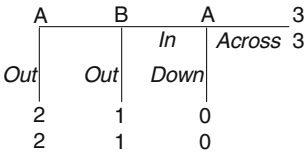
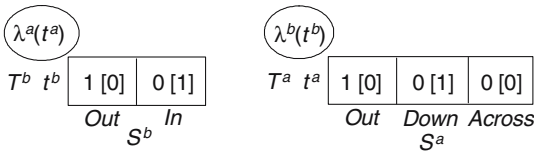


Fig. 14



(The claim is also true for a general type structure. This follows from Theorem 6.1(i) in Brandenburger et al. 2006.)

Does the same conclusion hold in all games? We will see in the next section.

9 Resolving the paradoxes II

Back to the tree, and the condition of RCSBR. It turns out that the result that RCSBR yields BI in Centipede was, indeed, special. In general, RCSBR need not yield the BI outcome in a PI game. (We will say later what makes Centipede special.)

Example 9.1 (A Third Coordination Game) Consider the coordination game in Fig. 13 and the associated CPS-based type structure in Fig. 14.

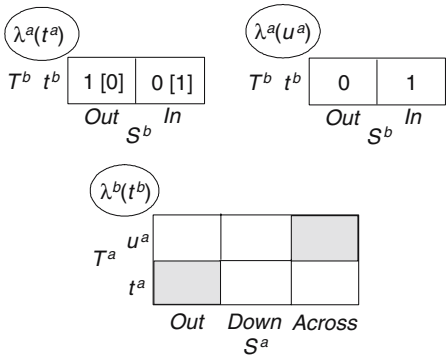
The rational strategy-type pairs are (Out, t^a) and (Out, t^b) for Ann and Bob respectively. Ann's type t^a strongly believes $\{(Out, t^b)\}$, and Bob's type t^b strongly believes $\{(Out, t^a)\}$. By induction, RCSBR holds at the state (Out, t^a, Out, t^b).

Here is a game of pure coordination (so that the BI outcome is even Pareto dominant), but the BI outcome need not arise under RCSBR. The key is to see that both $(Down, t^a)$ and $(Across, t^a)$ are irrational for Ann, since she (strongly) believes Bob plays *Out* . So, at his node, Bob can't believe Ann is rational. If he considers it sufficiently more likely Ann will play *Down* rather than *Across*, he will rationally play *Out* (as happens). In short, if Ann doesn't play *Out*, she is irrational and so 'all bets are off' as to what she will do. She could play *Down*.

The situation described in Example 9.1 may be surprising, at least at first blush, but there does not appear to be anything conceptually wrong with it. Indeed, it points to an interesting way in which the players in a game can literally be trapped by their beliefs – which here prevent them from getting their mutually preferred (3, 3) outcome.[13]

[13] The (3, 3) outcome does seem very salient in Fig. 13. But the game itself is only a partial description of the strategic situation. A full description includes a type structure – our claim is that if the type structure is as in Fig. 14, the saliency of (3, 3) disappears.

Fig. 15



This said, we still want to identify epistemic conditions which yield the BI outcome in any PI game. We will look at several routes.

Here is the first. Consider the following line of reasoning in Example 9.1 (it gets formalized below): If Ann forgoes the payoff of 2 she can get by playing *Out* at the first node, then surely she must be playing *Across* to get 3. Playing *Down* to get 0 makes little sense since this is lower than the payoff she gave up at the first node. But if Bob considers *Across* (sufficiently) more likely than *Down*, he will play *In*. Presumably then, Ann will indeed play *Across*, and the BI path results.

There is no contradiction with the previous analysis because in Fig. 14, Ann is irrational once she doesn't play *Out*, so we can't say Ann should then rationally play *Across* not *Down*. To make *Across* rational for Ann, we have to add more types to the structure. This key insight is due to Stalnaker (1998) and Battigalli and Siniscalchi (2002). To see how it works, add a second type u^a for Ann that is the 'reverse' of type t^a, as in Fig. 15. The rational strategy-type pairs for Ann are now (*Out*, t^a) and (*Across*, u^a), as shaded. If Bob strongly believes Ann is rational, then at his node he must assign probability 1 to Ann's playing *Across*. He will rationally play *In*. This means type t^a for Ann doesn't (strongly) believe Bob is rational. The non-BI scenario unravels.

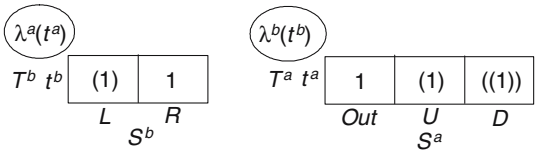
The solution concept that this line of argument yields is **extensive-form rationalizability** (**EFR**), due to Pearce (1984). Battigalli (1997) showed that EFR yields the BI outcome in a PI game (under an assumption ruling out certain payoff ties). So, we will get epistemic conditions for BI, as desired. But note that the reasoning above is also forward-induction (FI) reasoning à la Kohlberg and Mertens (1986, Sect. 2.3). EFR captures FI, too. (Interestingly, while Kohlberg and Mertens introduced FI in the context of non-PI games, we now see that it already arises in PI games – such as Fig. 13.)

To finish the epistemic analysis: Battigalli and Siniscalchi consider a **complete** CPS-based type structure, which contains, in a certain sense, every possible type for each player. Go back to Definition 4.1. A type structure as defined there is complete if each map λ^i is surjective – i.e., for each player i and every

Fig. 16

		B: *L*	B: *R*
	Out	2, 2	2, 2
A	*U*	0, 0	1, 3
	D	3, 1	0, 0

Fig. 17



(Borel) measure on $S^{-i} \times T^{-i}$, there is a type of player i with that measure.[14] A complete CPS-based type structure is defined analogously. Battigalli and Siniscalchi prove: *Fix a complete CPS-based type structure. If there is RCSBR at the state* $(s^1, t^1, \ldots, s^n, t^n)$*, then the strategy profile* $(s^1, \ldots, s^n)$ *is extensive-form rationalizable. Conversely, if the profile* $(s^1, \ldots, s^n)$ *is extensive-form rationalizable, then there is a state* $(s^1, t^1, \ldots, s^n, t^n)$ *at which there is RCSBR.*

Next, back to the matrix again. The answer to the question at the end of the previous section is that, in fact, RCAR need not yield an IA outcome.

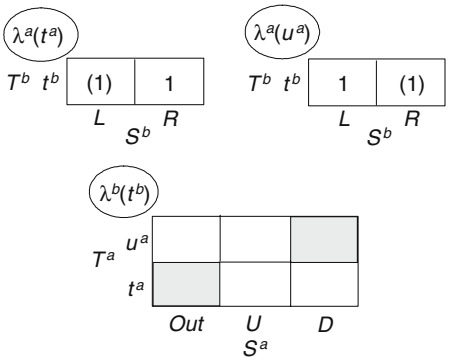
Example 9.2 (Battle of the Sexes With an Outside Option) Figure 16 is the strategic form of Battle of the Sexes With an Outside Option (Kohlberg and Mertens (1986, Sect. 2.3); Osborne and Rubinstein (1994, Ex.110.1)) and Fig. 17 is an associated LPS-based type structure.

The rational strategy-type pairs are (*Out*, t^a) and (R, t^b). (Both L and R give Bob a primary expected payoff of 2, but R gives Bob a secondary expected payoff of 3, versus 0 from L.) By induction, RCAR holds at (*Out*, t^a, R, t^b).

Yet the IA outcome is (3, 1). (Osborne and Rubinstein observe that IA in this game gives the Kohlberg and Mertens FI outcome.) As in Example 9.1, there is no conceptual problem with this non-IA scenario. Still, how can we get IA? What is needed is for Bob to consider D infinitely more likely than U, rather than vice versa. Then he will rationally play L, not R, and Ann presumably will play D. One reason to give U infinitely less weight than D is that it is eliminated on the first round of IA. Of course, we don't want just to assume the answer and require the weights to work this way. The key again is completeness. Notice that while choosing U can never be rational for Ann (for any type), choosing D can be rational. In Fig. 18, we have added a second type for Ann, which indeed makes D rational for her, and shaded her rational strategy-type pairs.

[14] Completeness is defined in Brandenburger (2003). A complete type structure will be uncountably infinite.

Fig. 18



If Bob assumes Ann is rational, he must consider the shaded states infinitely more likely than the unshaded ones – so he will play L, as desired.

For the general case, we need a definition of a complete LPS-based type structure. See Brandenburger et al. (2006) for a formal treatment that shows the following: *Fix a complete LPS-based type structure. If there is RmAR at the state* $(s^1, t^1, \ldots, s^n, t^n)$*, then the strategy profile* $(s^1, \ldots, s^n)$ *survives* $(m+1)$ *rounds of iterated admissibility. Conversely, if the profile* $(s^1, \ldots, s^n)$ *survives* $(m+1)$ *rounds of iterated admissibility, then there is a state* $(s^1, t^1, \ldots, s^n, t^n)$ *at which there is RmAR.*

Some observations: First, the result is stated for RmAR and not RCAR. See Sects. 10 – 11 below for the reason. Of course, for a given game, there is an m such that IA stabilizes after m rounds.

Remember that the 'philosophy' underneath admissibility is that a player should not rule out any strategies for the other players. The result says that underneath IA is the idea that a player should consider not only all strategies, but also all types, for the other players.

Next, IA yields the BI outcome in a PI game (ruling out certain payoff ties). See Marx and Swinkels (1997). So understanding IA gives a second epistemic analysis of BI, in addition to the EFR-based analysis above. (We say more in Sect. 12 below about strategic- vs. extensive-form analysis.)

Here is a third route to getting BI in PI games, different from the completeness route. Asheim (2001) develops an epistemic analysis using the properness concept (Myerson 1978). Go back to Example 9.1. The properness idea says that Bob's type t^b should view $(Across, t^a)$ as infinitely more likely than $(Down, t^a)$ since *Across* is the less costly 'mistake' for Ann, given her type t^a. Unlike the completeness route taken above, the irrationality of both *Down* and *Across* (given Ann's type t^a) is accepted. But the relative ranking of these 'mistakes' must be in the right order. With this ranking, Bob is irrational to play *Out* rather than *In*. Ann presumably will play *Across*, and we get BI again. Asheim (2001) formulates a general such result.

Finally, we point out once more that BI – and IA – aren't inevitable predictions of an epistemic analysis. These predictions rest on very specific conditions on the game, such as RCSBR, RmAR, and completeness.

10 Solution concepts

In analyzing our game-theoretic paradoxes, we have considered a number of epistemic conditions. Here, we organize these conditions and see that a unified picture emerges. Basically, the conditions are all characterized by various forms of iterated dominance. Some of these dominance concepts have only been understood – or even defined – via the epistemic program. So these discoveries and the overall scheme that emerges are another benefit of the program.

The first epistemic condition was RCBR (rationality and common belief of rationality). Then we looked at various 'refinements' of RCBR, as summarized in Table 1. Table 2 shows what is known about the characterization of these conditions. (Here $\approx$ means "is characterized by".)

Some comments on the table:

a. In the first row, IU is the set of iteratively undominated strategies (Sect. 4).

b. In the second row, $S^\infty W$ is the set of strategies that remain after one round of deletion of inadmissible (weakly dominated) strategies followed by iterated deletion of strongly dominated strategies (Dekel and Fudenberg 1990). Also in the second row, $S^\infty CD$ is the set of strategies that remain after one round of deletion of conditionally dominated strategies (Shimoji and Watson 1998) followed by iterated deletion of strongly dominated strategies. As a special case here, Ben Porath (1997) showed that in PI games satisfying a no-ties condition, RCIBR is characterized by $S^\infty W$. Example 8.1 is an instance of this. The $S^\infty W$ set in three-legged Centipede is $\{Out, Down\} \times \{In, Out\}$, so certainly includes the profile $(Down, In)$ we saw was possible under RCIBR.

c. In the third row, m-IA (resp. m-EFR) is the set of strategies that remain after m rounds of iterated admissibility (resp. m rounds of extensive-form rationalizability). We also record that EFR is equivalent round-for-round to iterated conditional dominance (ICD) – which is essentially "iterated strong dominance on the tree." (See Shimoji and Watson 1998.) Similar

Table 1

Matrix	Tree
RCBR (Rationality and common belief of rationality)	
RC1BR (Rationality and common 1st-level belief of rationality) **RCAR** (Rationality and common assumption of rationality) **LPS-Completeness**	**RCIBR** (Rationality and common initial belief of rationality) **RCSBR** (Rationality and common strong belief of rationality) **CPS-Completeness**

Table 2

Matrix	Tree
RCBR ≈ IU	
RC1BR ≈ S$^{\infty}$W (See Brandenburger 1992, Börgers 1994)	**RCIBR ≈ S$^{\infty}$CD** (See Ben Porath 1997, Battigalli and Siniscalchi 1999)
RmAR & Completeness ≈ $(m+1)$-IA (See Brandenburger et al. 2006) PI Games: m-IA yields BI outcome (for sufficiently large m) (See Marx and Swinkels 1997)	**RmSBR & Completeness ≈ $(m+1)$-EFR (= $(m+1)$-ICD)** (See Battigalli and Siniscalchi 2002) PI Games: m-EFR yields BI outcome (for sufficiently large m) (See Battigalli 1997)
RCAR & Completeness is impossible (See Brandenburger et al. 2006)	**RCSBR & Completeness ≈ EFR (= ICD)** (See Battigalli and Siniscalchi 2002) PI Games: EFR yields BI outcome
RCAR ≈ SAS (See Brandenburger et al. 2006) PI Games: RCAR yields a Nash-equilibrium outcome (See Brandenburger and Friedenberg 2004)	**RCSBR ≈ ?** PI Games: RCSBR yields a Nash-equilibrium outcome (See Friedenberg 2002)

to the second row, we get equality between the left-hand and right-hand solution concepts in the third row for the case of a generic tree. (See, e.g., Shimoji 2004.)

d. Note the impossibility result in the fourth row. RCAR is impossible in a complete (LPS-based) type structure. We come back to this in the next section.

e. In the fifth row, SAS stands for "self-admissible set," defined in Brandenburger et al. (2006). SAS's may be viewed as the weak-dominance analog to Pearce (1984) best-response sets (BRS's). But while the BRS's of a game are all contained in the IU set, the SAS's need not be contained in the IA set. We saw this in Battle of the Sexes With an Outside Option (Example 9.2). The profile (*Out*, *R*) was playable under RCAR, but even disjoint from the IA set. (In Example 8.2, we argued that an SAS outcome was an IA outcome – but this was special.)

f. Also in the fifth row, note that the characterization of RCSBR in general trees is open.

g. The table notes what various solution concepts yield in the special case of PI games. These statements are proved under various payoff restrictions. See the references for details.

h. Note, in particular, the result that in a PI game, RCSBR yields a Nash-equilibrium outcome. This is the 'real' reason why RCSBR gives BI in Centipede (Sect. 8). In Centipede, there is a unique Nash path and it coincides with the BI path. Of course, this isn't true in general – e.g., the Coordination Game in Example 9.1.

Much of the epistemic program in game theory can be thought of as studying 'refinements' of the basic RCBR condition on a game. (Table 2 shows what is known about some of these refinements – but it is certainly not exhaustive.) To some extent, the program can be seen as a response to the equilibrium refinements program of the 1980s. In that program, the starting point was Nash equilibrium. Various modifications of equilibrium were proposed, and attempts made to interpret these as reflecting one or another underlying notion of rationality (plus belief in rationality, etc.). For Mertens, a leading proponent, this direction of analysis was a conscious choice: "In this way, we may eventually reach an axiomatisation, and an interpretation in terms of rationality, without imposing any explicit preconception about what rationality exactly means, except for some general a priori requirement[s]" (Mertens 1989, p. 583). The epistemic program is different in two ways. It starts with explicit definitions of rationality, belief, etc., refines these conditions, and tries to work out implications for the play of a game. Also, Nash equilibrium is no longer the starting point, but a particular case (as noted in Sect. 4).

11 Paradox regained?

Naturally, the epistemic program has uncovered new foundational questions in game theory. The existence of structures containing all possible types of the players (Sect. 9) is one such issue. After all, such a structure sounds rather like the "sets of everything" that are well known to cause difficulties in mathematics (Russell's Paradox, etc.). Can such a structure actually exist?[15]

In the literature, various kinds of 'large' type structures have been considered. When types are associated with ordinary probabilities, existence results were obtained by Armbruster and Böge (1979), Böge and Eisele (1979), Mertens and Zamir (1985), and others. Battigalli and Siniscalchi (1999) and Brandenburger et al. (2006) give existence results for the cases of CPS's and LPS's respectively. But non-existence is also possible; see, e.g., Fagin et al. (1999), Heifetz and Samet (1999), Meier (2005), and Brandenburger and Keisler (2006).[16]

A way to understand the boundary between the existence and non-existence results is via mathematical logic. The epistemic approach to game theory says that players have beliefs about the game – about other players' strategies, beliefs, rationality, etc. Now, we add a specific language – i.e., logic – in which these

[15] Are complete type structures really needed for the theorems quoted in Sect. 9? (I am grateful to the referees for raising this issue.) Examples 9.1 and 9.2 show that an arbitrary type structure won't suffice, but also suggest that perhaps we can add just the 'right' types to the given structure, without adding all possible types. The difficulty with this approach is that what the right types are will depend on the game in question. (If tailoring epistemic conditions to a particular game, we could even require directly that the strategy profile we're interested in is played.) Completeness seems an appropriate game-independent condition.

[16] There are also existence results (e.g., Aumann 1999) and non-existence results (e.g., Fagin 1994; Heifetz and Samet (1998); Heifetz (1999), for knowledge structures (Sect. 13 below).

beliefs are formed. See the papers in *Special Issue on Interactive Epistemology* (this journal 1999) and the references there.

Analyzed this way, the boundary between existence and non-existence results turns on the expressibility of the language considered. In particular, the papers cited above (and others) that give existence results make various topological and measure-theoretic assumptions that, from a logical perspective, effectively restrict the language the players can use to form beliefs. With the restrictions, spaces of all beliefs become possible.

A largely open area is to find logics that allow us to carry out epistemic analyses like the ones discussed in the earlier sections. Such logics must be able to express concepts such as rationality, strong belief, assumption, etc., allow the existence of complete structures, and yield conclusions about solution concepts, as earlier. An analysis of this type would have the benefit of being much more explicit about the players' reasoning processes in games. Ewerhart (2002) and Board (2002; 2004) take steps in this direction – on the matrix and tree, respectively.

Let's go back to the pre-logical – i.e., topological and measure-theoretic – approach to getting complete type structures. Table 2 indicates that, even then, there is a 'limit to analysis.' For a given game matrix and any number m, RmAR is possible under completeness, but RCAR under completeness is impossible. True, neither condition is in any way necessary for a satisfactory analysis of a game. (In Sect. 9, we emphasized in particular that incomplete structures are meaningful and interesting.) But both conditions do seem basic to a 'fully rational' analysis of games, and the fact that both can't hold is definitely disturbing.

The analogous conditions on the tree – RCSBR and (CPS-based vs. LPS-based) completeness – are consistent. (Refer again to Table 2 and to Sect. 9.)

Why the difference? The basic reason appears to be that the strategic-form analysis is more demanding. In particular, it satisfies an invariance requirement, discussed next.

12 Strategic versus Extensive analysis [17]

Kohlberg and Mertens (1986, Sect. 2.4) argued that a 'fully rational' analysis of games should be invariant – i.e., should depend only on the fully reduced strategic form.[18] (See also Mertens (1989, p. 582) for further discussion.) In this they appealed to early results in game theory (Dalkey 1953; Thompson 1952) which established that two trees sharing the same reduced strategic form differ from each other by a (finite) sequence of elementary transformations of the tree, each of which can be argued to be 'strategically inessential.' Kohlberg and

[17] This section draws on material in "Common Assumption of Rationality in Games" by Brandenburger and Friedenberg (2002) This paper is superseded by Brandenburger et al. (2006).

[18] The strategic form after elimination of any (pure) strategies that are duplicates or convex combinations of other strategies.

Mertens added a fourth transformation involving convex combinations, to get to the fully reduced strategic form.[19]

As for how to ensure invariance, Kohlberg and Mertens (1986, Sect. 2.7) give the essential idea, although couched in terms of equilibrium. Here, we give a purely decision-theoretic version (which is then directly relevant to epistemic analysis). Fix a decision tree T – i.e., a two-player game tree, where one player (D) is the decision maker and the other player (N) is Nature, and we specify payoffs for D only. Let Λ be the matrix associated with T, where D chooses the row and N chooses the column. Say that T reduces to matrix M if Λ differs from M by the addition of duplicate rows or columns, or rows that are convex combinations of other rows.[20] We then have: *A row in M is admissible if and only if it is rational in every tree T that reduces to M.*

The forward direction uses standard arguments.[21] For the converse, let r be a row in M that is weakly dominated by a mixture of rows σ, and let C be the set of columns on which σ does strictly better than r. Consider the following tree T: (i) D moves first and chooses between the single move $\{r, \sigma\}$ and any of the other rows; (ii) N then moves, in ignorance of this move, and chooses a column; (iii) finally, if D chose $\{r, \sigma\}$ and N chose one of the columns from C, there is a single information set H at which D gets to choose between r and σ. The tree T reduces to M. Also, choosing σ at H will be strictly better for D than choosing r, so r isn't rational in T, as required.

So, in decision theory, admissibility implies invariance – in fact, is equivalent to it. If we build up our game analysis using a decision theory that satisfies admissibility, we can hope to get invariance at this level too. LPS-based decision theory satisfies admissibility. As wanted, the resulting strategic-form solution concepts in Table 2 – $S^{\infty}W$, m-IA, SAS – are all invariant in the Kohlberg and Mertens sense. (See Brandenburger and Friedenberg 2004.) The extensive-form concepts in Table 2 aren't.

Back to the epistemic conditions: RmAR and completeness yields an invariant prediction, because $(m+1)$-IA is invariant. Arguably then, the inconsistency of RCAR and completeness – as opposed to the consistency of RCSBR and completeness – is the price that has to be paid for having an invariant analysis.

In any event, it does seem that some basic requirement has to be given up in the search for the 'fully rational' analysis of games. If not exactly a paradox, this is certainly a surprising situation for game theory. To quote Mertens (1989, p. 583):

[19] The Dalkey–Thompson transformations can be replaced by the Elmes and Reny (1994) transformations, which preserve perfect recall (Kuhn 1953).

[20] Why not consider convex combination of columns? Under the (Anscombe and Aumann 1963) viewpoint, D's payoffs are really expected payoffs over objective lotteries. It is then natural to say that D can mix over these lotteries – creating a row that is a convex combination of other rows. But N does not mix.

[21] If r is an admissible row in M, then it is optimal under some full-support measure q on the columns of M. From q, define a full-support measure on the columns of Λ, and argue that r is also admissible in Λ. Use the new measure to define a CPS on T. Argue that if r isn't rational in T, given this CPS, then it is inadmissible in Λ.

> It is as if every time we think we finally get a hold on what rational behaviour means, we find ourselves having grasped only a shadow. Maybe this means there is excessive $\acute{\upsilon}\beta\rho\iota\varsigma$ in this endeavour: that rationality is something belonging to the gods themselves, and that should not be stolen from them. Maybe it is the tree of knowledge itself, that we should not touch?

Perhaps the problem of inconsistent requirements says we are allowed to know one thing – that rationality in its 'ultimate form' simply cannot be.

13 Knowledge-based approach

Throughout, we have focused on the epistemic literature that thinks of the players in a game as having beliefs about one another's strategies, beliefs about these beliefs, etc. As pointed out right at the start in Example 4.1, these beliefs don't have to be correct in any sense. In general, a player's type needn't even assign positive probability to the actual strategy-type pair of another player (or have that pair in its support in the infinite case).

Knowledge as usually formalized is different from belief, in that if a player knows an event E, then E indeed happens. Knowledge can be present in the belief-based approach, in the form of observation. If his information set H is reached, player i observes (and is correct) that the other players' strategies must be among those that allow H. These observations constitute the conditioning events in a CPS (refer back to Sect. 6). In the strategic-form approach, there is no (non-trivial) knowledge, just the sequence of hypotheses that makes up an LPS.

Philosophically, the overall view is that only observables are knowable. Unobservables are subject to belief, not knowledge. In particular, other players' strategies are unobservables, and only moves are observables.

Another strand in the literature does allow knowledge about the strategies chosen by other players. See, among others, Aumann (1995; 1998), Balkenborg and Winter (1997), Halpern (1999; 2001), Samet (1996), Stalnaker (1996), and also the exchange between Binmore (1996) and Aumann (1996).

There appear to be some connections between the belief-based and knowledge-based approaches, but also significant differences. Counterfactuals play an important role in a knowledge-based analysis, if we want to talk about what a player thinks at an information set that cannot be reached given what he knows. There may be an analogy to the role of extended probabilities in a belief-based analysis. But completeness is crucial to the belief-based approach, as we have seen, and an analogous concept does not appear to be present in the knowledge-based literature. Halpern (2001) is a good synthesis of the knowledge-based approach. We are not aware of any formal treatment of the relationship between this and the belief-based approach we have followed in this survey.

Finally, we mention some papers that use formalisms related to, but again different from, the ones we have covered. Feinberg (2005a;b) builds an extensive-form epistemic framework. His approach is 'local' rather than 'global,' since, instead of a CPS, he specifies at each information set separately what a

player believes there. Asheim and Perea (2005) is another epistemic analysis on the tree, but uses the idea of a "conditional LPS" (Blume et al. 1991a, Definition 4.2) rather than CPS's. (For each conditioning event E, take in sequence all hypotheses that give E positive probability, calculate conditionals, and in this way form an LPS concentrated on E.) Conceptually, conditional LPS's are the right object for epistemic analysis of "weak dominance on the tree"– as opposed to "strong dominance on the tree," which we noted earlier is what comes from a CPS-based analysis.

14 Conclusion

The epistemic program can be viewed as a methodical construction of game theory from its most basic elements – rationality and irrationality, belief (and knowledge), belief about belief, etc. It is a 'bottom-up' approach. In this, it is very different from the 'top-down' approach of the equilibrium refinements program (as noted earlier). It is also very different in that Nash equilibrium has played a much smaller role in the epistemic program. As we have seen, some of the most basic questions lead naturally to other solution concepts.

We have talked about some seemingly inherent limits to the epistemic analysis of games. Such limits aren't a sign of failure. Rather, finding such theoretical limits seems a sign that the epistemic program has reached a certain depth and maturity. Also, several of the examples we looked at involved scenarios that were 'a long way from' these theoretical limits. A big point of the epistemic program is that there isn't one right set of assumptions to make about a game. The inconsistency of certain conditions is important, but not the whole picture. The goal of the program is to be able to analyze many different sets of assumptions about games in a precise and uniform manner.

References

Anscombe F, Aumann R (1963) A definition of subjective probability. Ann Math Stat 34:199–205
Armbruster W, Böge W (1979) Bayesian game theory. In: Möschlin O, Pallaschke D (eds) Game theory and related topics. North-Holland, Amsterdam
Asheim G (2001) Proper rationalizability in lexicographic beliefs. Int J Game Theory 30:453–478
Asheim G, Perea A (2005) Sequential and quasi-perfect rationalizability in extensive games. Games Econ Behav 53:15–42
Aumann R (1987) Correlated equilibrium as an expression of Bayesian rationality. Econometrica 55:1–18
Aumann R (1995) Backward induction and common knowledge of rationality. Games Econ Behav 8:6–19
Aumann R (1996) Reply to Binmore. Games Econ Behav 17:138–146
Aumann R (1998) On the centipede game. Games Econ Behav 23:97–105
Aumann R (1999) Interactive epistemology I: knowledge. In: Lipman B (ed) Special issue on interactive epistemology. Int J Game Theory, 28:263–300
Aumann R, Brandenburger A (1995) Epistemic conditions for Nash equilibrium. Econometrica 63:1161–1180
Balkenborg D, Winter E (1997) A necessary and sufficient epistemic condition for playing backward induction. J Math Econ 27:325–345
Barrow J (1998) Impossibility: the limits of science and the science of limits. Oxford University Press, Oxford

Basu K (1990) On the existence of a rationality definition for extensive games. Int J Game Theory 19:33–44
Battigalli P (1997) On rationalizability in extensive games. J Econ Theory 74:40–61
Battigalli P, Siniscalchi M (1999) Hierarchies of conditional beliefs and interactive epistemology in dynamic games. J Econ Theory 88:188–230
Battigalli P, Siniscalchi M (2002) Strong belief and forward-induction reasoning. J Econ Theory 106:356–391
Ben Porath E (1997) Rationality, Nash equilibrium, and backward induction in perfect information game. Rev Econ Stud 64:23–46
Bernheim D (1987) Rationalizable strategic behavior. Econometrica 52:1007–1028
Bicchieri C (1988) Strategic behavior and counterfactuals. Synthese 76:135–169
Bicchieri C (1989) Self-refuting theories of strategic interaction: a paradox of common knowledge. Erkenntnis 30:69–85
Binmore K (1987) Modelling rational players I. Econ Philos 3:179–214
Binmore K (1996) A note on backward induction. Games Econ Behav 17:135–137
Blume L, Brandenburger A, Dekel E (1991a) Lexicographic probabilities and choice under uncertainty. Econometrica 59:61–79
Blume L, Brandenburger A, Dekel E (1991b) Lexicographic probabilities and equilibrium refinements. Econometrica 59:81–98
Board O (2002) Algorithmic characterization of rationalizability in extensive form games. Available at www.pitt.edu/∼ojboard
Board O (2004) Dynamic interactive epistemology. Games Econ Behav 49:49–80
Böge W, Eisele T (1979) On solutions of Bayesian games. Int J Game Theory 8:193–215
Bonanno G (1991) The logic of rational play in games of perfect information. Econ Philos 7:37–65
Börgers T (1994) Weak dominance and approximate common knowledge. J Econ Theory 64:265–276
Börgers T, Samuelson L (1992) Cautious utility maximization and iterated weak dominance. Int J Game Theory 21:13–25
Brandenburger A (1992) Lexicographic probabilities and iterated admissibility. In: Dasgupta P, Gale D, Hart O, Maskin E (eds) Economic analysis of markets and games. MIT Press, Cambridge, pp. 282–290
Brandenburger A (2003) On the existence of a 'complete' possibility structure. In: Basili M, Dimitri N, Gilboa I (eds) Cognitive processes and economic behavior. Routledge, London pp. 30–34
Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. Econometrica 55:1391–1402
Brandenburger A, Friedenberg A (2004) Notes on self-admissible sets. Available at www.stern.nyu.edu/∼abranden
Brandenburger A, Friedenberg A, Keisler HJ (2006) Admissibility in games. Available at www.stern.nyu.edu/∼ abranden
Brandenburger A, Keisler HJ (2006) An impossibility theorem on beliefs in games. Studia Logica 84:211–240
Dalkey N (1953) Equivalence of information patterns and essentially determinate games. In: Kuhn H, Tucker A (eds) Contributions to the theory of game, vol 2. Princeton University Press, New Jersey, pp. 217–244
Dekel E, Fudenberg D (1990) Rational behavior with payoff uncertainty. J Econ Theory 52:243–267
Elmes S, Reny P (1994) On the strategic equivalence of extensive form games. J Econ Theory 62:1–23
Ewerhart C (2002) Ex-ante justifiable behavior, common knowledge, and iterated admissibility. Available at http://mail.wiwi.uni-bonn.de/users/cewerhart
Fagin R (1994) A quantitative analysis of modal logic. J Symbol Logic 59:209–252
Fagin R, Geanakoplos J, Halpern J, Vardi M (1999) The hierarchical approach to modeling knowledge and common knowledge. In: Lipman B (ed) Special issue on interactive epistemology. Int J Game Theory 28:331–365
Feinberg Y (2005a) Subjective reasoning-dynamic games. Games Econ Behav 52:54–93
Feinberg Y (2005b) Subjective reasoning-solutions. Games Econ Behav 52:94–132
Friedenberg A (2002) When common belief is correct belief. Available at www.olin.wustl.edu/faculty/friedenberg

Gale D (1953) A Theory of *n*-person games with perfect information. Proc Nat Acad Sci, USA 39:496–501
Halpern J (1999) Hypothetical knowledge and counterfactual reasoning. Int J Game Theory 28:315–330
Halpern J (2001) Substantive rationality and backward induction. Games Econ Behav 37:425–435
Harsanyi J (1967–68) Games with incomplete information played by 'Bayesian' players, I–III. Manage Sci 14:159–182, 320–334, 486–502
Heifetz A (1999) How canonical is the canonical model? A comment on Aumann's interactive epistemology. In: Lipman B (ed) Special issue on interactive epistemology. Int J Game Theory 28:435–442
Heifetz A, Samet D (1998) Knowledge spaces with arbitrarily high rank. Games Econ Behav 22:260–273
Heifetz A, Samet D (1999) Coherent beliefs are not always types. J Math Econ 32:475–488
Hu H, Stuart HW (2001) An epistemic analysis of the Harsanyi transformation. Int J Game Theory 30:517–525
Kohlberg E, Mertens J-F (1986) On the strategic stability of equilibria. Econometrica 54:1003–1037
Kuhn H (1953) Extensive games and the problem of information. In: Kuhn H, Tucker A (eds) Contributions to the theory of Games, vol 2. Princeton University Press, New Jersey, pp. 193–216
Marx L, Swinkels J (1997) Order independence for iterated weak dominance. Games Econ Behav 18:219–245
Mas-Colell A, Whinston M, Green J (1995) Microeconomic theory. Oxford University Press, Oxford
Meier M (2005) On the nonexistence of universal information structures. J Econ Theory 122:132–139
Mertens J-F (1989) Stable equilibria – a reformulation. Part 1. Definition and basic properties. Math Oper Res 14:575–624
Mertens J-F, Zamir S (1985) Formulation of Bayesian analysis for games with incomplete information. Int J Game Theory 14:1–29
Myerson R (1978) Refinements of the Nash equilibrium concept. Int J Game Theory 1:73–80
Myerson R (1991) Game theory. Harvard University Press, Cambridge
Osborne M, Rubinstein A (1994) A course in game theory. MIT Press, Cambridge
Pearce D (1984) Rational strategic behavior and the problem of perfection. Econometrica 52:1029–1050
Rapaport A (1967) Escape from paradox. Sci Am 217:50–56
Reny P (1992) Rationality in extensive form games. J Econ Perspect 6:103–118
Rényi A (1955) On a new axiomatic theory of probability. Acta Math Acad Sci Hungar 6:285–335
Rosenthal R (1981) Games of perfect information, predatory pricing and the chain-store paradox. J Econ Theory 25:92–100
Samet D (1996) Hypothetical knowledge and games with perfect information. Games Econ Behav 17:230–251
Samuelson L (1992) Dominated strategies and common knowledge. Games Econ Behav 4:284–313
Shimoji M (2004) On the equivalence of weak dominance and sequential best response. Games Econ Behav 48:385–402
Shimoji M, Watson J (1998) Conditional dominance, rationalizability and game forms. J Econ Theory 83:161–195
Sorin S (1998) On the impact of an event. Int J Game Theory 27:315–330
Special Issue on Interactive Epistemology (1999) Lipman B (ed). Int J Game Theory 28
Stahl D (1995) Lexicographic rationalizability and iterated admissibility. Econ Lett 47:155–159
Stalnaker R (1996) Knowledge, belief and counterfactual reasoning in games. Econ Philos 12:133–163
Stalnaker R (1998) Belief revision in games: forward and backward induction. Math Soc Sci 36:31–56
Tan T, Werlang S (1988) The Bayesian foundations of solution concepts of games. J Econ Theory 45:370–391
Thompson F (1952) Equivalence of games in extensive form. Research Memorandum RM-759, The RAND Corporation