
Extensive Form Rationalizability

The heart has its reasons of which reason
knows nothing.

Blaise Pascal

The extensive form of a game is informationally richer than the normal form, since players gather information that allows them to update their subjective priors as the game progresses. For this reason, the study of rationalizability in extensive form games is more complex than the corresponding study in normal form games. There are two ways to use the added information to eliminate strategies that would not be chosen by a rational agent: backward induction and forward induction. The latter is relatively exotic (although more defensible), and will be addressed in Chapter 9. Backward induction, by far the most popular technique, employs the elimination of iterated weakly dominated strategies, arriving at the so-called *subgame perfect* Nash equilibria—the equilibria that remain Nash equilibria in all subgames. We shall call an extensive form game *generic* if it has a unique subgame perfect Nash equilibrium.

In this chapter we develop the tools of modal logic and present Robert Aumann’s famous proof (Aumann, 1995) that CKR implies backward induction. This theorem has been widely criticized, as well as widely misinterpreted. I will try to sort out the issues, which are among the most important in contemporary game theory.

5.1 Backward Induction and Dominated Strategies

Backward induction in extensive form games with *perfect information* (i.e., where each information set is a single node) operates as follows. Choose any terminal node $\tau \in T$ and find the parent node of this terminal node, say node v . Suppose player i chooses at v , and suppose i ’s highest payoff at v is attained at terminal node $\tau' \in T$. Erase all the branches from v so v becomes a terminal node, and attach the payoffs from τ' to the new terminal node v . Also record i ’s move at v , so you can specify i ’s equilibrium

strategy when you have finished the analysis. Repeat this procedure for all terminal nodes of the original game. When you are done, you will have an extensive form game that is one level less deep than the original game. Now repeat the process as many times as is possible. If the resulting game tree has just one possible move at each node, then when you reassemble the moves you have recorded for each player, you will have a Nash equilibrium.

We call this *backward induction* because we start at the terminal nodes of the game and move backward. Note that if players move at more than a single node, backward induction eliminates *weakly* dominated strategies, and hence can eliminate Nash equilibria that use weakly dominated strategies. Moreover, backward induction is *prima facie* much stronger than normal form rationalizability (§4.6), which is equivalent to the iterated elimination of strongly dominated strategies.

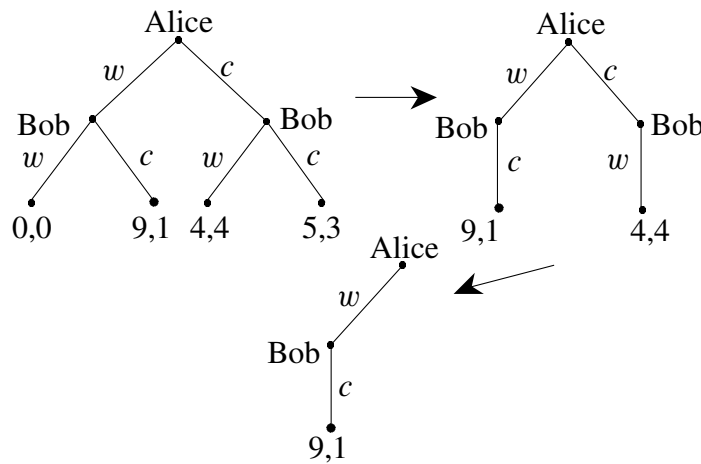


Figure 5.1. An Example of Backward Induction

For an example of backward induction, consider Fig. 5.1. We start with the terminal node labeled (0,0), and follow it back to the Bob node on the left. At this node, w is dominated by c because $1 > 0$, so we erase the branch where Bob plays w and its associated payoff. We locate the next terminal node in the original game tree, (4,4), and follow back to the Bob node on the right. At this node, c is dominated by w , so we erase the dominated node and its payoff. Now we apply backward induction to this smaller game tree—this time of course it's trivial. We find the first terminal node, (9,1), which leads back to Alice's choice node. Here c is dominated, so we erase that branch and its payoff. We now have our solution: Alice chooses w , Bob chooses cw , and the payoffs are (9,1).

It is clear from this example that by using backward induction and hence eliminating weakly dominated strategies, we have eliminated the Nash equilibrium c, ww . This is because when we assume Bob plays c in response to Alice's w , we have eliminated the weakly dominated strategies ww and wc for Bob. We call c, ww an *incredible threat*. Backward induction eliminates incredible threats.

5.2 Subgame Perfection

Let ν be an information set of an extensive form game \mathcal{G} that consists of a single node. Let \mathcal{H} be the smallest collection of nodes including ν such that if h' is in \mathcal{H} , then all of the successor nodes of h' are in \mathcal{H} and all nodes in the same information set as h' are in \mathcal{H} . We endow \mathcal{H} with the information set structure, branches, and payoffs inherited from \mathcal{G} , the players in \mathcal{H} being the subset of players of \mathcal{G} who move at some information set of \mathcal{H} . It is clear that \mathcal{H} is an extensive form game. We call \mathcal{H} a *subgame* of \mathcal{G} .

If \mathcal{H} is a subgame of \mathcal{G} with root node ν , then every pure strategy profile of \mathcal{G} that reaches ν has a counterpart s_H in \mathcal{H} , specifying that players in \mathcal{H} make the same choices with s_H at a node in \mathcal{H} as they do with s_G at the same node in \mathcal{G} . We call s_H the *restriction* of s_G to the subgame \mathcal{H} . Suppose $\sigma_G = \alpha_1 s_1 + \dots + \alpha_k s_k$ ($\sum_i \alpha_i = 1$) is a mixed strategy of \mathcal{G} that reaches the root node ν of \mathcal{H} , and let $I \subseteq \{1, \dots, k\}$ be the set of indices such that $i \in I$ iff s_i reaches h . Let $\alpha = \sum_{i \in I} \alpha_i$. Then, $\sigma_H = \sum_{i \in I} (\alpha_i / \alpha) s_i$ is a mixed strategy defined on \mathcal{H} , called the *restriction* of σ_G to \mathcal{H} . We have $\alpha > 0$ because σ_G reaches ν , and the coefficient α_i / α represents the probability of playing s_i , conditional on reaching h .

It is clear that if s_G is a pure strategy Nash equilibrium for a game \mathcal{G} , and if \mathcal{H} is a subgame of \mathcal{G} whose root node is reached using s_G , then the restriction s_H of s_G to \mathcal{H} must be a Nash equilibrium in \mathcal{H} . However, if the root node of \mathcal{H} is not reached by s_G , then the restriction of s_G to \mathcal{H} need *not* be a Nash equilibrium. This is because if a node is not reached by s_G , then the payoff to the player choosing at that node does not depend on his choice in \mathcal{G} , but it may depend on his choice in \mathcal{H} . We say a Nash equilibrium of an extensive form game is *subgame perfect* if its restriction to every subgame is a Nash equilibrium of the subgame.

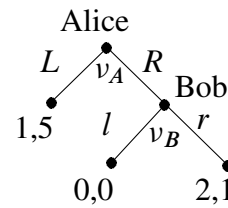
It is easy to see that a simultaneous move game has no proper subgames (a game is always a subgame of itself; we call the whole game an *improper* subgame), because all the nodes are in the same information set for at least

one player. Similarly, a game in which Nature makes the first move also has no proper subgames if there is at least one player who does not know Nature's choice.

At the other extreme, in a game of perfect information (i.e., for which all information sets are singletons), every nonterminal node is the root node of a subgame. This allows us to find the subgame perfect Nash equilibria of such games by backward induction, as described in §5.1. This line of reasoning shows that in general, backward induction consists of the iterated elimination of weakly dominated strategies and eliminates all non-subgame perfect Nash equilibria.

5.3 Subgame Perfection and Incredible Threats

The game to the right has a Nash equilibrium in pure strategies Rr , in which Alice earns 2 and Bob earns 1. This equilibrium is subgame perfect, because in the subgame starting with Bob's choice at v_B , r is payoff-maximizing for Bob. This equilibrium is also the one chosen by backward induction. However, there is a second Nash equilibrium, Ll , in which Alice earns 1 and Bob earns 5. Bob much prefers this equilibrium, and if he can somehow induce Alice to believe that he will play l , her best response is L . However Bob communicates to Alice his intention to play l , if Alice believes Bob is rational, she knows he in fact will play r if the game actually reaches v_B . Thus, Ll is thought to be an implausible Nash equilibrium, whereas the subgame perfect Nash equilibrium is held in high regard by game theorists.



5.4 The Surprise Examination

One summer a group of game theorists took an intensive, Monday through Friday, logic course. After several weeks, the professor announced that there would be a surprise examination one day the following week. Each student thought to himself “the exam cannot be given next Friday, because then it would not be a surprise.” Each then concluded that, for similar reasons, the exam could not be given next Thursday, next Wednesday, next Tuesday, or next Monday. Each student thus concluded that the professor was mistaken. The professor gave the exam the next Tuesday, and all of the students were surprised.

This is one version of a famous logic problem called The Surprise Examination or The Hanging Paradox. For an overview of the many proposed solutions to the problem, see Chow (1998). Interpretations vary widely, and there is no single accepted solution. There are a number of cogent analyses using standard logic and modal logic to show that the professor's statement is impermissively self-referential or self-contradictory, and because a false statement can validly imply anything, there is no paradox in the professor's prediction being correct.

Backward induction indicates that the exam cannot be given. But, if a student believes this, then it will be a surprise no matter what day it is given. Thus, the incoherence of backward induction should convince a rational student that the professor's prediction is indeed reasonable. But, what exactly is "incoherent" about backward induction? I present this paradox to indicate the danger of using the informal logic of backward induction. We develop a more analytically precise approach below.

5.5 The Common Knowledge of Logicality Paradox

Let us say an agent is *logical* in making inferences concerning a set of propositions if the agent rules out all statements that are inconsistent with this set. We then define *common knowledge of logicality* (CKL) for a set $i = 1, \dots, n$ of agents in the usual way: for any set of integers $i_1, \dots, i_k \in [1, \dots, n]$, i_1 knows that i_2 knows that ... knows that i_{k-1} knows that i_k is logical.

A father has \$690,000 to leave to his children, Alice and Bob, who do not know the size of his estate. He decides to give one child \$340,000 and the other \$350,000, each with probability 1/2. However, he does not want one child to feel slighted by getting a smaller amount, at least during his lifetime. So, he tells his children: "I will randomly pick two numbers, without replacement, from a set $S \subseteq [1, \dots, 100]$, assign to each of you randomly one of these numbers, and give you an inheritance equal to \$10,000 times the number you have been assigned. Knowing the number assigned to you will not allow you to conclude for sure whether you will inherit more or less than your sibling." The father, confident of the truth of his statement, which we take to be common knowledge for all three individuals, sets $S = \{34, 35\}$.

Alice ponders this situation, reasoning as follows, assuming common knowledge of logicality. Father knows that if $1 \in S$ or $100 \in S$, then

there is a positive probability one of these numbers will be chosen and assigned to me, in which case I would be certain of the relative position of my inheritance. Alice knows the father knows she is logical, so she knows that $1 \notin S$ and $100 \notin S$. But, Alice reasons that her father knows that she knows that he knows she is logical, so she concludes that the father knows that he cannot include 2 or 99 in S . But, Alice know this as well, by CKL, so she reasons that the father cannot include 3 or 98 in S . Completing this recursive argument, Alice concludes that S must be empty.

However, the father gave one child the number 34, and the other 35, neither child knowing for sure which has higher number. Thus, the father's original assertion was true, and Alice's reasoning was faulty. We conclude that *common knowledge of logicity is false* in this context. CKL fails when the father included 35 in S , because this is precluded by CKL.

CKL appears *prima facie* to be an innocuous extension of logicity, and indeed is not usually even mentioned in such problems, but in fact it leads to faulty reasoning and must be rejected. In this regard, CKL is much like CKR, which also appears to be an innocuous extension of rationality, but in fact is often counterindicated.

5.6 The Repeated Prisoner's Dilemma

Suppose Alice and Bob play the Prisoner's Dilemma, one stage of which is shown to the right, 100 times. Common sense tells us that players will cooperate for at least 95 rounds, and this is indeed supported by experimental evidence (Andreoni and Miller 1993).

	C	D
C	3,3	0,4
D	4,0	1,1

However, a backward induction argument indicates that players will defect on the very first round. To see this, note that the players will surely defect on round 100. But then, nothing they do on round 99 can help prolong the game, so they will both defect on round 99. Repeating this argument 99 times, we see that they will both defect on round 1.

Although in general backward induction removes weakly iterated dominated strategies, in this case, it removes only strongly iteratedly dominated strategies, so the only rationalizable strategy, according to the analysis of the previous chapter, is the universal defect Nash equilibrium. This presents a problem for the rationalizability concept that is at least as formidable as in the case of the normal form games presented in the previous chapter.

In this case, however, the extensive form provides an argument as to why the logic of backward induction is compromised. The backward induction

reasoning depends on CKR in precisely the same manner as in the previous chapter. However, in the current case, the first time either player chooses C, both players know that CKR is false. At the terminal nodes of the Repeated Prisoner's Dilemma, players have chosen C many times. Therefore, we cannot assume CKR at the terminal nodes, because these nodes could not be reached, given CKR. This critique of backward induction has been made by Binmore (1987), Bicchieri (1989), Pettit and Sugden (1989), Basu (1990), and Reny (1993), among others.

The critique, however, is incorrect. The backward induction argument is simply a classic example of *reductio ad absurdum*: assume a proposition and then show that the proposition is false. In this case, we assume CKR and we show by *reductio* that the 100th round will not be reached. There is no flaw in this argument. It is incoherent to base a critique of the proposition that CKR implies backward induction on what would happen if CKR were false.

The misleading attractiveness of this flawed critique of the proposition that CKR implies backward induction lies in the observation that the first time either player chooses C, both players know that CKR is false, and hence they are free to devise a *modus operandi* that serves their interests for the remainder of the game. For instance, both may employ the tit-for-tat strategy of playing C on one round, and copying one's partner's previous move on each subsequent round, except for playing universal D as the game nears the 100 round termination point.

This argument is completely correct, but is not a critique of the proposition that CKR implies backward induction. Indeed, assuming CKR, neither player will choose C in any period.

As I shall argue below, the problem with backward induction is that CKR is not generally a permissible assumption, and hence backward induction cannot be justified on rationality grounds.

5.7 The Centipede Game

In Rosenthal's centipede game, Alice and Bob, start out with \$2 each, and they alternate rounds. On the first round, Alice can defect (D) by stealing \$2 from Bob, and the game is over. Otherwise, Alice cooperates (C) by not stealing, and Nature gives Alice \$1. Then Bob can defect (D) and steal \$2 from Alice, and the game is over, or he can cooperate (C), and Nature gives

Bob \$1. This continues until one or the other defects, or 100 rounds have elapsed. The game tree is illustrated in Figure 5.2.

Formally, the reduced normal form of the centipede game (§5.7) can be described as follows. Alice chooses an odd number k_a between 1 and 101 and Bob chooses an even number k_b between 2 and 100, plus either C or D if $k_b = 100$. The lower of the two choices, say k^* , determines the payoffs. If $k^* = 100$, the payoff to $(k^*, C) = (52, 52)$ and the payoff to $(k^*, D) = (50, 53)$. Otherwise, if k^* is an odd number, the payoffs are $(4 + (k^* - 1)/2, (k^* - 1)/2)$, and if k^* is even, the payoffs are $(k^*/2, 3 + k^*/2)$. You can check that these generate exactly the payoffs as described above.

To determine the strategies in this game, note that Alice and Bob each has 50 places to move, and in each place each can play D or C. We can thus describe a strategy for each as a sequence of 50 letters, each of which is a D or C. This means there are $2^{50} = 1,125,899,906,842,624$ pure strategies for each. Of course, the first time a player plays D, what he does after that does not affect the payoff of the game, so the only payoff-relevant question is at what round, if any, a player first plays D. This leaves 51 strategies for each player.

We can apply backward induction to the game, finding that in the unique subgame perfect Nash equilibrium of this game, both players defect the first time they get to choose. To see this, note that on the final round, Bob will defect, and hence Alice will defect on her last move. But then Bob will defect on his next-to-last move, as will Alice on her next-to-last move. Similar reasoning holds for all rounds, proving that Bob and Alice will defect on their first move in the unique subgame perfect equilibrium.

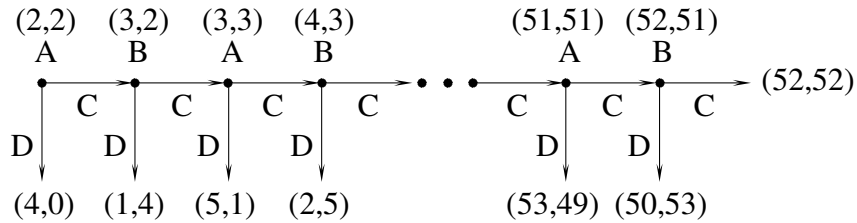


Figure 5.2. Rosenthal's Centipede Game

Now, of course, common sense tells you that this is not the way real players would act in this situation, and empirical evidence corroborates this intuition (McKelvey and Palfrey 1992). It may seem that the culprit is subgame perfection, because backward induction only finds subgame perfect

equilibria. This, however, is not the problem. The culprit is the Nash equilibrium criterion itself, because in *any* Nash equilibrium, Alice defects on round 1.

While backward induction does not capture how people really play the Centipede Game, normal form rationalizability does a better job, because it suggests that cooperating until near the end of the game does not conflict with CKR. This is because all pure strategies for Bob are normal-form rationalizable, except $k_b = (100, C)$, as are all pure strategies for Alice except $k_a = 101$ (i.e., cooperate in every round). To see this, we can show that there is a mixed strategy Nash equilibrium of the game where Bob uses $k_b = 2$ and any one of his other pure strategies except $(100, C)$, and Alice uses $k_a = 1$. This shows that all pure strategies for Bob except $(100, C)$ are rationalizable. Alice's $k_a = 101$ is strictly dominated by a mixed strategy using $k_a = 99$ and $k_a = 1$, but each of her other pure strategies is a best response to some rationalizable pure strategy of Bob. This shows that these pure strategies of Alice are themselves rationalizable.

This does not explain why real people cooperate until near the end of the game, but it does show that it does not conflict with CKR in the normal form game to do so. This is little consolation, however, since cooperating becomes compatible with CKR only by ignoring information that the players surely have—namely, that embodied in the extensive form structure of the game. CKR in the context of this additional information certainly does imply the validity of the backward induction argument, and hence of the assertion that CKR assures defection on round 1.

5.8 CKR Fails off the Backward Induction Path

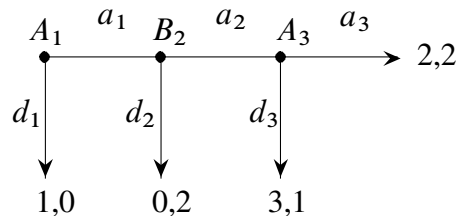


Figure 5.3. A Short Centipede Game

This section presents a formal epistemic argument supporting the contention that CKR is violated off the subgame perfect game path in a generic extensive form game. This is the thrust of Aumann's (1995) general proof that CKR implies backward induction, but we here present the proof for a

very simple game, where the intuition is relatively clear. Figure 5.3 depicts a very short centipede game (§5.7) played by Alice (A) and Bob (B), where Alice moves at A_1 and A_3 , and Bob moves at B_2 . Let R_A and R_B stand for “Alice is rational,” and “Bob is rational,” let \mathbf{K}_A and \mathbf{K}_B be knowledge operators, and let π_A and π_B be the payoffs for of Alice and Bob, respectively. We will have much more to say in later chapters concerning what it means to be “rational,” and what it means to assert that an agent “knows something.” For now, we simply assume that a rational agent always chooses a best response, we assume $\mathbf{K}p \Rightarrow p$; i.e., if an agent knows something, then it is true, and we assume $\mathbf{K}p \wedge \mathbf{K}(p \Rightarrow q) \Rightarrow \mathbf{K}q$; i.e., if an agent knows p and also knows that p implies q , then the agent knows q . We assume that both players know all the rules of the game, and since the game is of perfect information, when the game is at a particular node, it is common knowledge that this is the case.

We have $A_3 \wedge R_A \Rightarrow d_3$, which we read “at node A_3 , if Alice is rational then she will choose d_3 .” This is true because $a_3 \Rightarrow (\pi_A = 2)$, $d_3 \Rightarrow (\pi_A = 3)$, and since these implications simply follow from the rules of the game, they are known to Alice, so $\mathbf{K}_A(a_3 \Rightarrow (\pi_A = 2)) \wedge \mathbf{K}_A(d_3 \Rightarrow (\pi_A = 3))$. This assertion implies $A_3 \wedge R_A \Rightarrow d_3$. Now, if Bob knows that Alice is rational, and if the rules of the game are common knowledge, then $a_2 \Rightarrow \mathbf{K}_B a_2 \Rightarrow \mathbf{K}_B d_3 \Rightarrow \mathbf{K}_B(\pi_B = 1)$. Moreover, $d_2 \Rightarrow \mathbf{K}_B d_2 \Rightarrow \mathbf{K}_B(\pi_B = 3)$, so $B_2 \wedge R_B \wedge \mathbf{K}_B R_A \Rightarrow d_2$. Now, if Alice knows that Bob is rational at A_1 , and she knows that Bob knows that she is rational at B_2 , then $\mathbf{K}_A(a_1 \Rightarrow d_2)$, so $\mathbf{K}_A(a_1 \Rightarrow (\pi_A = 0))$. However $\mathbf{K}_A(d_1 \Rightarrow (\pi_A = 1))$. Hence, since Alice is rational at A_1 , she will choose d_1 . In short, we have

$$R_A \wedge \mathbf{K}_A(R_B \wedge \mathbf{K}_B R_A) \Rightarrow d_1. \quad (5.1)$$

We have thus shown that if there are two levels of mutual knowledge of rationality, then the backward induction solution holds. But, this presupposes that the set of assumptions is *consistent*; i.e., it assume that we cannot also prove, from these assumptions, that Alice will play a_1 . Note that if Alice plays a_1 , then the premise of (5.1) is false, and Bob knows this, which says

$$\neg \mathbf{K}_B R_A \vee \neg \mathbf{K}_B \mathbf{K}_A R_B \vee \neg \mathbf{K}_B \mathbf{K}_A \mathbf{K}_B R_A. \quad (5.2)$$

In words, if Alice choose a_1 , then either Bob does not know Alice is rational, or Bob does not know that Alice knows that Bob is rational, or Bob does not know that Alice knows that Bob knows that Alice is rational. One level of mutual knowledge, which implies $\mathbf{K}_B R_A$, eliminates the first alternative,

and two levels, which implies $\mathbf{K}_B \mathbf{K}_A R_B$, eliminates the second alternative. Thus, if Alice chooses a_1 , it must be the case that $\neg \mathbf{K}_B \mathbf{K}_A \mathbf{K}_B R_A$; i.e., Alice's choice violates third level mutual knowledge of rationality, and hence violates common knowledge of rationality.

We conclude no node after the first can be attained while conserving more than two levels of mutual knowledge of rationality. Nor is there anything special about this game. As we expressed in the previous section, and will prove in §5.11, in all finite extensive form games of perfect information with unique subgame perfect equilibria, the only nodes in the game at which common knowledge of rationality can hold are along the backward induction path of play. In the current case, there are just two such nodes, the root node and the first terminal node.

5.9 How to Play the Repeated Prisoner's Dilemma

In cases where a stage game is repeated a finite number of times, it is reasonable to assume Bayesian rationality (§1.5), avoid backward induction, and use decision theory to determine player behavior. Consider, for instance, a Prisoner's Dilemma (§2.10), the stage game of which is to right with $T > R > P > S$, repeated until one player defects or 100 rounds have been played. Backward induction implies that both players will defect on the very first round, and indeed, this is the only Nash equilibrium of the game. However, player 1 may say to himself, "if my partner and I both play D , we will each earn only P . I am willing to cooperate for at least 95 rounds, and if my partner is smart, he will also be willing to cooperate for many rounds. I suspect my partner will reason similarly. Thus we stand to earn on the order of $95R$. If I am wrong about my partner, I will lose only $S - P$, so it's worth a try, because if I am right, I will go home with a tidy bundle."

More formally, suppose I conjecture that my partner will cooperate up to round k , and then defect, with probability g_k . Then, I will choose a round m to defect that maximizes the expression

$$\pi_m = \sum_{i=1}^{m-1} ((i-1)R + S)g_i + ((m-1)R + P)g_m + ((m-1)R + T)(1 - G_m), \quad (5.3)$$

	C	D
C	R,R	S,T
D	T,S	P,P

where $G_m = g_1 + \dots + g_m$. The first term in this expression represents the payoff if my partner defects first, the second if we defect simultaneously, and the final term if I defect first. In many cases, maximizing this expression will suggest cooperating for many rounds for all plausible probability distributions. For instance, suppose g_k is uniformly distributed on the rounds $m = 1, \dots, 99$. Suppose, for concreteness, $(T, R, P, S) = (4, 2, 1, 0)$. Then, you can check by using equation (5.3) that it is a best response to cooperate up to round 98. Indeed, suppose you expect your opponent to defect on round 1 with probability 0.95, and otherwise defect with equal probability on any round from 2 to 99. Then it is still optimal to defect on round 98. Clearly the backward induction assumption is not plausible unless you think your opponent is highly likely to be an obdurate backward inductor.

The reasoning dilemma begins if I then say to myself “My partner is just as capable as I of reasoning as above, so he will also cooperate at least up to round 98. Thus, I should set $m = 97$. But, of course my partner also knows this, so he will surely defect on round 96, in which case I should surely defect on round 95.” This sort of self-contradictory reasoning shows that there is something faulty in the way we have set up the problem. If the $\{g_k\}$ distribution is reasonable, then I should use it. It is self-contradictory to use this distribution to show that it is the wrong distribution to use. But, my rational partner will know this as well, and I suspect he will revert to the first level of analysis, which says to cooperate at least up to round 95. Thus we two rational folks will cooperate for many rounds in this game rather than play the Nash equilibrium.

Suppose, however, that it is common knowledge that both I and my partner have the *same Bayesian prior* (§1.5) concerning when the other will defect. This is sometimes called *Harsanyi consistency* (Harsanyi 1967). Then, it is obvious that we will both defect on our first opportunity, because the backward induction conclusion now follows from a strictly Bayesian argument: the only prior that is compatible with common knowledge of common priors is defection on round one. However, there is no plausible reason for us to assume Harsanyi consistency in this case.

This argument reinforces our conclusion that *there is nothing sacrosanct about CKR*. Classical game theorists commonly argue that rationality *requires* that agents use backward induction, but this is simply not the case. If two players are rational and they know both are rational, and if each knows the other’s conjecture, then they will play the unique Nash equilib-

rium of the game (§8.4). But, as we have seen, we may each reasonably conclude that we do not know the other's conjecture, but we know enough to cooperate for many rounds.

5.10 The Modal Logic of Knowledge

The Savage model in decision theory is agnostic as to how an agent's subjective prior is acquired. How people play a game depends on their beliefs about the beliefs of the other players, including their beliefs about others' beliefs about the players, and so on. To deal analytically with this situation, we develop a formal model of what it means to say that an individual "knows" a fact about the world.

The states of nature consist of a finite universe Ω of possible worlds, subsets of which are called an *events*. Event E occurs at state ω if $\omega \in E$. When Alice is in state ω , she knows only that she is in a subset $\mathbf{P}_A\omega \subseteq \Omega$ of states; i.e., $\mathbf{P}_A\omega$ is the set of states Alice considers possible when the actual state is ω . We say that Alice *knows* the event E at state ω if $\mathbf{P}_A\omega \subseteq E$ because for every state ω' that Alice knows is possible, $\omega' \in E$.

Given a possibility operator \mathbf{P} , we define a corresponding *knowledge operator* \mathbf{K} by $\mathbf{K}E = \{\omega | \mathbf{P}\omega \subseteq E\}$. It is easy to check that the knowledge satisfies the following properties:

- | | | |
|-------|---|------------------------|
| (K1) | $\mathbf{K}\Omega = \Omega$ | omniscience |
| (K2a) | $\mathbf{K}(E \cap F) = \mathbf{K}E \cap \mathbf{K}F$ | |
| (K2b) | $E \subseteq F \Rightarrow \mathbf{K}E \subseteq \mathbf{K}F$ | |
| (K3) | $\mathbf{K}E \subseteq E$ | knowledge |
| (K4) | $\mathbf{K}E = \mathbf{K}\mathbf{K}E$ | transparency |
| (K5) | $\neg\mathbf{K}\neg\mathbf{K}E \subseteq \mathbf{K}E$ | negative introspection |

where " \neg ," means "not"; i.e., logical negation. Note that K2a implies K2b. To see this, assume K2a and $E \subseteq F$. Then, $\mathbf{K}E = \mathbf{K}(E \cap F) = \mathbf{K}E \cap \mathbf{K}F$, so $\mathbf{K}E \subseteq \mathbf{K}F$, which proves K2b. Property K3, often called the *axiom of knowledge*, asserts that what is known must be true (if we drop this principle, we get a model of *belief* rather than *knowledge*), and follows directly from P1. Property K4, called the *axiom of transparency*, says that if you know something, then you know that you know it. Property K5 says that if you do not know something, then you know that you do not know it. This is not a very intuitive statement, but it allows us to specify the properties of the knowledge operator *syntactically* without regard to its *semantic* interpretation in terms of possible worlds and the possibility

operator $\mathbf{P}\omega$. We show that K5 follows from P1 and P2 by extending the definition of the \mathbf{P} operator from states to events by $\mathbf{P}E = \bigcup_{\omega \in E} \mathbf{P}\omega$, then the knowledge and possibilities operators are *dual* in the sense that $\neg\mathbf{K}\neg E = \mathbf{P}E$ and $\neg\mathbf{P}\neg E = \mathbf{K}E$ for any event $E \subseteq \Omega$. To see the first,

$$\begin{aligned} \neg\mathbf{K}\neg E &= \{\omega | \mathbf{P}\omega \not\subseteq \neg E\} = \{\omega | \mathbf{P}\omega \cap E \neq \emptyset\} \\ &= \{\omega' | \omega' \in \bigcup_{\omega \in E} \mathbf{P}\omega\} = \mathbf{P}E. \end{aligned}$$

To see the second, suppose $\omega \in \neg\mathbf{P}\neg E$. We must show $\mathbf{P}\omega \subseteq E$. If this is false, then $\mathbf{P}\omega \cap \neg E \neq \emptyset$, which implies $\omega \in \mathbf{P}\neg E$, which is a contradiction. To prove K5, which can be written $\mathbf{P}\mathbf{K}E \subseteq \mathbf{K}E$, suppose $\omega \in \mathbf{P}\mathbf{K}E$. Then $\mathbf{P}\omega \subseteq \mathbf{K}E$, so $\omega \in \mathbf{K}E$. This argument can be reversed to prove equality.

As we saw in §4.1, we can recover the possibility operator $\mathbf{P}\omega$ for an individual from his knowledge operator \mathbf{K} , because

$$\mathbf{P}\omega = \bigcap \{E | \omega \in \mathbf{K}E\}. \quad (5.4)$$

To verify this equation, note that if $\omega \in \mathbf{K}E$, then $\mathbf{P}\omega \subseteq E$, so the left hand side of (5.4) is contained in the right hand side. Moreover, if ω' is not in the right hand side, then $\omega' \notin E$ for some E with $\omega \in \mathbf{K}E$, so $\mathbf{P}\omega \subseteq E$, so $\omega' \notin \mathbf{P}\omega$. Thus right hand side of (5.4) is contained in the left.

We say an event E is *self-evident* to an agent if he knows E at each state $\omega \in E$. Thus E is self-evident exactly when $\mathbf{K}E = E$, which means $\mathbf{P}\omega \subseteq E$ for every $\omega \in E$. Clearly, Ω itself is self-evident, and if E and F are self-evident, then $E \cap F$ is self-evident. Thus, for each state ω , $\mathbf{P}\omega$ is the *minimal* self-evident event containing ω . Every self-evident event is the union of minimal self-evident events. The minimal self-evident events coincide with the cells of the partition \mathcal{P} .

5.11 Backward Induction and Extensive Form CKR

In this section we show how a little modal logic can clarify issues concerning rational behavior and choice. We take the case of backward induction, presenting Aumann's (1995) proof that common knowledge of rationality in generic extensive form games of perfect information is possible only at nodes of the game tree that lie along the backward induction path of play.

Consider a finite generic extensive form epistemic game of perfect information \mathcal{G} (a game is generic if, for each player, no two payoffs at terminal nodes are equal). A pure strategy profile s assigns an action s^v at each non-terminal node v . Indeed, if s_i is the pure strategy profile of player i , and if

i moves at v , then $s^v = s_i^v$. We denote by b the unique backward induction strategy profile. Thus, if player i moves at node v , then

$$\pi_i^v(b) > \pi_i^v(b/a^v) \quad \text{for } a^v \neq b^v, \quad (5.5)$$

where $\pi_i^v(s)$ is the payoff of strategy profile s to player i , starting from node v (even if, starting from the beginning of the game, v would not be reached), and s/t^v denotes the strategy profile s with the player who chooses at v replacing his action s_i^v with action t at v .

To specify Bayesian rationality in this framework, suppose players choose pure strategy profile $\mathbf{s}(\omega)$ in state ω . We then say i is Bayesian rational if, for every node v at which i chooses, and for every pure strategy $t_i \in S_i$, we have

$$R_i \subseteq \neg \mathbf{K}_i \{ \omega \in \Omega \mid \pi_i^v(\mathbf{s}/t_i) > \pi_i^v(\mathbf{s}) \}; \quad (5.6)$$

i.e., i does not know that there is a better strategy than $\mathbf{s}_i(\omega)$ at v . Common knowledge of rationality, which we write as CKR, means R_i is common knowledge for all players i . Note that this definition is somewhat weaker than Bayesian rationality, which requires that agents have subjective priors over events, and maximize utility subject to this prior.

Let $I^v \subseteq \Omega$ be the event that b^v is chosen at node v . Thus

$$I^v = \{ \omega \in \Omega \mid \mathbf{s}(\omega)^v = b^v \}, \quad (5.7)$$

so the event I that the backward induction path is chosen is simply

$$I = \bigcap_v I^v.$$

The assertion that common knowledge of rationality implies backward induction is then simply

THEOREM 5.1 $CKR \subseteq I$.

Proof: We first show that at every terminal node v , $CKR \subseteq I^v$. We have

$$\begin{aligned} CKR &\subseteq R_i \subseteq \neg \mathbf{K}_i \{ \omega \in \Omega \mid \pi_i^v(\mathbf{s}/b_i) > \pi_i^v(\mathbf{s}) \} \\ &= \neg \mathbf{K}_i \{ \omega \in \Omega \mid \pi_i^v(b) > \pi_i^v(b/\mathbf{s}_i) \} \\ &= \neg \mathbf{K}_i \{ \omega \in \Omega \mid \mathbf{s}_i^v \neq b^v \} = \neg \mathbf{K}_i \neg I^v. \end{aligned}$$

The first line follows from (5.6) with $t_i = b_i$. The second line follows from the fact that v is a terminal node at which i chooses, so $\pi_i^v(b) = \pi_i^v(\mathbf{s}/b_i)$ and $\pi_i^v(\mathbf{s}) = \pi_i^v(b/\mathbf{s}_i)$.

Because i chooses at v , I^v is a union of cells of i 's knowledge partition, so I^v is self-evident to i , and hence $I^v = \mathbf{K}_i I^v$. Thus we have $\text{CKR} \subseteq \neg \mathbf{K}_i \neg \mathbf{K}_i I^v$. By negative introspection (K5), this implies $\text{CKR} \subseteq \neg \neg \mathbf{K}_i I^v = \mathbf{K}_i I^v = I^v$.

This argument proves that at every state compatible with CKR, players must make the backward induction move at each terminal node. Note that this argument does not commit what we called in chapter 5 the “fallacy of backward induction,” because this argument does not assume that a terminal node is reached that could not be reached if players used backward induction. Indeed, this argument does not assume anything about what nodes are reached or not reached.

The rest of the proof proceeds by mathematical induction. Assume that $\text{CKR} \subseteq I^w$ for all nodes w that follow a node v , where player i chooses. We can then write $\text{CKR} \subseteq \mathbf{K}_i I^{>v} = \bigcap_{w>v} \mathbf{K}_i I^w$, where $w > v$ means that node w follows node v in the game tree. We then have

$$R_i \subseteq \neg \mathbf{K}_i \{\omega \in \Omega \mid \pi_i^v(\mathbf{s}/b_i) > \pi_i^v(\mathbf{s})\},$$

by (5.6) with $t_i = b_i$. Now $\pi_i^v(\mathbf{s})$ depends only on s^v and $s^{>v}$, the restriction of s to the nodes following v . Thus we can write

$$R_i \cap I^{>v} \subseteq \neg \mathbf{K}_i \{\omega \in \Omega \mid \pi_i^v(b) > \pi_i^v(b/s^v)\} \cap I^{>v} = \neg \mathbf{K}_i \neg I^v \cap I^{>v},$$

where the first inclusion follows from the fact that for $\omega \in I^{>v}$, $\pi_i^v(\mathbf{s}/b_i) = \pi_i^v(b)$ and $\pi_i^v(\mathbf{s}) = \pi_i^v(b/s^v)$. Thus,

$$\text{CKR} \subseteq R_i \cap I^{>v} \subseteq \neg \mathbf{K}_i \neg I^v \cap I^{>v} \subseteq I^v \cap I^{>v},$$

where the argument for the final inclusion is as before. ■

This theorem does not claim that rational agents will always play the subgame perfect equilibrium. Rather, it claims that if a player makes a move to a node that is not along the backward induction path of play, then common knowledge of rationality cannot obtain at that node or any subsequent node of the game tree. There is nothing irrational about a player making such a move, as he may have some notion as to how rational agents will play the game based on considerations other than CKR.

Another way of saying this is that CKR is a condition, not a premise (§4.14). In some cases CKR will hold, but not because CKR implies the outcome, but rather the outcome implies CKR.

5.12 Rationality and Extensive Form CKR

Between 1987 and 1993, several influential papers questioned the classical game theoretic argument that in an extensive form game of perfect information with a single subgame perfect Nash equilibrium, rational agents must play this equilibrium. Aumann (1995) was seen by many game theorists as a futile and inadequate response to these critics in defense of the conventional wisdom. The central criticism of Aumann's analysis was stated as follows by Binmore (1996):

What keeps a rational player on the equilibrium path is his evaluation of what would happen if he were to deviate. But, if he were to deviate, he would behave irrationally. Other players would then be foolish if they were not to take this evidence of irrationality into account in planning their responses to the deviation... Aumann... is insistent that his conclusions say nothing whatever about what players would do if vertices of the game tree off the backward induction path were to be reached. But, if nothing can be said about what would happen off the backward-induction path, then it seems obvious that nothing can be said about the rationality of remaining on the backward-induction path." (p. 135).

Similarly Ben-Porath (1997) asserts that

Aumann assumes that in every vertex x there is common knowledge that a player will play rationally in the subgame that starts at x . This is assumed even for vertices x that cannot be reached if there is CKR at the beginning. Thus, the assumption is that a player i will ignore the fact that another player j behaved in a way which is consistent with CKR. (p.43)

One correction is clearly in order. If a rational player were to deviate from the equilibrium path, says Binmore, "he would behave irrationally." The correct statement is that if a player deviated from the equilibrium path, he would violate CKR, not rationality. That said, although there may be versions of CKR that are vulnerable to this critique, Aumann's version, presented in §5.11, is not.

This argument should not be seen, however, as a defense of CKR. Aumann himself, in all his writings, states clearly that CKR is not dictated by the norms of social interaction among rational agents. CKR is not a

strengthening of Bayesian rationality. Rather, CKR is powerful and often highly implausible assumption concerning the communality of mental representations across Bayesian rational agents.

A major attraction of epistemic game theory lies in its allowing us to replace arguments about where a proposition ψ is true and false in a game by an analysis of the set of states $\omega \in \Omega$ in which $\psi(\omega)$ holds. Thus, the conclusion of Aumann's argument, equation 5.1, must be read as "in every state ω in which CKR holds, the backward induction path is chosen by $s(\omega)$." Similarly, "CKR fails off the backward induction path" should be read "in every state ω for which $s(\omega)$ is not the backward induction path, CKR fails."

The bottom line is that the critics, from Binmore (1987) to Reny (1993) are correct in stating that rationality does not imply backward induction. But Aumann (1995) is also correct in stating that in every state that CKR holds, the backward induction path is followed.

5.13 On the Nonexistence of CKR

The proof that CKR implies backward induction in Aumann (1995) is followed by the proof of the following

THEOREM 5.2 *In every game of perfect information, there is a knowledge system such that $CKR \neq \emptyset$.*

The proof is trivial. Assume Ω has exactly one state, in which each agent's strategy is the backward induction strategy.

The more interesting question, however, is: what are the characteristics of knowledge systems for which $CKR \neq \emptyset$, and are there plausible knowledge systems for which $\Omega = CKR$? The answers to these questions are, to my knowledge, unknown.

It is easy, however, to construct a realistic epistemic game in which $CKR = \emptyset$. For instance, consider the situation described in §5.9. The game is the 100-round Repeated Prisoner's Dilemma, and each player has a subjective prior that includes a probability distribution over the strategies of the potential partners, and chooses a strategy that maximizes his expected payoff subject to this conjecture. Unless all players' conjectures lead to defecting on round 1, $CKR = \emptyset$ for this epistemic game. Nothing, of course, constrains rational agents to hold such a pattern of conjectures, so $CKR = \emptyset$ should be considered the default situation.

More generally, in any epistemic game that has a perfect information extensive form and a unique subgame perfect Nash equilibrium \mathbf{s}^* , the priors $p_i(\cdot|\omega)$ fully determine the probability each player places on the occurrence of \mathbf{s}^* , namely $p_i([\mathbf{s} = \mathbf{s}^*]|\omega)$. This is surely zero unless $\mathbf{s}_i(\omega) = \mathbf{s}^*$. Moreover, we must have $p_i([\mathbf{s} = \mathbf{s}^*]|\omega) = 1$ if $\omega \in \text{CKR}$, which is a restriction on subjective priors that has absolutely no justification in general, although it can be justified in certain cases (e.g., the one- or two-round Prisoner's Dilemma or the game in § 5.3).

It might be suggested that a plausible strategy selection mechanism epistemically justified by some principle other than CKR might succeed in selecting out the subgame perfect equilibrium—for instance extensive form rationality as proposed by Pearce (1984) and Battigalli (1997). However, this selection mechanism is not epistemically grounded at all. There are alternative, epistemically grounded selection mechanisms for extensive form games, such as Fudenberg, Kreps and Levine (1988), Börgers (1994), and Ben-Porath (1997), but these mechanisms do not justify backward induction.