Chapter 3
Extensive Games

Game theorists may disagree about the suitability of one or another normal form solution concept—there are many—but a brief inspection of the literature will show that they rarely disagree about the correct epistemic characterisation of each solution concept. For extensive games, while several solution concepts are available, the most obvious candidate is certainly in most cases backward induction (the subgame-perfect (Nash) equilibrium). Ironically, however, game theorists widely disagree about its correct epistemic characterisation, and the disagreement centres on the question of whether or not common true belief about rationality leads to backward induction. Robert Aumann defends a position in favour of this implication, Philip Reny objects, and both positions are taken by various other theorists.[1]

The purpose of this chapter is not to substantiate one line of argument or another. Rather, by analysing the logical form of the arguments *à la* Aumann and *à la* Reny, I will point out that important modelling assumptions have been overlooked. Devising a logical formalism that captures the two different interpretations of what extensive games in fact model, I will argue that the argument *à la* Aumann assumes the one-shot interpretation in combination with a principle of rationality with which it is incompatible, and I will argue that the argument *à la* Reny assumes the many-moment interpretation in combination with implausible belief revision policies.[2]

[1] Robert Aumann, 'Backward Induction and Common Knowledge of Rationality', *Games and Economic Behavior*, 8 (1995), 6–19, ibid., 'Reply to Binmore', *Games and Economic Behavior*, 17 (1996), 138–146, and ibid., 'On the Centipede Game', *Games and Economic Behavior*, 23 (1998), 97–105. Philip Reny, 'Two Papers on the Theory of Strategic Behaviour', Ph.D. diss. (Princeton University, 1988), 'Common Knowledge and Games with Perfect Information', in A. Fine and J. Leplin (eds.), *Proceedings of the Philosophy of Science Association: Volume 2* (East Lansing, Mich.: Philosophy of Science Association, 1989), 363–369, ibid., 'Common Belief and the Theory of Games with Perfect Information', *Journal of Economic Theory*, 59 (1993), 257–274, and ibid., 'Rational Behaviour in Extensive-Form Games', *Canadian Journal of Economics*, 28 (1995), 1–16.

[2] Joseph Halpern, 'Substantive Rationality and Backward Induction', *Games and Economic Behavior*, 37 (2001), 321–339 compares Aumann's approach to backward induction with Stalnaker's game models. In contrast to the comparison in the current chapter, Halpern does not distinguish

B. de Bruin, *Explaining Games*, Synthese Library 346,
DOI 10.1007/978-1-4020-9906-9_3, © Springer Science+Business Media B.V. 2010

3.1 The One-Shot Interpretation

3.1.1 The Epistemic Characterisation Result

Common true belief about rationality and utility, in extensive form games, yields backward induction, or so Robert Aumann and others claim.[3] To understand this result, define the *normal form* $\mathrm{nf}(\Gamma)$ of an extensive game Γ as a triple $(I,(A_i)_i,(v_i)_i)$ where I collects the players of Γ, A_i all strategies player i has in Γ, and $v_i\colon \prod_i A_i \rightarrow \mathbb{R}$ are utility functions such that

$$v_i(1_{k_1},\dots,i_{k_i},\dots,N_{k_N}) = u_i(O(1_{k_1},\dots,i_{k_i},\dots,N_{k_N})),$$

where O is a function mapping a tuple of strategies to the terminal node of the extensive game that is reached when the players play according to these strategies. I write $u_i(k,l)$ for $v_i(O(k,l))$. If $\mathrm{nf}(\Gamma) = (I,(A_i)_i,(v_i)_i)$, and if X_1, X_2, and so on, are sets of strategies satisfying $X_i \subseteq A_i$ for all i, then the *subspan* of $\mathrm{nf}(\Gamma)$ with respect to $\prod_i X_i$ is the triple $(I,(X_i)_i,(v_i|_{X_i})_i)$ obtained from $\mathrm{nf}(\Gamma)$ by removing for all i the strategies in the complement of X_i (with respect to A_i) and modifying the utility functions correspondingly.

In line with the previous chapter, notation $\mathrm{nsd}_i(X_1,\dots,X_N)$ is extended to extensive game-playing settings for the strategies that are not strictly dominated for player i in the subspan of $\mathrm{nf}(\Gamma)$ with respect to $\prod_i X_i$; that is, strategies for which there is no strategy in X_i which does strictly better against any combination of opponents' strategies. In general, I am interested in dominance relations in subspans of the normal form of subgames of some underlying game generated by certain decision nodes; that is, in constructs of the form $\mathrm{nsd}_i^x(X_1,\dots,X_N)$, where $X_i \subseteq A_i$, for some decision node x.

To compute such sets—the idea is simpler than the construction—consider the subgame Γ_x of Γ generated by some decision node x, construct its normal form $\mathrm{nf}(\Gamma_x)$, delete from $\mathrm{nf}(\Gamma_x)$, for all j, the strategies not coinciding on Γ_x with any strategy from X_j, find out which of the remaining strategies in the resulting subspan of $\mathrm{nf}(\Gamma_x)$ are strictly undominated, and then take all strategies from A_i coinciding on

between one-shot and many-moment interpretations of extensive game-play, nor does he give an inductive and implicit axiomatisation of rationality.

[3] 'Backward Induction and Common Knowledge of Rationality' contains a defence of this claim, based on the one-shot interpretation. Assuming a many-moment perspective, Aumann defended the same claim for smaller classes of extensive games in 'On the Centipede Game', a result discovered independently by Wlodek Rabinowicz, 'Grappling with the Centipede: Defence of Backward Induction for BI-terminating Games', *Economics and Philosophy*, 14 (1999), 95–126, John Broome and Wlodek Rabinowicz, 'Backwards Induction in the Centipede Game', *Analysis*, 59 (1999), 237–242. Magnus Jiborn and Wlodek Rabinowicz, 'Backward Induction without Full Trust in Rationality', in W. Rabinowicz (ed.), *Value and Choice: Some Common Themes in Decision Theory and Moral Philosophy: Volume 2* (Lund: Lund Philosophy Reports, 2001), 101–120 prove a many-moment characterisation for the Centipede on the basis of sufficiently strong, but not necessarily full beliefs about rationality.

Γ_x with such a strictly undominated strategy. For weak dominance, define nwd_i and its relativisations similarly.

The set containing the strategies that coincide with the backward induction strategy on the subgame generated by x is written BI_i^x, and in the logic the set of corresponding proposition letters is written BI_i^x. Assuming that the extensive game is *generic* in the sense that no player is indifferent between any two terminal nodes, BI_i^x contains all strategies prescribing the uniquely optimal action at x if x is an immediate predecessor of a terminal node. Reasoning back to the root of the game, BI_i^z collects all strategies prescribing the uniquely optimal action at decision node z under the assumption that at decision nodes $y \succ z$ higher up in the game tree all players j plays a strategy from BI^j. For the root ρ of the game, BI_i^ρ is a singleton also written BI_i. For terminal nodes x the convention is applied that $\mathsf{BI}_i^x = A_i$.

As I have suggested, while the mathematical differences between normal form games and extensive games strongly suggest that the former model situations of simultaneous and independent choice and the latter model temporally extended situations of sequential choice, we are by no means obliged to make such an interpretation. The one-shot interpretation, in fact, holds to the view that playing an extensive game is playing its normal form; that is, whenever players play an extensive game, what they actually do is choose, at one point in time, their strategies for the entire game. This does not completely determine a unique one-shot interpretation, and hence there is room for difference of opinion about the relevant kind of rationality principles. One can invoke some aspects of the sequential structure of the game to ascertain whether a strategy is rational or not, for while a strategy maps all decision nodes of a player to actions, choosing, for some decision node, one action over another implies that certain decision nodes will not be reached—the terminology is admittedly inappropriate in a one-shot context—and whether these unreached nodes are relevant or not determines different notions of rationality.

To capture the differences, I call a strategy *on-path rational* whenever the rationality depends only on what happens on the actual path through the extensive game, and I call it *off-path rational* just in case it prescribes rational actions at every decision node, reached or unreached. Aumann, for instance, asserts that

> each player chooses a *strategy*, in the usual game-theoretic sense of the term...; that is, he decides what to do at each of his vertices x in the game tree, whether or not x is reached.[4]

This seems to demonstrate that he adopts a one-shot conception of extensive game-play. Yet he also writes that ‘when deciding what to do at x, the player considers the situation *from that point on:* he acts *as if* x is reached’, to conclude that

> it is this feature that distinguishes the current analysis from a strategic [i.e., normal] form analysis.[5]

[4] ‘Backward Induction and Common Knowledge of Rationality’, 7 (notation changed). As I am more interested here in what Aumann wishes to model, than in the resulting model itself, the stress lies on his verbal statements rather than on his formalism. This also applies to my treatment of Reny in the next section.

[5] Ibid. (emphasis in original, notation changed).

Attributing the one-shot interpretation to Aumann and also accepting this conclusion, there seems to be a difference between the one-shot interpretation and playing the normal form of an extensive game, and this is, in fact, more or less how Aumann seems to view it. He does accept the main tenet of the one-shot interpretation that the objects of choice of an extensive game are the strategies of its normal form, but he contrasts his view with 'strategic form analysis'. The reason for this is that he holds the view that in order to evaluate the rationality of a strategy, one has to go beyond the information of the normal form and inspect the prescriptions of the strategy at all decision nodes of the underlying extensive game; that is, Aumann adopts a one-shot interpretation of extensive game-play with an off-path conception of rationality. This is underscored by the statement that a rational player,

> no matter where he finds himself—at which vertex—[,]. . . will not knowingly continue with a strategy that yields him less than he could have gotten with a different strategy,[6]

as well as by the remark that

> for each of his vertices *x* and strategies *k*, it is not the case that [player] *i* knows that *k* would yield him a higher conditional payoff at *x* than the strategy he chooses.[7]

All in all, Aumann adopts a one-shot interpretation with off-path rationality. This is the same as playing normal form games as far as the objects of choice are concerned, but it is different with respect to the rationality principle. This is a view that I discarded in Chapter 1, but it will return in the discussion later on.

There is a problem, though. In some sense, the phrase about players who continue 'not knowingly' suggests many-moment game-playing situations. Instead of taking Aumann to put forward an incongruent claim here, I take knowledge to refer to beliefs that, in one-shot game-playing situations, players have about the moves that the full strategies of their opponents prescribe in (certain) subgames. The last quotation means, in that case, that no rational player will choose a full strategy that prescribes suboptimal moves at some decision node given the beliefs that the player has, in that one-shot game-playing situation, about what her opponents' choices of full strategies will prescribe in the subgame generated by that decision node.

3.1.1.1 An Explicit Formalisation of Rationality

I will present a direct formalisation to make this precise. While it has the advantage of staying close to the sources, it makes the proof of the epistemic characterisation theorem cumbersome, and dependent on heavy logical axioms. I will therefore subsequently turn to an alternative formalisation.

Following the above quotation quite literally, we have

$$\bigwedge_{x}\bigwedge_{k}\neg\bigvee_{l}\bigvee_{m}(\Box_i\mathbf{i}_k\wedge\Box_i\mathbf{j}_l\wedge\mathbf{u}_i^x(k,l)<\mathbf{u}_i^x(m,l)),$$

[6] Ibid.

[7] Ibid. 10 (notation changed).

which says that if player i believes that her opponent is playing his lth strategy, and i herself believes she is playing her kth strategy (and she is really playing her kth strategy), then there is no better strategy m than her kth strategy. This is equivalent with

$$\bigwedge_x \bigwedge_k \bigwedge_l \bigwedge_m \neg(\Box_i \mathbf{i}_k \wedge \Box_i \mathbf{j}_l \wedge \mathbf{u}_i^x(k,l) < \mathbf{u}_i^x(m,l)),$$

and with

$$\bigwedge_x \bigwedge_k \bigwedge_l ((\Box_i \mathbf{i}_k \wedge \Box_i \mathbf{j}_l) \rightarrow \bigwedge_m (\mathbf{u}_i^x(k,l) \geq \mathbf{u}_i^x(m,l))),$$

which brings out nicely the similarity with other rationality notions. This motivates the following two axioms as renderings of Aumann's notion of rationality

ANRat $\mathbf{anrat}_i^x \leftrightarrow \bigwedge_k \bigwedge_l ((\Box_i \mathbf{i}_k^x \wedge \Box_i \mathbf{j}_l^x) \rightarrow \bigwedge_m (\mathbf{u}_i^x(k,l) \geq \mathbf{u}_i^x(m,l)).$
AFRat $\mathbf{afrat}_i \leftrightarrow \bigwedge_{\rho \preceq x} \mathbf{anrat}_i^x.$

To prove the epistemic characterisation theorem, the proof system $_\Gamma\mathbf{KT_{EC}45afrat}$ is used containing Prop, Dual, K, T, 4, 5, E, C, the proof rules modus ponens, necessitation and induction, all axioms for one-shot game-playing situations, plus the two rationality axioms.

Theorem 3.1 (Aumann, 1995) *Let Γ be a finite N-person generic extensive form game with perfect information. Assume that the following two conditions are true.*

1. *There is common true belief among the players about the utility functions of all players.*
2. *It is common true belief among the players that they are rational.*

Then the backward induction outcome is reached.

To give some impression of how this theorem is proven, it can be demonstrated that for all players i we have

$$\forall x \vdash \mathbf{Cafrat} \rightarrow BI_i^x.$$

The rule of necessitation, the K-axiom, and some propositional logic and some aggregation of proofs makes it possible to derive from the relevant inductive hypothesis that

$$\mathbf{Cafrat} \rightarrow (\Box_i \bigwedge_{x \prec y} BI_i^y \wedge \Box_i \bigwedge_{x \prec y} BI_j^y).$$

This can be used to show that we have

$$\mathbf{Cafrat} \rightarrow \neg\Box_i \neg BI_i^x, \tag{3.1}$$

from which, first

$$\mathbf{Cafrat} \rightarrow \neg\Box_i \neg\Box_i BI_i^x$$

by means of the T- and the KnStrat-axiom, and second,

$$\mathbf{Cafrat} \rightarrow BI_i^x,$$

by means of the 5-axiom, classical negation and the T-axiom.

The idea underlying a proof of 3.1 is to derive a contradiction from $\mathbf{Cafrat} \wedge \Box_i \neg BI_i^x$, and it is here that the rather manipulated inductive hypothesis is used. The contradiction becomes apparent as soon as it is observed that on the basis of this assumption, one would show

$$\mathbf{Cafrat} \rightarrow (\mathbf{afrat} \wedge \Box_i \bigwedge_{x \prec y} BI_i^y \wedge \Box_i \bigwedge_{x \prec y} BI_j^y \wedge \Box_i \neg BI_i^x),$$

which could be specialised to get a statement expressing that from **Cafrat** it follows that for some strategy k player i knows, first, that she plays k in the subgame generated by x, second, that k is not the inductive strategy in that subgame, and, third, that k yields strictly less than the inductive strategy in that subgame. This means that, first, knowledge of one's strategy is needed, second, some technical axioms about equivalence of strategies (taking care of subgames), and third, the fact that a generic game's utility function is an injection is essential for obtaining strict inequality. Moreover, a result about inductive strategies is needed to the effect that in generic games the inductive strategy is strictly better in some subgame given that opponents play inductively in that subgame. This is a statement with the same structural function as Lemma 3.1 (used below), connecting the BI_i^x and inequality statements with u_i^x about utility.

3.1.1.2 An Implicit, Inductive Formalisation of Rationality

This is all fine, but it is slightly cumbersome. Moreover, a proof system is used in which the T-axiom (veridicality), the 4-axiom (positive introspection) and the 5-axiom (negative introspection) figure. There is, however, no need to use such heavy machinery—endangering the general format of epistemic characterisation results in the way witnessed by that of the Nash equilibrium—once we adopt an inductive and implicit axiomatisation of rationality by means of the following three axioms.

NRat$_{bas}$ $\mathbf{nrat}_i^x \rightarrow \mathrm{nsd}_i^x(A_i, A_j)$.
NRat$_{ind}$ $(\mathbf{nrat}_i^x \wedge \Box_i X_i \wedge \Box_i X_j) \rightarrow \mathrm{nsd}_i^x(X_i, X_j)$.
FRat $\mathbf{frat}_i \leftrightarrow \bigwedge_{\rho \preceq x} \mathbf{nrat}_i^x$.

These axioms need some explanation. First, a preliminary remark about applying on-path rationality to subgames is appropriate. On the one-shot interpretation, objects of choice (full strategies) are always functions mapping all decision nodes of a player to actions, and consequently beliefs are about the full strategies opponents choose. It makes perfect sense, though, to speak about the on-path rationality of a strategy in any subgame, for one can consider the restriction to the subgame of a strategy to evaluate its rationality as a course of action in the subgame and in light of the restrictions to the subgame of the strategies one expects one's opponents to play. This idea is captured in the first two axioms.

What is rational often depends on one's beliefs, but not always; and as I have argued, this forms the basis of the implicit and inductive formalisation of rationality. The first-axiom captures the base case without beliefs. It states that player i, if on-path rational in the subgame Γ_x generated by x, never chooses a strategy which prescribes bad actions independently of what her opponents play. If player i is on-path rational in Γ_x, she does not choose any strategy of which the restriction to Γ_x coincides with a strategy strictly dominated in the normal form of Γ_x.

The second axiom states that, if she is on-path rational in Γ_x, player i never plays a strategy that is strictly dominated in the normal form of Γ_x from which those strategies (both of her opponents as well as herself) have been removed that she believes will not be chosen. The beliefs are represented by sets X_i and X_j of strategies. Finally, the third axiom states that player i is off-path rational in the entire game if she is on-path rational in all of its subgames. The proof system ${}_{\Gamma}\mathbf{K}_{\mathbf{EC}}\mathbf{frat}$ used consists of axioms Prop, Dual, K, E, C, the proof rules modus ponens, necessitation and induction, all axioms for one-shot game-playing situations, plus the three above rationality axioms. A formalisation of the assumptions can be found in Table 3.1.

Table 3.1

Assumptions	
preferences	$\bigwedge_{i,k,l} \mathbf{u}_i(k,l) = \mathbf{r}_{i,k,l}$
principles	$\bigwedge_i \mathbf{frat}_i$
beliefs	
preferences	$\mathbf{C} \bigwedge_{i,k,l} \mathbf{u}_i(k,l) = \mathbf{r}_{i,k,l}$
principles	$\mathbf{C} \bigwedge_i \mathbf{frat}_i$
performed action	–
Solution Concept	
player i	BI_i
Proof System	${}_{\Gamma}\mathbf{K}_{\mathbf{EC}}\mathbf{frat}$

To prove the theorem we need two lemmas. To collect all propositional formulae for backward induction strategies in any subgame Γ_x, first take all strategies that prescribe backward induction actions in all real subgames of Γ_x to obtain the set $\bigcap_{x \prec y} BI_i^y$. Some of the strategies in this set, however, do not prescribe the backward induction action at x, and therefore attention has to be restricted to those elements for which there is no strictly better alternative given that all players take backward induction actions at decision nodes $y \succ x$. This is establishes the first lemma.

Lemma 3.1 $BI_i^x = \mathrm{nsd}_i^x(\bigcap_{x \prec y} BI_1^y, \dots, \bigcap_{x \prec y} BI_N^y) \cap \bigcap_{x \prec y} BI_i^y$.

In the formalism proposed, the formula $\bigwedge_{x \prec y} \bigvee BI_i^y$ states that at any $y \succ x$ player i plays according to backward induction (if y is a decision node of hers). The intersection $\bigcap_{x \prec y} BI_i^y$ not being empty (it contains all strategies available to i in Γ that prescribe backward induction actions in Γ_x), it is simple to observe that $\vdash_{\Gamma \mathbf{K}_{\mathbf{C}}\mathbf{frat}} \bigwedge_{x \prec y} \bigvee BI_i^y \rightarrow \bigvee \bigcap_{x \prec y} BI_i^y$. With the convention to omit disjunction symbols in front of sets of propositional formulae, this establishes the second lemma.

Lemma 3.2 $\vdash_{\Gamma\mathbf{K_C frat}} \bigwedge_{x\prec y} BI_i^y \rightarrow \bigcap_{x\prec y} BI_i^y$.

I can now turn to a proof of Theorem 3.1 in the alternative formalism. In fact, the theorem can now be phrased in terms of true belief rather than knowledge.

Proof I prove the result for $N = 2$ with players i and $j \neq i$. For more than two players one only needs to add the relevant conjuncts and to expand nsd_i^x to a function taking three or more arguments. Clearly $\mathbf{frat}_i \rightarrow \mathbf{nrat}_i^x$ by axiom FRat, and the case of a decision node x with depth $d(x) = 1$ reduces to axiom NRat_{bas} with an application of Lemma 3.1. Let $d(x) > 1$. The inductive hypothesis gives for every $y \succ x$

$$(\mathbf{Cfrat} \wedge \mathbf{frat}) \rightarrow BI_i^y.$$

Because I consider finite games, the proofs for all $y \succ x$ and both players i and j can be aggregated into

$$(\mathbf{Cfrat} \wedge \mathbf{frat}) \rightarrow (\bigwedge_{x\prec y} BI_i^y \wedge \bigwedge_{x\prec y} BI_j^y),$$

and, applying Lemma 3.2, into

$$(\mathbf{Cfrat} \wedge \mathbf{frat}) \rightarrow (\bigcap_{x\prec y} BI_i^y \wedge \bigcap_{x\prec y} BI_j^y).$$

An application of the necessitation rule for i and the K-axiom, together with some propositional reasoning, yields

$$\mathbf{Cfrat} \rightarrow (\Box_i \bigcap_{x\prec y} BI_i^y \wedge \Box_i \bigcap_{x\prec y} BI_j^y).$$

Since $\mathbf{frat}_i \rightarrow \mathbf{nrat}_i^x$, we find

$$(\mathbf{Cfrat} \wedge \mathbf{frat}) \rightarrow (\mathbf{nrat}_i^x \wedge \Box_i \bigcap_{x\prec y} BI_i^y \wedge \Box_i \bigcap_{x\prec y} BI_j^y).$$

Applying the NRat_{ind}-axiom gives

$$(\mathbf{Cfrat} \wedge \mathbf{frat}) \rightarrow \mathrm{nsd}_i^x(\bigcap_{x\prec y} BI_i^y, \bigcap_{x\prec y} BI_j^y).$$

Invoking the inductive hypothesis again, and applying Lemma 3.2, the consequent of this formula can be made somewhat more precise in

$$(\mathbf{Cfrat} \wedge \mathbf{frat}) \rightarrow \mathrm{nsd}_i^x(\bigcap_{x\prec y} BI_i^y, \bigcap_{x\prec y} BI_j^y) \cap \bigcap_{x\prec y} BI_i^y,$$

which is

$$(\mathbf{Cfrat} \wedge \mathbf{frat}) \rightarrow BI_i^x$$

by Lemma 3.1.

3.1.2 Discussion

I have presented two versions of the one-shot interpretation of extensive game-play: one with on-path rationality and one with off-path rationality. Aumann was seen to adopt the latter version. I do not believe, however, that the latter version is conceptually consistent; off-path rationality is strictly incompatible with the true spirit of the one-shot interpretation. The reason is that there is no sensible rationale to take care of what would happen at unreached, off-path nodes in a situation in which the objects of choice are strategies from the normal form of an extensive game. In a one-shot situation, it just does not make sense to talk about nodes being reached or not. The game-playing situation is a strategic predicament in which the players choose a strategy that fixes a complete plan of action for the entire game. Temporal deliberation is senseless, as is thinking about players having beliefs at various points in a temporally extended sequence of decision moments. No nodes are reached or unreached. There is only one decision moment and the outcome of the game is determined on the basis of the strategies the players choose at that precise decision moment.

Does the fact that the one-shot interpretation leaves no room for rationality notions transcending the normal form entail that the epistemic characterisation result of backward induction fails to be significant, or that backward induction cannot be epistemically characterised in a one-shot interpretation? I answer the first question in the affirmative. There is no sense to any epistemic characterisation that presupposes the one-shot interpretation together with a form of rationality that goes beyond on-path rationality by using the specific structural properties of extensive games. The second question, however, need not be answered in the affirmative. It is not difficult to see that once you rephrase the NRat-axioms in terms of weak rather than strict dominance, backward induction can be characterised on the basis of on-path rationality at the root of the game. Given an extensive game with perfect information Γ, let proof system ${}_{\Gamma}\mathbf{K_C Nrat'}$ consist of the following axioms: Prop, Dual, K, C, the proof rules modus ponens, necessitation and induction, all axioms for one-shot game-playing situations for Γ, plus the following two rationality axioms.

NRat'_{bas} $\quad \mathbf{Nrat'}^{x}_{i} \rightarrow \mathrm{nwd}^{x}_{i}(A_i, A_j).$
NRat'_{ind} $\quad (\mathbf{Nrat'}^{x}_{i} \wedge \Box_i X_i \wedge \Box_i X_j) \rightarrow \mathrm{nwd}^{x}_{i}(X_i, X_j).$

The following theorem captures the relation between backward induction and common true belief about rationality in terms of weak dominance.

Theorem 3.2 *Let Γ be a finite generic N-person extensive game with perfect information. Then*

$$\vdash_{{}_{\Gamma}\mathbf{K_C Nrat'}} (\mathbf{CNrat'}^{\rho} \wedge \mathbf{Nrat'}^{\rho}) \rightarrow \bigwedge_{i} BI^{\rho}_{i}.$$

To prove this theorem, first observe that on the level of the normal form of the extensive game, the relevant solution concept is iterated weak dominance. Although the actual outcome of a process of iterative elimination of weakly dominated strategies depends on the exact definition of the elimination algorithm, a lemma due to Hervé Moulin shows this to be irrelevant for current purposes.[8]

Lemma 3.3 *Let Γ be the a finite generic N-person extensive game with perfect information, and let* $\mathrm{nf}(\Gamma)$ *be its normal form. Then*

1. *Any* natural *algorithm for iterated weak dominance yields precisely one strategy profile in* $\mathrm{nf}(\Gamma)$.
2. *All of these algorithms yield the same strategy profile.*
3. *The strategies from this profile correspond to the backward induction strategies of* Γ.

The proof of Theorem 3.2 is a direct analogue of the proof of Theorem 3.1. This precipitates us, of course, into the problem encountered earlier concerning the dubious status of the solution concept of iterated weak dominance. I will not reiterate this point, but instead proceed to the many-moment interpretation.

3.2 The Many-Moment Interpretation

3.2.1 The Inconsistency Result

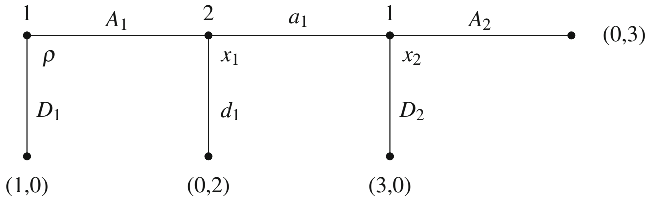

Fig. 3.1

Common true belief about rationality and utility, in extensive form games, does not yield backward induction. Or so Philip Reny and others claim.[9] To defend this

[8] *Game Theory for the Social Sciences* (New York: New York University Press, 1986).

[9] Philip Reny, art. cit. Kaushik Basu, ‘Strategic Irrationality in Extensive Games’, *Mathematical Social Sciences*, 15 (1988), 247–260 investigates temporality aspects. Elchanan Ben-Porath, ‘Rationality, Nash Equilibrium and Backwards Induction in Perfect-Information Games’, *Review of Economic Studies*, 64 (1997), 23–46 develops an account in terms of common certainty. Christina Bicchieri, ‘Common Knowledge and Backward Induction: A Solution to the Paradox’, in M. Vardi (ed.), *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge: March 7–9, 1988, Pacific Grove, California* (Los Altos, Calif.: Morgan Kaufmann,

claim, Reny proves an inconsistency result in the context of the many-moment interpretation of extensive game-play where, in contrast to the one-shot interpretation, games are seen as models of a temporal succession of many decision moments. Referring to the game shown in Figure 3.1, for instance, Reny writes that

> I claim that if player one does not take the dollar and end the game in the first round [does not play D_1], but instead leaves it so that player 2 must decide whether or not to take the two dollars [whether or not to play d_1], then it is no longer possible for rationality to be common knowledge. (i.e. At [*sic*] player two's information set, it is not possible for rationality to be common knowledge).[10]

Such reasoning makes no sense in the one-shot interpretation, according to which no decision nodes are reached at all. On the contrary, strategies are chosen which may or may not induce a path through the game tree to reach some decision node. But it does not make sense to talk about the beliefs of the players at those decision nodes. The players have beliefs at the moment they choose their strategy, but the game stops after that. Reny, by contrast, considers beliefs of players at some decision moment—typical of the many-moment interpretation. Such beliefs describe the expectations of a player about what actions will be taken at decision nodes in the subgame generated by the current decision node.

In principle, as I have argued, the many-moment interpretation leaves two possibilities open. The beliefs at some decision moment can, first, be viewed as dependent on what happened before, and be sensitive to the history of the decision moment in the sense, for instance, that a player could decide to believe her opponent to be irrational if the current decision node can only be reached by the irrational play of her opponent. Second, the beliefs can be viewed as completely insensitive to history. It will be seen that Reny's inconsistency result presupposes a history-sensitive view of belief formation.

Before presenting a formalisation of Reny's inconsistency theorem criticising the epistemic characterisation of backward induction in terms of common true belief about rationality and utility, I should spell out the purpose of the present section, and set down some notation. In the preceding section I developed a logical framework for the one-shot interpretation of extensive games and applied it to Aumann's result on common knowledge of rationality and backward induction. Likewise, I here de-

1988), 381–393, and ead., 'Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge', *Erkenntnis*, 30 (1989), 69–85 develop the view that players can be seen as possessing theories reflecting the epistemic set-up of game-playing. Kenneth Binmore, 'Rationality and Backward Induction', *Journal of Economic Methodology*, 4 (1997), 23–41 zooms in on counterfactual reasoning. Thorsten Clausing, 'Doxastic Conditions for Backward Induction', *Theory and Decision*, 54 (2003), 315–336 sets up a truly doxastic system related to ours. Robert Stalnaker, 'Belief Revision in Games: Forward and Backward Induction', *Mathematical Social Sciences*, 36 (1998), 31–56 studies these issues from the perspective of game models and belief revision theory. More philosophical essays on backward induction include Philip Pettit and Robert Sugden, 'The Backward Induction Paradox', *The Journal of Philosophy*, 86 (1989), 169–182, Jordan Howard Sobel, 'Backward-Induction Arguments: A Paradox Regained', *Philosophy of Science*, 60 (1993), 114–133, and Roy Sorensen, 'Paradoxes of Rationality', in A. Mele (ed.), *The Handbook of Rationality* (Oxford: Oxford University Press, 2004).

[10] Reny, 'Common Knowledge and Games with Perfect Information', 364–365.

velop a logical framework for the many-moment interpretation. Rather than using it to characterise backward induction from the point of view of the many-moment interpretation, I turn to an inconsistency theorem which denies that common knowledge of rationality entails backward induction. It should be stressed, though, that the many-moment perspective has also been adopted to defend the implication of backward induction by common knowledge of rationality. While such forms of defence characterise backward induction in subclasses of extensive games only, they do not make the dubious assumptions that underlie the inconsistency result that I ultimately reject in this section.[11]

Recall that I use super-scripted beliefs to indicate beliefs at decision moments at which the respective decision is reached. Then first define, on the basis of beliefs $\mathbf{P}_i^x(\mathbf{i}_k^x) = \mathbf{p}_k$ and $\mathbf{P}_i^x(\mathbf{j}_l^x) = \mathbf{p}_l$ an auxiliary notion of the expected utility conditional on reaching some immediate successor $y \succ x$ as

$$\mathrm{EU}_i(y, \mathbf{P}_i^x) = \sum_{k,l} \mathbf{p}_k \mathbf{p}_l \mathbf{u}_i^y(k,l).$$

Then define $\mathrm{EU}_i(k, x, \mathbf{P}_i^x)$, the intended interpretation being the expected utility of playing according to the kth strategy at the decision moment at which node x is reached, as

$$\mathrm{EU}_i(k, x, \mathbf{P}_i^x) = \mathrm{EU}_i(y, \mathbf{P}_i^x)$$

for the immediate successor y that is reached when at x player i plays according to his kth strategy.

3.2.1.1 An Explicit Formalisation of Rationality

To formalise rationality, the principle of expected utility maximisation can be relativised to subgames thus.

RRat $\mathbf{rrat}_i^x \leftrightarrow ((\Box_i^x \bigwedge_{k,l} \mathbf{u}_i^x(k,l) = \mathbf{r}_{i,k,l} \wedge \bigwedge_k \mathbf{P}_i^x(\mathbf{i}_k^x) = \mathbf{p}_k \wedge \bigwedge_l \mathbf{P}_i^x(\mathbf{j}_l^x) = \mathbf{p}_l \wedge \mathbf{i}_m(x)) \rightarrow \bigwedge_k \mathrm{EU}_i(m, x, \mathbf{P}_i^x) \geq \mathrm{EU}_i(k, x, \mathbf{P}_i^x))$.

The antecedent of the right hand side contains a condition on the beliefs about the utility structure, and on probabilistic beliefs about what player i herself and her opponent j will play. In the consequent it is stated that i will maximise her expected utility given her beliefs.

[11] Many-moment advocates of the implication of backward induction by common knowledge of rationality (or even by a weakening of these assumptions) include Robert Aumann, ‘On the Centipede Game’, *Games and Economic Behavior*, 23 (1998), 97–105, Wlodek Rabinowicz, ‘Grappling with the Centipede: Defence of Backward Induction for BI-terminating Games’, *Economics and Philosophy*, 14 (1999), 95–126, John Broome and Wlodek Rabinowicz, ‘Backwards Induction in the Centipede Game’, *Analysis*, 59 (1999), 237–242, and Magnus Jiborn and Wlodek Rabinowicz, ‘Backward Induction without Full Trust in Rationality’, in W. Rabinowicz (ed.), *Value and Choice: Some Common Themes in Decision Theory and Moral Philosophy: Volume 2* (Lund: Lund Philosophy Reports, 2001), 101–120.

We need additional axioms, however, to fix the belief formation policies of the players. First, players do not revise their beliefs during game-play as long as this does not lead to inconsistency.

StratPers $\quad \bigwedge_i \bigwedge_j \mathbf{P}_i^x(\mathbf{j}_k^z) = \mathbf{P}_i^y(\mathbf{j}_k^z),$

for $x \preceq y \preceq z$. This persistence axiom states that if $x \preceq y \preceq z$, then the beliefs that player i has at x about the action of her opponent or herself at z will be the same at y. Of course, if game-play has passed z and the beliefs have been contradicted, then i will have different beliefs. But as long as z has not been reached the beliefs remain constant.

While this axiom concerns beliefs about strategies, we need another axiom that involves beliefs about rationality. It states that a player never gives up her beliefs about someone's rationality as long as that person has not moved; in more technical language, the axiom states that if i believes at x that j is rational at some future node y, then i will not change that belief as long as j has not moved.

RatPers $\quad \bigwedge_i \bigwedge_j (\Box_i^x \mathbf{rrat}_j^y \leftrightarrow \Box_i^x \mathbf{rrat}_j^z),$

where $x \preceq y \prec z$, $\iota(z) = j$, and no u with $\iota(u) = j$ exists such that $y \prec u \prec z$. It is left to the reader to verify that a striking consequence of this is that either i believes j to be rational everywhere, or nowhere.

Table 3.2

Assumptions	
preferences	$\bigwedge_{i,k,l} \mathbf{u}_i(k,l) = \mathbf{r}_{i,k,l}$
principles	–
beliefs	
preferences	$\mathbf{C} \bigwedge_{i,k,l} \mathbf{u}_i(k,l) = \mathbf{r}_{i,k,l}$
principles	–
performed action	–
Inconsistency	
node x	$\neg \mathbf{C}^x \mathbf{rrat}^x$
Proof System	${}_{\Gamma}\mathbf{KD_{EC}rrat}$

To prove the inconsistency result, the proof system ${}_{\Gamma}\mathbf{KD_{EC}Prrat}$ with Prop, Dual, K, D, E, C, the linear (in)equality axioms, the Kolmogorov axioms, the interrelation axioms, the proof rules modus ponens, necessitation and induction, all axioms for many-moment game-playing situations, the rationality axiom and the two persistence axioms are used. The formalisation of the assumptions can be found in Table 3.2.

Theorem 3.3 (Reny, 1988) *There is an extensive form game with perfect information such that for all game-playing situations that consist of at least two decision* *moments there cannot be common true belief, at the second decision moment, among the players that they are rational.*

Reny's original proof involves an argument to the effect that no game-playing situation of the game shown in Figure 3.1 can have common belief about rationality at its second moment, because every second decision moment would be a moment at which at x_1 player 2 has to move. In this original game, there is no way for both players to play on and gain (only one will gain from playing on) and hence the suggestion might arise that the inconsistency result is not too surprising after all. However, the result can be proven in a case where both players would gain from playing on, too, and to underline this I prove the result using the game shown in Figure 3.2 rather than Reny's original game shown in Figure 3.1. In fact, Reny provides an argument which reveals that the class of games for which inconsistency results can be proven is fairly large.[12]

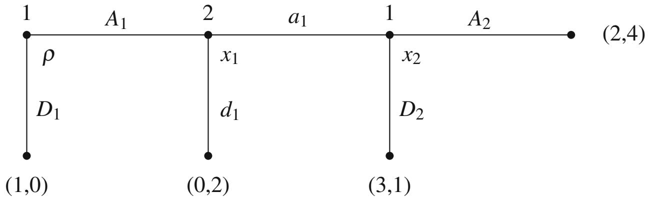

Fig. 3.2

Proof I make use of the two interrelation axioms Cons and KnProb, because all relevant beliefs in this proof all have probability one, and I use obvious notation to refer to strategies. I first prove

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \rightarrow \Box_2^{x_1}\Box_1^{\rho} d_1. \tag{3.2}$$

To do that, it suffices to show

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \rightarrow \Box_1^{\rho} d_1, \tag{3.3}$$

because a simple argument using the rule of necessitation for $\Box_2^{x_1}$ concludes the proof. Because of the StratPers-axiom, however, to prove 3.3, it suffices to show

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \rightarrow \Box_1^{x_1} d_1, \tag{3.4}$$

and to show 3.4, in turn, it is shown that

[12] 'Common Belief and the Theory of Games with Perfect Information', *Journal of Economic Theory*, 59 (1993), 269.

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to (\Box_1^{x_1}\mathbf{rrat}_2^{x_1} \wedge \Box_1^{x_1}\Box_2^{x_1}D_2), \tag{3.5}$$

and then apply necessitation for $\Box_1^{x_1}$ to an instance of the rationality axiom to get

$$(\Box_1^{x_1}\mathbf{rrat}_2^{x_1} \wedge \Box_1^{x_1}\Box_2^{x_1}D_2) \to \Box_1^{x_1}d_1.$$

The remainder of the proof of 3.2 is devoted to showing 3.5. Clearly we have

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to \Box_1^{x_1}\mathbf{rrat}_2^{x_1}.$$

To prove

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to \Box_1^{x_1}\Box_2^{x_1}D_2,$$

observe that with the RatPers-axiom for $\Box_2^{x_1}$ and necessitation for $\Box_1^{x_1}$ it can be shown that

$$\Box_1^{x_1}\Box_2^{x_1}\mathbf{rrat}_1^{x_1} \to \Box_1^{x_1}\Box_2^{x_1}\mathbf{rrat}_1^{x_2},$$

because x_2 is a successor of x_1 at which 1 moves for which in addition no y with $x_1 \succ y \succ x_2$ exists at which it is 2's turn. Hence

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to \Box_1^{x_1}\Box_2^{x_1}\mathbf{rrat}_1^{x_2}.$$

Applying the rationality axioms concludes the proof of 3.2.

Observe now that it is an easy consequence of the RatPers-axiom that

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to \Box_2^{x_1}\mathbf{rrat}_1^{\rho}, \tag{3.6}$$

and that

$$(\Box_2^{x_1}\Box_1^{\rho}d_1 \wedge \Box_2^{x_1}\mathbf{rrat}_1^{\rho}) \to \Box_2^{x_1}\neg A_1. \tag{3.7}$$

follows directly from the rationality axiom plus an appropriate application of the rule of necessitation that. All this is used to prove

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to \bot. \tag{3.8}$$

The KnWhere-axiom gives $\Box_2^{x_1}A_1$. Hence it suffices to show that

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to \neg\Box_2^{x_1}A_1.$$

Combining 3.2 and 3.6 gives

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to (\Box_2^{x_1}\Box_1^{\rho}d_1 \wedge \Box_2^{x_1}\mathbf{rrat}_1^{\rho})$$

to which application of 3.7 gives

$$\mathbf{C}^{x_1}\mathbf{rrat}^{x_1} \to \Box_2^{x_1}\neg A_1.$$

An application of the D-axiom finishes the proof.

3.2.2 Discussion

Given the game-theoretic view of rationality as expected utility maximisation, the question to ask is not so much whether the RRat-axiom is plausible, but whether the belief persistence principles embodied in StratPers and RatPers are plausible. Let me distinguish the plausibility of the principles in general, and the plausibility of the specific instances in the proof of Theorem 3.3.

Let me start with the general plausibility of the StratPers-axiom:

$$\bigwedge_i \bigwedge_j \mathbf{P}_i^x(\mathbf{j}_k^z) = \mathbf{P}_i^y(\mathbf{j}_k^z),$$

for $x \preceq y \preceq z$. A possible argument in favour of this principle runs as follows. If at x player i believes that at some $z \succeq x$ her opponent j chooses action a, say, then there is no need for i to revise her beliefs at some intermediate y (satisfying $x \preceq y \preceq z$, that is) as long as i has not received any contradictory information on the way from x to y. But information contradicting that j chooses a can only be information that j chooses, at z, an action different from a. Player i has not received such information at the intermediate y, and consequently she will not need to revise her beliefs at y. Arguably, this yields a defence of StratPers.

Yet this argument overlooks subtle ways of obtaining pertinent information. A reason for player i's belief that j chooses a at z may be her belief that at z player j chooses rationally. If, however, on the path from x to y, player i has seen j choosing irrationally, this reason is probably no longer available. Player i might revise her beliefs in such a way that at z player j plays irrationally, too—not choosing a. She need not change her beliefs, but she may change them, and that is sufficient to make StratPers problematic.

This is a general problem with the StratPers-axiom. It completely ignores the fact that the reasons players have for particular beliefs may change over time, and that consequently they have to reconsider (or revise) their beliefs, even if they are not directly contradicted by observed facts.

Similar arguments work against the general plausibility of the RatPers-axiom:

$$\bigwedge_i \bigwedge_j (\Box_i^x \mathbf{rrat}_j^y \leftrightarrow \Box_i^x \mathbf{rrat}_j^z),$$

where $x \preceq y \prec z$, $\iota(z) = j$, and no u with $\iota(u) = j$ exists such that $y \prec u \prec z$. Imagine that only irrational play on the part of j may get her from y to z. Although i believes, at x, that j will not take that irrational route, it is still questionable whether i should maintain that even though j plays irrationally, she will return to playing rationally at node z.

Now it may be that the use of the belief persistence axioms in the proof of Theorem 3.3 is harmless. StratPers is used to prove

$$\Box_1^{x_1} d_1 \rightarrow \Box_1^{\rho} d_1$$

in the proof of 3.2. The general difficulty that displays is clearly revealed. The reasons for the belief $\Box_1^{x_1} d_1$ are the beliefs $\Box_1^{x_1} \mathbf{rrat}_2^{x_1} \wedge \Box_1^{x_1} \Box_2^{x_1} D_2$. This is so because $(\Box_1^{x_1} \mathbf{rrat}_2^{x_1} \wedge \Box_1^{x_1} \Box_2^{x_1} D_2) \rightarrow \Box_1^{x_1} d_1$ is obtained by necessitation on an instance of the RRat-axiom. These reasons, perhaps available at x_1, may not be available at ρ, though. That is, it may be doubted whether $\Box_1^{\rho} \mathbf{rrat}_2^{x_1} \wedge \Box_1^{\rho} \Box_2^{x_1} D_2$. One way to substantiate doubt would concern the second conjunct. As inspection of the proof of 3.2 shows, the reasons for $\Box_1^{x_1} \Box_2^{x_1} D_2$ involve, among other things, player 1's beliefs about 2's beliefs about the rationality of player 1 at x_2, or $\Box_1^{x_1} \Box_2^{x_1} \mathbf{rrat}_1^{x_2}$. The question whether these may figure as reasons for the beliefs at the root of the game (reasons for $\Box_1^{\rho} d_1$) then boils down to the question whether these reasons were already available at the root of the game; that is, whether $\Box_1^{\rho} \Box_2^{x_1} \mathbf{rrat}_1^{x_2}$ follows from $\Box_1^{x_1} \Box_2^{x_1} \mathbf{rrat}_1^{x_2}$.

It may come as an anticlimax that there do not seem to be any serious problems here. It is about player 1 imagining (at the root and at x_1) what player 2 will or does believe at x_1 about player 1 at x_2. But player 1 will not have obtained any new information about player 2's beliefs (at x_1) while going from the root to x_1. At the root, player 1 imagines player 2's beliefs at x_1, and at x_1 player 1 again imagines player 2's beliefs at x_1. There is no difference between these cases. There would have been a difference had the statement compared player 2's beliefs at the root with his beliefs at x_1. But that is not the issue here. Consequently, StratPers causes no harm to the plausibility of the assumptions of Theorem 3.3.

To turn to the RatPers-axiom, it is first used to arrive at

$$\Box_1^{x_1} \Box_2^{x_1} \mathbf{rrat}_1^{x_1} \rightarrow \Box_1^{x_1} \Box_2^{x_1} \mathbf{rrat}_1^{x_2}$$

in the proof of 3.5. You may find this problematic as it involves the rationality of player 1 at a decision moment where she need not choose any action. But apart from that there do not seem to be reasons to doubt this line of reasoning. Player 1 does not move at x_1, so player 2, if he believes that 1 is rational at the decision moment corresponding to x_1, has no reason to say that 1 would not be rational at the possible succeeding decision moment. Player 1 believes all this, and consequently the RatPers-axiom is unproblematic here.

Yet it is also used to prove 3.6,

$$\mathbf{C}^{x_1} \mathbf{rrat}^{x_1} \rightarrow \Box_2^{x_1} \mathbf{rrat}_1^{\rho},$$

and here I can point to something dubious: a belief revision policy, forced upon player 2, that is excessively rigid. It excludes, for instance, sensible dealings with a situation of the following kind. Player 2 has actually arrived at x_1, so player 1 has moved across. While player 2 considers this to be irrational, he also believes it to be an accident or a mistake. At x_2, that is, player 2 believes that player 1 was irrational at the first decision moment, but he also believes at x_2 that player 1 is rational at the second decision moment (and perhaps even the third). This kind of subtle belief revision policy is excluded by RatPers. Either a player is believed to be rational everywhere or irrational everywhere.

In summary, the proof of Theorem 3.3 boils down to showing that there is a contradiction between having arrived at x_1 and there being common true belief about rationality at x_1. Such a contradiction can only be shown if, from the fact that there is common belief about rationality at x_1, it can be derived that one cannot be at x_1, more specifically, that one cannot be at x_1 because it can only be reached irrationally. That can only be demonstrated successfully if, from common belief about rationality at x_1, something follows about the beliefs and rationality at ρ.

But there is nothing in the concept of common belief at some decision moment that obliges me to interpret it in such a temporally extended way, and to adopt corresponding belief revision policies. In other words, there is nothing to disallow a game-playing situation in which at the first decision moment there is no common belief about rationality, while there is in the second. The RatPers axiom (together with the StratPers-axiom) exclude that possibility. This means that they are too strict. That being the case, the inconsistency result only works under very heavy, if not implausible assumptions, and shares this fate with the epistemic characterisation result it set out to criticise.