FOUNDATIONS OF STRATEGIC EQUILIBRIUM

JOHN HILLAS AND ELON KOHLBERG

Contents

Date: June, 1996.

Prepared for the *Handbook of Game Theory with Economic Applications* edited by Robert J. Aumann and Sergiu Hart. We are grateful to Jean-François Mertens from whom we have learnt much of what is written here. The detailed comments of Robert Aumann clarified much of the expression, and not a little of the content, of the chapter. We are also grateful to the reviewers, one of whom, in particular, made a large number of detailed and helpful comments.

1. Introduction

The central concept of noncooperative game theory is that of the *strategic equilibrium* (or Nash equilibrium, or noncooperative equilibrium). In this chapter we discuss some of the conceptual issues surrounding this concept and its refinements. Many of these issues have received increasing attention in the last few years. We are not at all comprehensive in covering the approaches that have been taken to the question. In particular, we focus almost exclusively on the "purely rational" approach and say almost nothing about other approaches. We do survey some of the issues and approaches that do not fit neatly into the core of our argument in Section 15.

Though we do examine the idea of correlated equilibrium in Section 6, most of our discussion is in the context of independent strategies. Aumann (1987a) and, in a somewhat different style, Nau and McCardle (1990) argue that correlated equilibrium is a better expression of rationality in games. We don't take any position on that issue here. However many of the issues and concepts we do discuss in terms of the refinements of equilibrium are not as well developed for correlated equilibrium as they are in the setting where stochastic independence of the solution is assumed. (See however our discussion of the work of Aumann and Brandenburger (1991) in Section 15.1 for one example of a theorem in which this independence is proved rather than assumed.)

There is some literature on the refinement of correlated equilibrium. Myerson (1986a and 1986b) was one of the first to examine issues of perfection and sequential rationality in the context of correlated equilibrium. More recently Dhillon and Mertens (1991) reexamined the issue of perfection. There are also a number of comments in the work of Mertens (1987,1989,1991b) on strategic stability indicating how the theory would be developed allowing correlation.

This chapter is rather informal. Not everything is defined precisely and there is little use, except in examples, of symbols. We hope that this will not give our readers any problem. We do not seek to give an introduction to the concept of strategic equilibrium here. Indeed we believe that one shouldn't worry too much about the foundations of this concept, or perhaps any concept, until one has some familiarity with its use and power. Readers who need or want formal definitions of the concepts we discuss here could consult the chapters by Hart (1992) and van Damme (1994) in this Handbook.

2. Pre-equilibrium Ideas

Before discussing the idea of equilibrium in any detail we shall look at some weaker conditions. We might think of these conditions as necessary implications of assuming that the players know the game, including the rationality and knowledge of the others. Of course, if each player knows the knowledge of the others then all that they know will be common knowledge, in the sense of Aumann (1976).

2.1. **Iterated Dominance and Rationalizability.**

Consider the problem of a player in some game. Except in the most trivial cases the set of strategies that he will be prepared to play will depend on his assessment of what the other players will do. However it *is* possible to say a little. If some strategy was strictly preferred by him to another strategy s whatever he thought the other players would do, then he surely would not play s. And this remains true if it was some lottery over his

strategies that was strictly preferred to s. We call a strategy such as s a *strictly dominated strategy*.

Perhaps we could say a little more. A strategy s would surely not be played unless there was some assessment of the manner in which the others might play that would lead s to be (one of) the best. This is clearly at least as restrictive as the first requirement. (If s is best for some assessment it cannot be strictly worse than some other strategy for all assessments.) In fact, if the set of assessments of what the others might do is convex (as a set of probabilities on the vectors of pure strategies of the others) then the two requirements are equivalent. This will be true if there is only one other player, or if a player's assessment of what the others might do permits correlation. However, the set of product distributions over the product of two or more players' pure strategy sets is not convex.

Thus we have two cases: one in which we eliminate strategies that are strictly dominated, or equivalently never best against some distribution on the vectors of pure strategies of the others; and one in which we eliminate strategies that are never best against some product of distributions on the pure strategy sets of the others. In either case we have identified a set of strategies that we argue a rational player would not play. But since everything about the game, including the rationality of the players, is assumed to be common knowledge no player should put positive weight, in his assessment of what the other players might do, on such a strategy. And we can again ask: Are there any strategies that are strictly dominated when we restrict attention to the assessments that put weight only on those strategies of the others that are not strictly dominated. If so, a rational player who knew the rationality of the others would surely not play such a strategy. And similarly for strategies that were not best responses against some assessment putting weight only on those strategies of the others that are best responses against some assessment by the others.

And we can continue for an arbitrary number of rounds. If there is ever a round in which we don't find any new strategies that will not be played by rational players commonly knowing the rationality of the others, we would never again "eliminate" a strategy. Thus, since we start with a finite number of strategies, the process must eventually terminate. We call the strategies that remain *iteratively undominated* or *correlatedly rationalizable* in the first case; and *rationalizable* in the second case. The term rationalizable strategy and the concept were introduced by Bernheim (1984) and Pearce (1984). The term correlatedly rationalizable strategy and the concept were explicitly introduced by Brandenburger and Dekel (1987), who also show the equivalence of this concept to what we are calling iteratively undominated strategies, though both the concept and this equivalence are alluded to by Pearce (1984).

The issue of whether or not the assessments of one player of the strategies that will be used by the others should permit correlation has been the topic of some discussion in the literature. Aumann (1987) argues strongly that they should. Others have argued that there is at least a case to be made for requiring the assessments to exhibit independence. For example, Bernheim (1986) argues as follows.

> Aumann has disputed this view [that assessments should exhibit independence]. He argues that there is no *a priori* basis to exclude

> any probabilistic beliefs. Correlation between opponents' strategies may make perfect sense for a variety of reasons. For example, two players who attended the same "school" may have similar dispositions. More generally, while each player knows that his decision does not directly affect the choices of others, the substantive information which leads him to make one choice rather than another also affects his beliefs about other players' choices.
>
> Yet Aumann's argument is not entirely satisfactory, since it appears to make our theory of rationality depend upon some ill-defined "dispositions" which are, at best, extra-rational. What is the "substantive information" which disposes an individual towards a particular choice? In a pure strategic environment, the only available substantive information consists of the features of the game itself. This information is the same, regardless of whether one assumes the role of an outside observer, or the role of a player with a particular "disposition". Other information, such as the "school" which a player attended, is simply extraneous. Such information could only matter if, for example, different schools taught different things. A "school" may indeed teach not only the information embodied in the game itself, but also "something else"; however, differences in schools would then be substantive only if this "something else" was substantive. Likewise, any apparently concrete source of differences or similarities in dispositions can be traced to an amorphous "something else", which does not arise directly from considerations of rationality.

This addresses Aumann's ideas in the context of Aumann's arguments concerning correlated equilibrium. Without taking a position here on those arguments, it does seem that in the context of a discussion of rationalizability the argument for independence is not valid. In particular, even if one accepts that one's opponents actually choose their strategies independently and that there is nothing substantive that they have in common outside their rationality and the roles they might have in the game, another player's assessment of what they are likely to play could exhibit correlation.

Consider the three player game of 1. Here Players 1 and 2 play a game having two pure strategy equilibria. This game has been much discussed in the literature as the "Stag Hunt" game. (See Aumann (1990), for example.) For our purposes here all that is relevant is that it is not clear which equilibrium that Players 1 and 2 will play and that which equilibrium is more likely has something to do with how rational players in general would think about playing a game. If players in general tend to "play safe" then the equilibrium (B, R) seems likely, while if they tend to coordinate on efficient equilibria then (T, L) seems likely. Player 3 has a choice that does not affect the payoffs of Players 1 and 2, but whose value to him does depend on the choices of 1 and 2. If Players 1 and 2 play (B, R) then Player 3 does best by choosing E, while if they play (T, L) then Player 3 does best by choosing W.

Now suppose that Player 3 knows that the other players were independently randomly chosen to play the game and that they have no further information about each other and that they choose their strategies independently. Now if it is common knowledge that the players were rational and if there were a commonly known

	L	R
T	$9, 9, 1$	$0, 7, 0$
B	$7, 0, 0$	$8, 8, 1$

W

	L	R
T	$9, 9, 0$	$0, 7, 0$
B	$7, 0, 0$	$8, 8, 1$

E

Figure 1

distribution of what players in such a situation did then such a distribution should form a Nash equilibrium. On the other hand if they don't know the distributions then it seems natural to allow them to have nondegenerate distributions over the distributions of what rational players commonly knowing the rationality of the others do in such a game. Now, the action taken by Player 1 will, in general, give Player 3 some information on which he will update his distribution over the distributions of what rational players do. And this will lead to correlation in his assessment of what Players 1 and 2 will do in the game. Indeed, in this setting, requiring independence essentially amounts to requiring that players be certain about things about which we are explicitly allowing them to be wrong.

We indicated earlier that the set of product distributions over two or more players' strategy sets was not convex. This is correct, but somewhat incomplete. It is possible to put a linear structure on this set that would make it convex. In fact, this is exactly what we do in the proof of the existence of equilibrium below. What we mean is that if we think of the product distributions as a subset of the set of all probability distributions on the vectors of pure strategies and use the linear structure that is natural for that latter set then the set of product distributions is not convex. If instead we use the product of the linear structures on the spaces of distributions on the individual strategy spaces, then the set of product distributions will indeed be convex. The nonconvexity however reappears in the fact that with this linear structure the expected payoff function is no longer linear—or even quasi-concave.

2.2. **Strengthening Rationalizability.** There have been a number of suggestions as to how to strengthen the notion of rationalizability. Many of these involve some form of the iterated deletion of (weakly) dominated strategies, that is, strategies such that some other (mixed) strategy does at least as well whatever the other players do, and strictly better for some choices of the others. The difficulty with such a procedure is that the order in which weakly dominated strategies are eliminated can affect the outcome at which one arrives. Now it is certainly possible to give a definition that unambiguously determines the order but such a definition implicitly rests on the assumption that the other players will view a strategy eliminated in a later round as infinitely more likely than one eliminated earlier.

Having discussed in the previous section the reasons for rejecting the requirement that a player's beliefs over the choices of two of the other players be independent we shall not again discuss the issue but shall allow correlated beliefs in all of the definitions we discuss. This has two implications. The first is that our statement below of Pearce's notion of cautiously rationalizable strategies will not be his original definition but rather the suitably modified one. The second is that we shall be able

to simplify the description by referring simply to rounds of deletions of dominated strategies rather than the somewhat more complicated notions of rationalizability.

Even before the definition of rationalizability, Moulin (1979) suggested using as a solution an arbitrarily large number of rounds of the elimination of all weakly dominated strategies for all players. Moulin actually proposed this as a solution only when it led to a set of strategies for each player such that whichever of the allowable strategies the others were playing the player would be indifferent among his own allowable strategies.

A somewhat more sophisticated notion is that of *cautiously rationalizable strategies* defined by Pearce (1984). The set of such strategies is the set obtained by the following procedure. One first eliminates all strictly dominated strategies, and does this for an arbitrarily large number of rounds until one reaches a game in which there are no strictly dominated strategies. One then has a single round in which all weakly dominated strategies are eliminated. One then starts again with another (arbitrarily long) sequence of rounds of elimination of strictly dominated strategies, and again follows this with a single round in which all weakly dominated strategies are removed, and so on. For a finite game such a process ends after a finite number of rounds.

Each of these definitions has a certain apparent plausibility. Nevertheless they are not well motivated. Each depends on an implicit assumption that a strategy eliminated at a later round is much more likely than a strategy eliminated earlier. And this in turn depends on an implicit assumption that in some sense the strategies deleted at one round are equally likely. For suppose we could split one of the rounds of the elimination of weakly dominated strategies and eliminate only part of the set. This could completely change the entire process that follows.

	X	Y	Z
A	$1,1$	$1,0$	$1,0$
B	$0,1$	$1,0$	$2,0$
C	$1,0$	$1,1$	$0,0$
D	$0,0$	$0,0$	$1,1$

Figure 2

Consider, for example, the game of Figure 2. The only cautiously rationalizable strategies are A for Player 1 and X for Player 2. In the first round strategies C and D are (weakly) dominated for Player 1. After these are eliminated strategies Y and Z are strictly dominated for Player 2. And after these are eliminated strategy B is strictly dominated for Player 1. However, if (A, X) is indeed the likely outcome then perhaps strategy D is, in fact, much less likely than strategy C, since, given that 2 plays X, strategy C is one of Player 1's best responses, while D is not. Suppose that we start by eliminating just D. Now, in the second round only Z is strictly dominated. Once Z is eliminated, we eliminate B for Player 1, but nothing else. We are left with A and C for Player 1 and X and Y for Player 2.

There seems to us one slight strengthening of rationalizability that is well motivated. It is one round of elimination of weakly dominated strategies followed by an

arbitrarily large number of rounds of elimination of strictly dominated strategies. This solution is obtained by Dekel and Fudenberg (1990), under the assumption that there is some small uncertainty about the payoffs, by Börgers (1994), under the assumption that rationality was "almost" common knowledge, and by Ben-Porath (1995) for the class of generic extensive form games with perfect information.

The papers of Dekel and Fudenberg and of Börgers use some approximation to the common knowledge of the game and the rationality of the players in order to derive, simultaneously, admissibility and some form of the iterated elimination of strategies. Ben-Porath obtains the result in extensive form games because in that setting a natural definition of rationality implies more than simple ex ante expected utility maximization. An alternative justification is possible. Instead of deriving admissibility, we include it in what we mean by rationality. A choice s is admissibly rational against some conjecture c about the strategies of the others if there is some sequence of conjectures putting positive weight on all possibilities and converging to c such that s is maximizing against each conjecture in the sequence. Now common knowledge of the game and of the admissible rationality of the players gives precisely the set we described. The argument is essentially the same as the argument that common knowledge of rationality implies correlated rationalizability.

3. The Idea of Equilibrium

In the previous section we examined the extent to which it is possible to make predictions about players' behavior in situations of strategic interaction based solely on the common knowledge of the game, including in the description of the game the players' rationality. The results are rather weak. In some games these assumptions do indeed restrict our predictions, but in many they imply few, if any, restrictions.

To say something more, a somewhat different point of view is productive. Rather than starting from only knowledge of the game and the rationality of the players and asking what implications can be drawn, one starts from the supposition that there is some established way in which the game will be played and asks what properties this manner of playing the game must satisfy in order not to be self-defeating, that is, so that a rational player, knowing that the game will be played in this manner, does not have an incentive to behave in a different manner. This is the essential idea of a strategic equilibrium, first defined by John Nash (1950,1951).

There are a number of more detailed stories to go along with this. The first was suggested by von Neumann and Morgenstern (1944), even before the first definition of equilibrium by Nash. It is that players in a game are advised by game theorists on how to play. In each instance the game theorist, knowing the player's situation, tells the player what the theory recommends. The theorist does offer a (single) recommendation in each situation and all theorists offer the same advice. One might well allow these recommendations to depend on various "real-life" features of the situation that are not normally included in our models. One would ask what properties the theory should have in order for players to be prepared to go along with its recommendations. This idea is discussed in a little more detail in the introduction of the chapter on "Strategic Equilibrium" in this Handbook (van Damme, 1994).

Alternatively, one could think of a situation in which the players have no information beyond the rules of the game. We'll call such a game a *Tabula-Rasa game.* If

a player's optimal choice depends on the actions of the others then, since he doesn't know those actions we might argue that he will form some probabilistic assessment of them. We might go on to argue that since the players have precisely the same information they will form the same assessments about how choices will be made. Again, one could ask what properties this common assessment should have in order not to be self-defeating. The first to make the argument that players having the same information should form the same assessment was Harsanyi (1967-1968, Part III). Aumann (1974, p. 92) labeled this view the Harsanyi doctrine.

Yet another approach is to think of the game being preceded by some stage of "pre-play negotiation" during which the players may reach a non-binding agreement as to how each should play the game. One might ask what properties this agreement should have in order for all players to believe that everyone will act according to the agreement. One needs to be a little careful about exactly what kind of communication is available to the players if one wants to avoid introducing correlation. Bárány (1992) and Forges (1990) show that with at least four players and a communication structure that allows for private messages any correlated equilibrium "is" a Nash equilibrium of the game augmented with the communication stage. There are also other difficulties with the idea of justifying equilibria by pre-play negotiation. See Aumann (1990).

Rather than thinking of the game being played by a fixed set of players one might think of each player as being drawn from a population of rational individuals who find themselves in similar roles. The specific interactions take place between randomly selected members of these populations, who are aware of the (distribution of) choices that had been made in previous interactions. Here one might ask what distributions are self-enforcing, in the sense that if players took the past distributions as a guide to what the others choices were likely to be, the resulting optimal choices would (could) lead to a similar distribution in the current round. One already finds this approach in Nash (1950).

A somewhat different approach sees each player as representing a whole population of individuals, each of whom is "programmed" (for example, through his genes) to play a certain strategy. The players themselves are not viewed as rational, but they are assumed to be subject to "natural selection," that is, to the weeding out of all but the payoff-maximizing programs. Evolutionary approaches to game theory were introduced by Manyard Smith and Price (1973).

In most of the rest of this chapter we shall consider only interpretations that involve rational players. We shall return to discuss the evolutionary interpretation briefly in Section 15.2.

3.1. **Self-Enforcing Plans.** One interpretation of equilibrium sees the focus of the analysis as being the actual strategies chosen by the players, that is, their plans in the game. An equilibrium is defined to be a self-enforcing profile of plans. At least a necessary condition for a profile of plans to be self-enforcing is that each player, given the plans of the others should not have an alternate plan that he strictly prefers.

This is the essence of the definition of an equilibrium. As we shall soon see, in order to guarantee that such a self-enforcing profile of plans exists we must consider not only deterministic plans, but also random plans. That is, as well as being permitted to plan what to do in any eventuality in which he might find

himself, a player is explicitly thought of as planning to use some lottery to choose between such deterministic plans.

Such randomizations have been found by many to be somewhat troubling. Arguments are found in the literature that such "mixed strategies" are less stable than pure strategies. (There is, admittedly, a precise sense in which this is true.) And, in the early game theory literature there is discussion as to what precisely it means for a player to choose a mixed strategy and why players may choose to use such strategies. See, for example, the discussion in Luce and Raiffa (1957, pp. 74–76).

Harsanyi (1973) provides an interpretation that avoids this apparent instability and, in the process, provides a link to the interpretation of Section 3.2. Harsanyi considers a model in which there is some small uncertainty about the players' payoffs. This uncertainty is independent across the players, but each player knows his own payoff. The uncertainty is assumed to be of a form such that it is represented by a probability distribution with a continuously differentiable density. If for each player and each vector of pure actions (that is, strategies in the game without uncertainty) the probability that the payoff is close to some particular value is high, then we might consider the game close to the game in which the payoff is exactly that value. Conversely, we might consider the game in which the payoffs are known exactly to be well approximated by the game with small uncertainty about the payoffs.

Harsanyi shows that in a game with such uncertainty about payoffs all equilibria are essentially pure, that is, each player plays a pure strategy with probability 1. Moreover, with probability 1, each player is playing his unique best response to the strategies of the other players; and the expected mixed actions of the players will be close to an equilibrium of the game without uncertainty. Harsanyi also shows, modulo a small technical error later corrected by van Damme (1991), that any regular equilibrium can be approximated in this way by pure equilibria of a game with small uncertainty about the payoffs. We shall not define a regular equilibrium here. Nor shall we give any of the technical details of the construction, or any of the proofs. The reader should instead consult Harsanyi (1973), van Damme (1991), or van Damme (1994).

3.2. **Self-Enforcing Assessments.** Let us consider again Harsanyi's construction described in the previous section. In an equilibrium of the game with uncertainty no player consciously randomizes. Given what the others are doing the player has a strict preference for one of his available choices. However the player does not know what the others are doing. He knows that they are not randomizing—like him they have a strict preference for one of their available choices—but he does not know precisely their payoffs. Thus, if the optimal actions of the others differ as their payoffs differ, the player will have some probabilistic assessment of the actions that the others will take. And, since we assumed that the randomness in the payoffs was independent across players, this probabilistic assessment will also be independent, that is, it will be a vector of mixed strategies of the others.

The mixed strategy of a player does not represent a conscious randomization on the part of that player, but rather the uncertainty in the minds of the others as to how that player will act. We see that even without the construction involving uncertainty about the payoffs, we could adopt this interpretation of a mixed strategy. This interpretation has been suggested and promoted by Bob Aumann for some time (for example Aumann, 1987a, Aumann and Brandenburger, 1995) and

is, perhaps, becoming the preferred interpretation among game theorists. There is nothing in this interpretation that compels us to assume that the assessments over what the other players will do should exhibit independence. Nevertheless, we shall *assume* that the assessments are independent.

Thus the focus of the analysis becomes, not the choices of the players, but the assessments of the players about the choices of the others. The basic consistency condition that we impose on the players' assessments is this: A player reasoning through the conclusions that others would draw from their assessments, should not be led to revise his own assessment.

More formally, Aumann and Brandenburger (1995, p. 1177) show that if each player's assessment of the choices of the others is independent across the other players, if any two players have the same assessment as to the actions of a third, and if these assessments, the game, and the rationality of the players are all mutually known, then the assessments constitute a strategic equilibrium. (A fact is *mutually known* if each player knows the fact and knows that the others know it.) We discuss this and related results in more detail in Section 15.1.

4. The Mixed Extension of a Game

Before discussing exactly how rational players assess their opponents' choices, we must reflect on the manner in which the payoffs represent the outcomes. If the players are presumed to quantify their uncertainties about their opponents' choices, then in choosing among their own strategies they must, in effect, compare different lotteries over the outcomes. Thus for the description of the game it no longer suffices to ascribe a payoff to each outcome, but it is also necessary to ascribe a payoff to each lottery over outcomes. Such a description would be unwieldy unless it could be condensed to a compact form.

One of the major achievements of von Neumann and Morgenstern (1944) was the development of such a compact representation ("cardinal utility"). They showed that if a player's ranking of lotteries over outcomes satisfied some basic conditions of consistency, then it was possible to represent that ranking by assigning numerical "payoffs" just to the outcomes themselves, and by ranking lotteries according to their expected payoffs. See the chapter by Fishburn (1994) in this handbook for details.

Assuming such a scaling of the payoffs, one can expand the set of strategies available to each player to include not only definite ("pure") choices but also probabilistic ("mixed") choices, and extend the definition of the payoff functions by taking the appropriate expectations. The strategic form obtained in this manner is called the *mixed extension* of the game.

Recall from Section 3.2 that we consider a situation in which each player's assessment of the strategies of the others can be represented by a product of probability distributions on the others' pure strategy sets, that is, by a mixed strategy for each of the others. And that any two players have the same assessment about the choices of a third.

Denoting the (identical) assessments of the others as a probability distribution (mixed strategy) over a player's (pure) choices, we may describe the consistency condition as follows: Each player's mixed strategy must place positive probability only on those pure strategies which maximize the player's payoff given the others'

mixed strategies. Thus a profile of consistent assessments may be viewed as a strategic equilibrium in the mixed extension of the game.

Let us consider an example (Figure 3). Player 1 chooses the row and player 2 (simultaneously) chooses the column. The resulting (cardinal) payoffs are indicated in the appropriate box of the matrix, with player 1's payoff appearing first.

	L	R
T	$2,0$	$0,1$
B	$0,1$	$1,0$

Figure 3

What probabilities could characterize a self-enforcing assessment? A (mixed) strategy for player 1 (that is, an assessment by 2 of how 1 would play) is a vector $(x, 1-x)$, where x lies between 0 and 1 and denotes the probability of playing T. Similarly, a strategy for 2 is a vector $(y, 1-y)$. Now, given x, the payoff-maximizing value of y is indicated in Figure 3a, and given y the payoff-maximizing-value of x is indicated in Figure 3b. When the figures are combined as in Figure 3c, it is evident that the game possesses a single equilibrium, namely $x = \frac{1}{2}, y = \frac{1}{3}$. Thus in a self-enforcing assessment Player 1 must assign a probability of $\frac{1}{3}$ to 2's playing L, and player 2 must assign a probability of $\frac{1}{2}$ to 1's playing T.

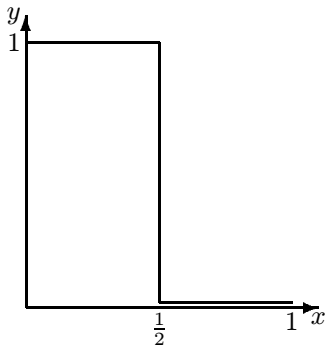


Figure 3a

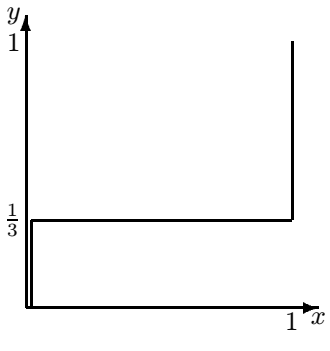


Figure 3b

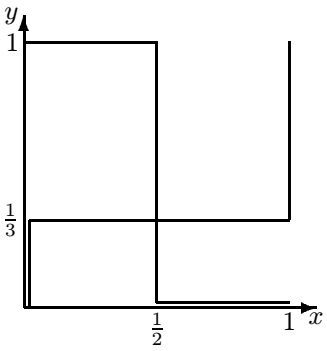


Figure 3c

Note that these assessments do not imply a recommendation for action. For example, they give player 1 no clue as to whether he should play T or B (because the expected payoff to either strategy is the same). But this is as it should be: it is impossible to expect rational deductions to lead to definite choices in a game like Figure 3, because whatever those choices would be they would be inconsistent with their own implications. (Figure 3 admits no pure-strategy equilibrium.) Still, the assessments do provide the players with an a priori evaluation of what the play of the game is worth to them, in this case $\frac{2}{3}$ to 1 and $\frac{1}{2}$ to 2.

The game of Figure 3 is an instance in which our consistency condition completely pins down the assessments that are self-enforcing. In general, we cannot expect such a sharp conclusion.

	L	C	R
T	$8,5$	$0,0$	$6,3$
B	$0,0$	$7,6$	$6,3$

Figure 4

Consider, for example, the game of Figure 4. There are three equilibrium outcomes: (8,5), (7,6) and (6,3) (for the latter, the probability of T must lie between .5 and .6). Thus, all we can say is that there are three different consistent ways in which the players could view this game. Player 1's assessment would be: Either that Player 2 was (definitely) going to play L, or that layer 2 was going to play C, or that Player 2 was going to play R. Which one of these assessments he would, in fact, hold is not revealed to us by means of equilibrium analysis.

5. The Existence of Equilibrium

We have seen that, in the game of Figure 3, for example, there may be no pure self-enforcing plans. But what of mixed plans, or self-enforcing assessments? Could they too be refuted by an example? The main result of non-cooperative game theory states that no such example can be found.

Theorem 1 (Nash 1950, 1951)**.** *The mixed extension of every finite game has at least one strategic equilibrium.*

(A game is *finite* if the player set as well as the set of strategies available to each player is finite.)

Sketch of Proof. The proof may be sketched as follows. (It is a multi-dimensional version of Figure 3c.) Consider the set-valued mapping (or correspondence) that maps each strategy profile, x, to all strategy profiles in which each player's component strategy is a *best response* to x (that is, maximizes the player's payoff given that the others are adopting their components of x). If a strategy profile is contained in the set to which it is mapped (is a *fixed point*) then it is an equilibrium. This is so because a strategic equilibrium is, in effect, defined as a profile that is a best response to itself.

Thus the proof of existence of equilibrium amounts to a demonstration that the "best response correspondence" has a fixed point. The fixed-point theorem of Kakutani (1941) asserts the existence of a fixed point for every correspondence from a convex and compact subset of Euclidean space into itself, provided two conditions hold. One, the image of every point must be convex. And two, the graph of the correspondence (the set of pairs (x, y) where y is in the image of x) must be closed.

Now, in the mixed extension of a finite game, the strategy set of each player consists of all vectors (with as many components as there are pure strategies) of non-negative numbers that sum to 1; that is, it is a simplex. Thus the set of all strategy profiles is a product of simplices. In particular, it is a convex and compact subset of Euclidean space.

Given a particular choice of strategies by the other players, a player's best responses consist of all (mixed) strategies that put positive weight only on those pure

strategies that yield the highest expected payoff among all the pure strategies. Thus the set of best responses is a subsimplex. In particular, it is convex.

Finally, note that the conditions that must be met for a given strategy to be a best response to a given profile are all weak polynomial inequalities, so the graph of the best response correspondence is closed.

Thus all the conditions of Kakutani's theorem hold, and this completes the proof of Nash's theorem. □

Nash's theorem has been generalized in many directions. Here we mention two.

Theorem 2 (Fan 1952, Glicksberg 1952)**.** *Consider a strategic form game with finitely many players, whose strategy sets are compact subsets of a metric space, and whose payoff functions are continuous. Then the mixed extension has at least one strategic equilibrium.*

(Here "mixed strategies" are understood as Borel probability measures over the given subsets of pure strategies).

Theorem 3 (Debreu 1952)**.** *Consider a strategic form game with finitely many players, whose strategy sets are convex compact subsets of a Euclidean space, and whose payoff functions are continuous. If, moreover, each payoff function is quasi-concave in the player's own strategy, then the game has at least one strategic equilibrium.*

(A real-valued function on a Euclidean space is *quasi-concave* if, for each number a, the set of points at which the value of the function is at least a is convex.)

Theorem 2 may be thought of as identifying conditions on the strategy sets and payoff functions so that the game is like a finite game, that is, can be well approximated by finite games. Theorem 3 may be thought of as identifying conditions under which the strategy spaces are like the mixed strategy spaces for the finite games and the payoff functions are like expected utility.

6. Correlated Equilibrium

We have argued that a self-enforcing assessment of the players' choices must constitute an equilibrium of the mixed extension of the game. But our argument has been incomplete: we have not explained why it is sufficient to assess each player's choice separately.

Of course, the implicit reasoning was that, since the players' choices are made in ignorance of one another, the assessments of those choices ought to be independent. In fact, this idea is subsumed in the definition of the mixed extension, where the expected payoff to a player is defined for a product distribution over the others' choices.

Let us now make the reasoning explicit. We shall argue here in the context of a Tabula-Rasa game, as we outlined in Section 3. Let us call the common assessment of the players implied by the Harsanyi doctrine the *rational assessment*. Consider the assessment over the pure-strategy profiles by a rational observer who knows as much about the players as they know about each other. We claim that observation of some player's choice, say Player 1's choice, should not affect the observer's assessment of the other players' choices. This is so because, as regards the other players, the observer and Player 1 have identical information and—by the

Harsanyi doctrine—also identical analyses of that information, so there is nothing that the observer can learn from the player.

Thus, for any player, the conditional probability, given that player's choice, over the others' choices is the same as the unconditional probability. It follows (Aumann and Brandenburger 1995, Lemma 4.6, p. 1169) that the observer's assessment of the choices of all the players must be the product of his assessments of their individual choices. In making this argument we have taken for granted that the strategic form encompasses all the information available to the players in a game. This assumption, of "completeness," ensured that a player had no more information than was available to an outside observer.

Aumann (1974, 1987) has argued against the completeness assumption. His position may be described as follows: It is impractical to insist that every piece of information available to some player be incorporated into the strategic form. This is so because the players are bound to be in possession of all sorts of information about random variables which are strategically irrelevant (that is, which cannot affect the outcome of the game). Thus he proposes to view the strategic form as an incomplete description of the game, indicating the available "actions" and their consequences; and to take account of the possibility that the actual choice of actions may be preceded by some unspecified observations by the players.

Having discarded the completeness assumption (and hence the symmetry in information between player and observer), we can no longer expect the rational assessment over the pure-strategy profiles to be a product distribution. But what can we say about it? That is, what are the implications of the rational assessment hypothesis itself? Aumann (1987) has provided the answer. He showed that a distribution on the pure-strategy profiles is consistent with the rational assessment hypothesis if and only if it constitutes a correlated equilibrium.

Before going into the details of Aumann's argument, let us comment on the significance of this result. At first blush, it might have appeared hopeless to expect a direct method for determining whether a given distribution on the pure-strategy profiles was consistent with the hypothesis: after all, there are endless possibilities for the players' additional observations, and it would seem that each one of them would have to be tried out. And yet, the definition of correlated equilibrium requires nothing but the verification of a finite number of linear inequalities.

Specifically, a distribution over the pure-strategy profiles constitutes a *correlated equilibrium* if it imputes positive marginal probability only to such pure strategies, s, as are best responses against the distribution on the others' pure strategies obtained by conditioning on s. Multiplying throughout by the marginal probability of s, one obtains linear inequalities (if s has zero marginal probability, the inequalities are vacuous).

Consider, for example, Figure 3. Denoting the probability over the ijth entry of the matrix by p_{ij}, the conditions for correlated equilibrium are as follows:

$$2p_{11} \geq p_{12},\ p_{22} \geq 2p_{21},\ 2p_{21} \geq p_{11},\ \text{and}\ p_{12} \geq p_{22}$$

There is a unique solution: $p_{11} = p_{21} = 1/6$, $p_{12} = p_{22} = 1/3$. So in this case, the correlated equilibria and the Nash equilibria coincide.

For an example of a correlated equilibrium that is not a Nash equilibrium, consider the distribution $1/2(T, L) + 1/2(B, R)$ in Figure 5. The distribution over II's choices obtained by conditioning on T is L with probability 1, and so T is a best response. Similarly for the other pure strategies and the other player.

	L	R
T	$3,1$	$0,0$
B	$0,0$	$1,3$

Figure 5

For a more interesting example, consider the distribution that assigns weight $1/6$ to each non-zero entry of the matrix in Figure 6. (This example is due to Moulin and Vial (1978)). The distribution over II's choices obtained by conditioning on T is C with probability $1/2$ and R with probability $1/2$, and so T is a best response (it yields 1.5 while M yields 0.5 and B yields 1). Similarly for the other pure strategies of Player 1, and for Player II.

	L	C	R
T	$0,0$	$1,2$	$2,1$
M	$2,1$	$0,0$	$1,2$
B	$1,2$	$2,1$	$0,0$

Figure 6

It is easy to see that the correlated equilibria of a game contain the convex hull of its Nash equilibria. What is somewhat less obvious is that the containment may be strict (Aumann 1974), even in payoff space. The game of Figure 6 illustrates this: the unique Nash equilibrium assigns equal weight, $1/3$, to every pure strategy, and hence gives rise to the expected payoffs $(1, 1)$; whereas the correlated equilibrium described above gives rise to the payoffs $(1.5, 1.5)$.

Let us now sketch the proof of Aumann's result. By the rational assessment hypothesis, a rational observer can assess in advance the probability of each possible list of observations by the players. Furthermore, he knows that the players also have the same assessment, and that each player would form a conditional probability by restricting attention to those lists that contain his actual observations. Finally, we might as well assume that the player's strategic choice is a function of his observations (that is, that if the player must still resort to a random device in order to decide between several payoff-maximizing alternatives, then the observer has already included the various outcomes of that random device in the lists of the possible observations).

Now, given a candidate for the rational assessment of the game, that is, a distribution over the matrix, what conditions must it satisfy if it is to be consistent with some such assessment over lists of observations? Our basic condition remains as in the case of Nash equilibria: by reasoning through the conclusions that the players would reach from their assessments, the observer should not be led to revise his own assessment. That is, conditional on any possible observation of a player, the pure strategy chosen must maximize the player's expected payoff.

As stated, the condition is useless for us, because we are not privy to the rational observer's assessment of the probabilities over all the possible observations. However, by lumping together all the observations inducing the same choice, s, (and by noting that, if s maximized the expected payoff over a number of disjoint events then it would also maximize the expected payoff over their union), we obtain a condition that we can check: that the choice of s maximizes the player's payoff against the conditional distribution given s. But this is precisely the condition for correlated equilibrium.

To complete the proof, note that the basic consistency condition has no implications beyond correlated equilibria: Any correlated equilibrium satisfies this condition relative to the following assessment of the players' additional observations: Each player's additional observation is the name of one of his pure strategies, and the probability distribution over the possible lists of observations (that is, over the entries of the matrix) is precisely the distribution of the given correlated equilibrium.

For the remainder of this chapter we shall restrict our attention to uncorrelated strategies. The issues and concepts we discuss concerning the refinement of equilibrium are not as well developed for correlated equilibrium as they are in the setting where stochastic independence of the solution is assumed.

7. The Extensive Form

The strategic form is a convenient device for defining strategic equilibria: it enables us to think of the players as making single, simultaneous choices. However, actually to describe "the rules of the game," it is more convenient to present the game in the form of a tree.

The *extensive form* is a formal representation of the rules of the game. It consists of a rooted tree whose nodes represent decision points (an appropriate label identifies the relevant player), whose branches represent moves and whose endpoints represent outcomes. Each player's decision nodes are partitioned into *information sets* indicating the player's state of knowledge at the time he must make his move: the player can distinguish between points lying in different information sets but cannot distinguish between points lying in the same information set. Of course, the actions available at each node of an information set must be the same, or else the player could distinguish between the nodes according to the actions that were available. This means that the number of moves must be the same and that the labels associated with moves must be the same.

Random events are represented as nodes (usually denoted by open circles) at which the choices are made by Nature, with the probabilities of the alternative branches included in the description of the tree.

The information partition is said to have *perfect recall* (Kuhn 1953) if the players remember whatever they knew previously, including their past choices of moves. In other words, all paths leading from the root of the tree to points in a single information set, say player i's, must intersect the same information sets of player i and must display the same choices by player i.

The extensive form is "finite" if there are finitely many players, each with finitely many choices at finitely many decision nodes. Obviously, the corresponding strategic form is also finite (there are only finitely many alternative "books of instructions"). Therefore, by Nash's theorem, there exists a mixed-strategy equilibrium.

But a mixed strategy might seem a cumbersome way to represent an assessment of a player's behavior in an extensive game. It specifies a probability distribution over complete plans of action, each specifying a definite choice at each of the player's information sets. It may seem more natural to specify an independent probability distribution over the player's moves at each of his information sets. Such a specification is called a *behavioral strategy*.

Is nothing lost in the restriction to behavioral strategies? Perhaps, for whatever reason, rational players do assess their opponents' behavior by assigning probabilities to complete plans of action, and perhaps some of those assessments cannot be reproduced by assigning independent probabilities to the moves?

Kuhn's Theorem (1953) guarantees that, in a game with perfect recall, nothing, in fact, is lost. It says that in such a game every mixed strategy of a player in a tree is equivalent to some behavioral strategy, in the sense that both give the same distribution on the endpoints, whatever the strategies of the opponents.

For example, in the (skeletal) extensive form of Figure 7, while it is impossible to reproduce by means of a behavioral strategy the correlations embodied in the mixed strategy $.1TLW+.1TRY+.5BLZ+.1BLW+.2BRX$, nevertheless it is possible to construct an equivalent behavioral strategy, namely $((.2, .8), (.5, 0, .5), (.125, 0, .25, .625))$.

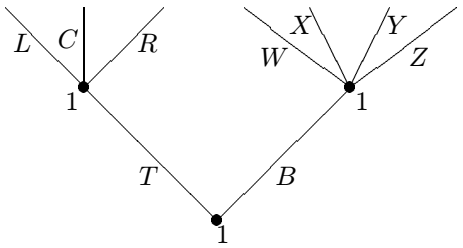


Figure 7

To see the general validity of the theorem, note that the distribution over the endpoints is unaffected by correlation of choices that anyway cannot occur at the same "play" (that is, on the same path from the root to an endpoint). Yet this is precisely the type of correlation that is possible in a mixed strategy but not in a behavioral strategy. (Correlation among choices lying on the same path is possible also in a behavioral strategy. Indeed, this possibility is already built into the structure of the tree: if two plays differ in a certain move, then (because of perfect recall) they also differ in the information set at which any later move is made and so the assessment of the later move can be made dependent on the earlier move.)

Kuhn's Theorem allows us to identify each equivalence class of mixed strategies with a behavioral strategy—or sometimes, on the boundary of the space of behavioral strategies, with an equivalence class of behavioral strategies. Thus, in games with perfect recall, for the strategic choices of payoff-maximizing players there is no difference between the mixed extension of the game and the "behavioral extension" (where the players are restricted to their behavioral strategies). In particular, the equilibria of the mixed and of the behavioral extension are equivalent, so either may be taken as the set of candidates for the rational assessment of the game.

Of course, the equivalence of the mixed and the behavioral extensions implies the existence of equilibrium in the behavioral extension of any finite game with

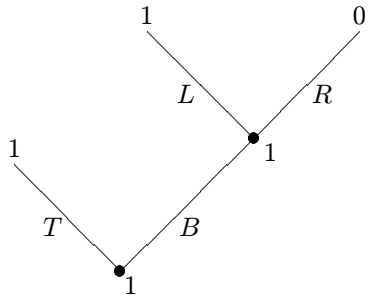


Figure 8

perfect recall. It is interesting to note that this result does not follow directly from either Nash's theorem or from Debreu's theorem. The difficulty is that the convex structure on the behavioral strategies does not reflect the convex structure on the mixed strategies, and therefore the best-reply correspondence need not be convex. For example, in Figure 8, the set of optimal strategies contains (T, R) and (B, L) but not their behavioral mixture $(\frac{1}{2}T + \frac{1}{2}B, \frac{1}{2}L + \frac{1}{2}R)$. This corresponds to the difference between the two linear structures on the space of product distributions on strategy vectors that we discussed at the end of Section 2.1.

8. Refinement of Equilibrium

Let us review where we stand: Assuming that in any game there is one particular assessment of the players' strategic choices that is common to all rational decision makers ("the rational assessment hypothesis"), we can deduce that that assessment must constitute a strategic equilibrium (which can be expressed as a profile of either mixed or behavioral strategies).

The natural question is: Can we go any further? That is, when there are multiple equilibria, can any of them be ruled out as candidates for the self-enforcing assessment of the game? At first blush, the answer seems to be negative. Indeed, if an assessment stands the test of individual payoff maximization, then what else can rule out its candidacy?

And yet, it turns out that it *is* possible to rule out some equilibria. The key insight was provided by Selten (1965). It is that irrational assessments by two different players might each make the other look rational (that is, payoff maximizing).

A typical example is that of Figure 9. The assessment (T, R) certainly does not appear to be self-enforcing. Indeed, it seems clear that Player 1 would play B rather than T (because he can bank on the fact that Player 2—who is interested only in his own payoff—will consequently play L). And yet (T, R) constitutes a strategic equilibrium: Player 1's belief that Player 2 would play R makes his choice of T payoff-maximizing, while Player 2's belief that Player 1 would play T makes his choice of R (irrelevant hence) payoff-maximizing.

Thus Figure 9 provides an example of an equilibrium that can be ruled out as a self-enforcing assessment of the game. (In this particular example there remains only a single candidate, namely (B, L).) By showing that it is sometimes possible

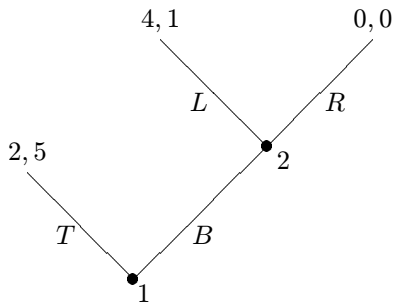


Figure 9

to narrow down the self-enforcing assessments beyond the set of strategic equilibria, Selten opened up a whole field of research: the *refinement* of equilibrium.

8.1. **The Problem of Perfection.** We have described our project as identifying the self-enforcing assessments. Thus we interpret Selten's insight as being that not all strategic equilibria are, in fact, self-enforcing. We should note that this is not precisely how Selten described the problem. Selten explicitly sees the problem as the prescription of disequilibrium behavior in "unreached parts of the game" (Selten 1975, p. 25). Harsanyi and Selten (1988, p. 17) describe the problem of imperfect equilibria in much the same way.

Kreps and Wilson (1982) are a little less explicit about the nature of the problem but their description of their solution suggests that they agree. "A sequential equilibrium provides at each juncture an equilibrium in the subgame (of incomplete information) induced by restarting the game at that point" (p. 864).

Also Myerson seems to view his definition of proper equilibrium as addressing the same problem as addressed by Selten. However he discusses examples and defines a solution explicitly in the context of normal form games where there are no unreached parts of the game. He describes the program as eliminating those equilibria that "may be inconsistent with our intuitive notions about what should be the outcome of a game" (Myerson 1978, p. 73).

The description of the problem of the refinement of strategic equilibrium as looking for a set of necessary and sufficient conditions for self-enforcing behavior does nothing without specific interpretations of what this means. Nevertheless it seems to us to tend to point us in the right direction. Moreover it does seem to delineate the problem to some extent. Thus in the game of Figure 10 we would be quite prepared to concede that the equilibrium (B, R) was unintuitive while at the same time claiming that it was quite self-enforcing.

8.2. **Equilibrium Refinement versus Equilibrium Selection.** There is a question separate from but related to that of equilibrium refinement. That is the question of equilibrium selection. Equilibrium refinement is concerned with establishing necessary conditions for reasonable play, or perhaps necessary and sufficient conditions for "self-enforcing." Equilibrium selection is concerned with narrowing the prediction, indeed to a single equilibrium point. One sees a problem with some of the equilibrium points, the other with the *multiplicity* of equilibrium points.

	L	R
T	$10, 10$	$0, 0$
B	$0, 0$	$1, 1$

Figure 10

The central work on equilibrium selection is the book of Harsanyi and Selten (1988). They take a number of positions in that work with which we have explicitly disagreed (or will disagree in what follows): the necessity of incorporating mistakes; the necessity of working with the extensive form; the rejection of forward induction type reasoning; the insistence on subgame consistency. We are, however, somewhat sympathetic to the basic enterprise. Whatever the answer to the question we address in this chapter there will remain in many games a multiplicity of equilibria, and thus some scope for selecting among them. And the work of Harsanyi and Selten will be a model for those who undertake this enterprise.

9. Admissibility and Iterated Dominance

It is one thing to point to a specific equilibrium, like (T, R) in Figure 9, and claim that "clearly" it cannot be a self-enforcing assessment of the game; it is quite another matter to enunciate a principle that would capture the underlying intuition.

One principle that immediately comes to mind is *admissibility*, namely that rational players never choose dominated strategies. (As we discussed in the context of rationalizability in Section 2.2 a strategy is dominated if there exists another strategy yielding at least as high a payoff against any choice of the opponents and yielding a higher payoff against some such choice.) Indeed, admissibility rules out the equilibrium (T, R) of Figure 9 (because R is dominated by L).

Furthermore, the admissibility principle immediately suggests an extension, *iterated dominance*: If dominated strategies are never chosen, and if all players know this, all know *this*, and so on, then a self-enforcing assessment of the game should be unaffected by the (iterative) elimination of dominated strategies. Thus, for example, the equilibrium (T, L) of Figure 11 can be ruled out even though both T and L are admissible. (See Figure 11a.)

At this point we might think we have nailed down the underlying principle separating self-enforcing equilibria from ones that are not self-enforcing. (Namely, that rational equilibria are unaffected by deletions of dominated strategies). However, nothing could be further from the truth: First, the principle cannot possibly be a general property of self-enforcing assessments, for the simple reason that it is self-contradictory; and second, the principle fails to weed out *all* the equilibria that appear not to be self-enforcing.

On reflection, one realizes that admissibility and iterated dominance have somewhat inconsistent motivations. Admissibility says that whatever the assessment of how the game will be played, the strategies that receive zero weight in this assessment nevertheless remain relevant, at least when it comes to breaking ties. Iterated dominance, on the other hand, says that some such strategies, those that receive zero weight because they are inadmissible, are irrelevant and may be deleted.

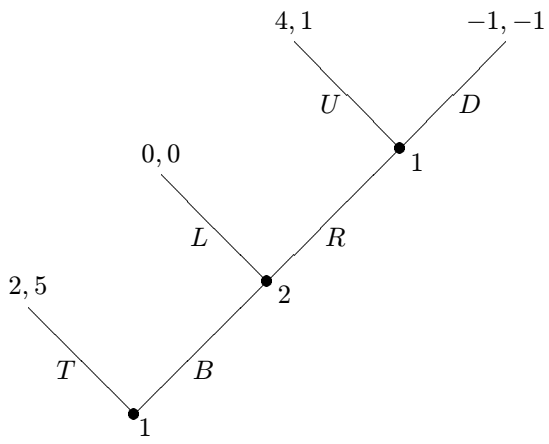


Figure 11

	L	R
T	$2,5$	$2,5$
BU	$0,0$	$4,1$
BD	$0,0$	$-1,-1$

Figure 11a

	L	R
T	$2,0$	$1,0$
M	$0,1$	$0,0$
B	$0,0$	$0,1$

Figure 12

To see that this inconsistency in motivation actually leads to an inconsistency in the concepts, consider the game of Figure 12. If a self-enforcing assessment were unaffected by the elimination of dominated strategies then Player 1's assessment of Player 2's choice would have to be L (delete B and then R) but it would also have to be R (delete M and then L). Thus the assessment of the outcome would have to be both $(2,0)$ and $(1,0)$.

To see the second point, consider the game of Figure 13. As is evident from the normal form of Figure 13a, there are no dominance relationships among the strategies, so all the equilibria satisfy our principle, and in particular those in which Player 1 plays T (for example, $(T, .5L + .5C)$). And yet those equilibria appear not to be self-enforcing: indeed, if Player 1 played B then he would be faced with the

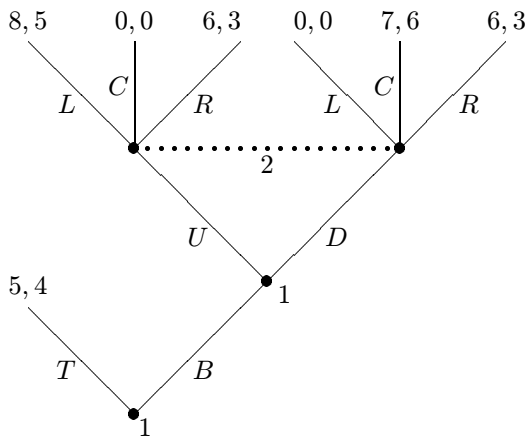


Figure 13

	L	C	R
T	$5,4$	$5,4$	$5,4$
BU	$8,5$	$0,0$	$6,3$
BD	$0,0$	$7,6$	$6,3$

Figure 13a

game of Figure 4, which he assesses as being worth more than 5 (recall that any self-enforcing assessment of the game of Figure 4 has an outcome of either $(8,5)$ or $(7,6)$ or $(6,3)$); thus Player 1 must be expected to play B rather than T.

In the next section, we shall concentrate on the second problem, namely how to capture the intuition ruling out the outcome $(5,4)$ in Figure 13.

10. Backward Induction

10.1. **The Idea of Backward Induction.** Selten (1965,1975) proposed several ideas that may be summarized as the following *principle of backward induction*:

A self-enforcing assessment of the players' choices in a game tree must be consistent with a self-enforcing assessment of the choices from any node (or, more generally, information set) in the tree onwards.

This is a multi-person analog of "the principle of dynamic programming" (Bellman 1957), namely that an optimal strategy in a one-person decision tree must induce an optimal strategy from any point onward.

The force of the backward induction condition is that it requires the players' assessments to be self-enforcing even in those parts of the tree that are ruled out by their own assessment of earlier moves. (As we have seen, the equilibrium condition

by itself does not do this: one can take the "wrong" move at a node whose assessed probability is zero and still maximize one's expected payoff.)

The principle of backward induction indeed eliminates the equilibria of the games of Figures 9 and 11 that do not appear to be self-enforcing. For example, in the game of Figure 11 a self-enforcing assessment of the play starting at Player 1's second decision node must be that Player 1 would play U, therefore the assessment of the play starting at Player 2's decision node must be BU, and hence the assessment of the play of the full game must be BRU, that is, it is the equilibrium (BU, R).

Backward induction also eliminates the outcome $(5, 4)$ in the game of Figure 13. Indeed, any self-enforcing assessment of the play starting at Player 1's second decision node must impute to Player 1 a payoff greater than 5, so the assessment of Player 1's first move must be B.

And it eliminates the equilibrium (T, R, D) in the game of Figure 14 (which is taken from Selten 1975). Indeed, whatever the self-enforcing assessment of the play starting at Player 2's decision node, it certainly is not (R, D) (because, if Player 2 expected Player 3 to choose D, then he would maximize his own payoff by choosing L rather than R).

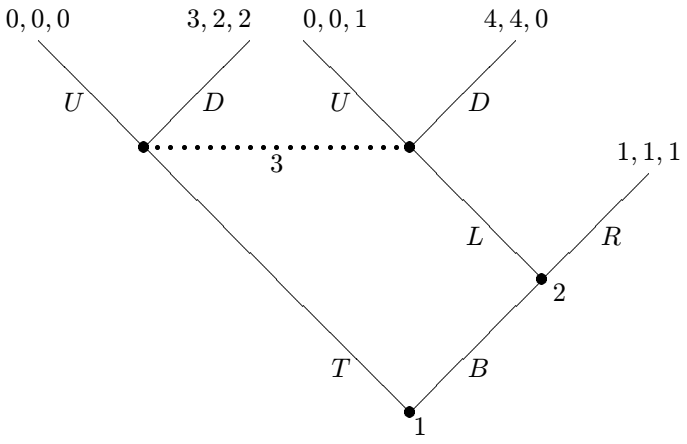


Figure 14

There have been a number of attacks (Basu 1988, 1990, Ben-Porath 1995, Reny 1992a, 1993) on the idea of backward induction along the following lines. The requirement that the assessment be self-enforcing implicitly rests on the assumption that the players are rational and that the other players know that they are rational, and indeed, on higher levels of knowledge. Also, the requirement that a self-enforcing assessment be consistent with a self-enforcing assessment of the choices from any information set in the tree onwards seems to require that the assumption be maintained at that information set and onwards. And yet, that information set might only be reached if some player has taken an irrational action. In such a case the assumption that the players are rational and that their rationality is known to the others should not be assumed to hold in that part of the tree. For example, in the game of Figure 15 there seems to be no compelling reason why Player 2, if called on to move, should be assumed to know of Player 1's rationality. Indeed,

since he has observed something that contradicts Player 1 being rational, perhaps Player 2 *must* believe that Player 1 is not rational.

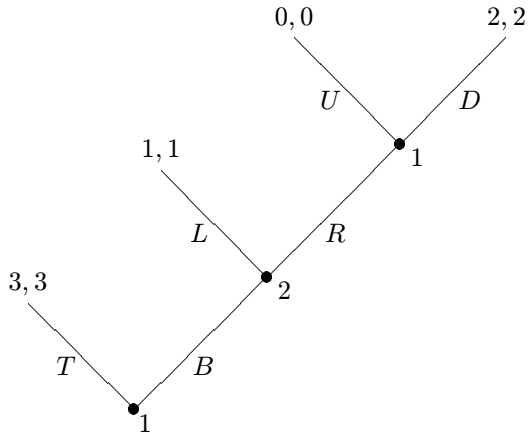


Figure 15

The example however does suggest the following observation: the part of the tree following an irrational move is anyway irrelevant (because a rational player is sure not to take such a move), so whether or not rational players can assess what would happen there has no bearing on their assessment of how the game would actually be played (for example, in the game of Figure 15 the rational assessment of the outcome is $(3, 3)$, regardless of what Player 2's second choice might be). While this line of reasoning is quite convincing in a situation like that of Figure 15, where the irrationality of the move B is self evident, it is less convincing in, say Figure 11. There, the rationality or irrationality of the move B becomes evident only after consideration of what would happen if B *were* taken, that is, only after consideration of what in retrospect appears "counterfactual" (Binmore 1990).

One approach to this is to consider what results when we no longer assume that the players are known to be rational following a deviation by one (or more) from a candidate self-enforcing assessment. Reny (1992a) and, though it is not the interpretation they give, Fudenberg, Kreps, and Levine (1988) show that such a program leads to few restrictions beyond the requirements of strategic equilibrium. We shall discuss this a little more at the end of Section 10.4.

And yet, perhaps one can make some argument for the idea of backward induction. The argument of Reny (1992a), for example, allows a deviation from the candidate equilibrium to be an indication to the other players that the assumption that all of the players were rational is not valid. In other words the players are no more sure about the nature of the game than they are about the equilibrium being played. We may recover some of the force of the idea of backward induction by requiring the equilibrium to be robust against a little strategic uncertainty.

Thus we argue that the requirement of backward induction results from a series of tests. If indeed a rational player, in a situation in which the rationality of all players is common knowledge, would not take the action that leads to a certain information set being reached then it matters little what the assessment prescribes

at that information set. To check whether the hypothesis that a rational player, in a situation in which the rationality of all players is common knowledge, wouldn't take that action we suppose that he would and see what could arise.

If all self-enforcing assessments of the situation following a deviation by a particular player would lead to him deviating then we reject the hypothesis that such a deviation contradicts the rationality of the players. And so, of course, we reject the candidate assessment as self-enforcing.

If however our analysis of the situation confirms that there is a self-enforcing assessment in which the player, if rational, would not have taken the action, then our assessment of him not taking the action is confirmed. In such a case we have no reason to insist on the results of our analysis following the deviation. Moreover, since we assume that the players are rational and our analysis leads us to conclude that rational players will not play in this part of the game we are forced to be a little imprecise about what our assessment says in that part of the game. This relates to our discussion of sets of equilibria as solutions of the game in Section 13.2.

This modification of the notion of backward induction concedes that there may be conceivable circumstances in which the common knowledge of rationality of the players would, of necessity, be violated. It argues, however, that if the players are sure enough of the nature of the game, including the rationality of the other players, that they abandon this belief only in the face of truly compelling evidence, that the behavior in such circumstances is essentially irrelevant.

The principle of backward induction is completely dependent on the extensive form of the game. For example, while it excludes the equilibrium (T, L) in the game of Figure 11, it does not exclude the same equilibrium in the game of Figure 11a (that is, in an extensive form where the players simultaneously choose their strategies).

Thus one might see an inconsistency between the principle of backward induction and von Neumann and Morgenstern's reduction of the extensive form to the strategic form. We would put a somewhat different interpretation on the situation. The claim that the strategic form contains sufficient information for strategic analysis is not a denial that some games have an extensive structure. Nor is it a denial that valid arguments, such as backward induction arguments, can be made in terms of that structure. Rather the point is that, were a player, instead of choosing through the game, required to decide in advance what he will do, he could consider in advance any of the issues that would lead him to choose one way or the other during the game. And further, these issues will affect his incentives in precisely the same way when he considers them before playing as they would had he considered them during the play of the game.

In fact, we shall see in Section 13.6 that the sufficiency of the normal form substantially strengthens the implications of backward induction arguments. We put off that discussion for now. We do note however that others have taken a different position. Selten's position, as well as the position of a number of others, is that the reduction to the strategic form is unwarranted, because it involves loss of information. Thus Figures 13 and 13a represent fundamentally different games, and $(5, 4)$ is indeed not self-enforcing in the former but possibly is self-enforcing in the latter. (Recall that this equilibrium cannot be excluded by the strategic-form arguments we have given to date, such as deletions of dominated strategies, but can be excluded by backward induction in the tree).

10.2. **Subgame Perfection.** We now return to give a first pass at giving a formal expression of the idea of backward induction. The simplest case to consider is of a node such that the part of the tree from the node onwards can be viewed as a separate game (a "subgame"), that is, it contains every information set which it intersects (in particular, the node itself must be an information set).

Because the rational assessment of any game must constitute an equilibrium, we have the following implication of backward induction (*subgame perfection,* Selten 1965):

The equilibrium of the full game must induce an equilibrium on every subgame.

The subgame-perfect equilibria of a game can be determined by working from the ends of the tree to its root, each time replacing a subgame by (the expected payoff of) one of its equilibria. We must show that indeed a profile of strategies obtained by means of step-by-step replacement of subgames with equilibria constitutes a subgame-perfect equilibrium. If not, then there is a smallest subgame in which some player's strategy fails to maximize his payoff (given the strategies of the others). But this is impossible, because the player has maximized his payoff given his own choices in the subgames of the subgame, and those he is presumed to have chosen optimally.

For example, in the game of Figure 11, the subgame whose root is at Player 1's second decision node can be replaced by $(4, 1)$, so the subgame whose root is at Player 2's decision node can also be replaced by this outcome, and similarly for the whole tree.

Or in the game of Figure 13, the subgame (of Figure 4) can be replaced by one of its equilibria, namely $(8, 5), (7, 6)$ or $(6, 3)$. Since any of them give Player 1 more than 5, Player 1's first move must be B. Thus all three outcomes are subgame perfect, but the additional equilibrium outcome, $(5, 4)$, is not.

Because the process of step-by-step replacement of subgames by their equilibria will always yield at least one profile of strategies, we have the following result.

Theorem 4 (Selten 1965)**.** *Every game tree has at least one subgame-perfect equilibrium.*

Subgame perfection captures only one aspect of the principle of backward induction. We shall consider other aspects of the principle in Sections 10.3 and 10.4.

10.3. **Sequential Equilibrium.** To see that subgame perfection does not capture all that is implied by the idea of backward induction it suffices to consider quite simple games. While subgame perfection clearly isolates the self-enforcing outcome in the game of Figure 9 it does not do so in the game of Figure 16, in which the issues seem largely the same. And we could even modify the game a little further so that it becomes difficult to give a presentation of the game in which subgame perfection has any bite. (Say, by having Nature first decide whether Player 1 obtains a payoff of 5 after M or after B and informing Player 1, but not Player 2.)

One way of capturing more of the idea of backward induction is by explicitly requiring players to respond optimally at all information sets. The problem is, of course, that, while in the game of Figure 16 it is clear what it means for Player 2 to respond optimally, this is not generally the case. In general, the optimal choice for a player will depend on exactly which node of his information set has been reached. And, at an out of equilibrium information set this may not be determined by the strategies of his opponents.

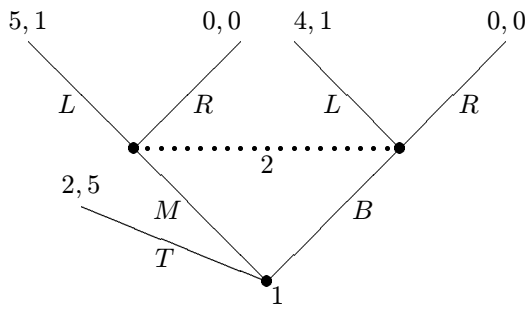


Figure 16

The concept of sequential equilibrium recognizes this by defining an equilibrium to be a pair consisting of a behavioral strategy and a system of beliefs. A *system of beliefs* gives, for each information set, a probability distribution over the nodes of that information set. The behavioral strategy is said to be *sequentially rational* with respect to the system of beliefs if, at every information set at which a player moves, it maximizes the conditional payoff of the player, given his beliefs at that information set and the strategies of the other players. A system of beliefs is said to be *consistent* with a behavioral strategy if it is the limit of a sequence of beliefs each being the actual conditional distribution on nodes of the various information sets induced by a sequence of completely mixed behavioral strategies converging to the given behavioral strategy. A *sequential equilibrium* is a pair such that the strategy is sequentially rational with respect to the beliefs and the beliefs are consistent with the strategy.

The idea of a strategy being sequentially rational appears quite straightforward and intuitive. However the concept of consistency is somewhat less natural. Kreps and Wilson (1982) attempted to provide a more primitive justification for the concept, but, as was shown by Kreps and Ramey (1987) this justification was fatally flawed. Kreps and Ramey suggest that this throws doubt on the notion of consistency. (They also suggest that the same analysis casts doubt on the requirement of sequential rationality. At an unreached information set there is some question of whether a player should believe that the future play will correspond to the equilibrium strategy. We shall not discuss this further.)

Recent work has suggested that the notion of consistency is a good deal more natural than Kreps and Ramey suggest. In particular Kohlberg and Reny (1994) show that it follows quite naturally from the idea that the players' assessments of the way the game will be played reflects certainty or stationarity in the sense that it would not be affected by the actual realizations observed in an identical situation. Related ideas are explored by Battigalli (1994) and Swinkels (1994). We shall not go into any detail here.

The concept of sequential equilibrium is a strengthening of the concept of subgame perfection. Any sequential equilibrium is necessarily subgame perfect, while the converse is not the case. For example, it is easy to verify that in the game of Figure 16 the unique sequential equilibrium involves I choosing M. And a similar

result holds for the modification of that game involving a move of Nature discussed above.

Notice also that the concept of sequential equilibrium, like that of subgame perfection, is quite sensitive to the details of the extensive form. For example in the extensive form game of Figure 17 there is a sequential equilibrium in which Player 1 plays T and Player 2 plays L. However in a very similar situation (we shall later argue strategically identical)—that of Figure 18—there is no sequential equilibrium in which Player 1 plays T. The concepts of extensive form perfect equilibrium and quasi-perfect equilibrium that we discuss in the following section also feature this sensitivity to the details of the extensive form. In these games they coincide with the sequential equilibria.

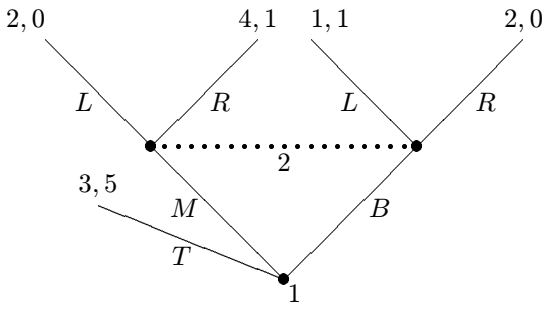


Figure 17

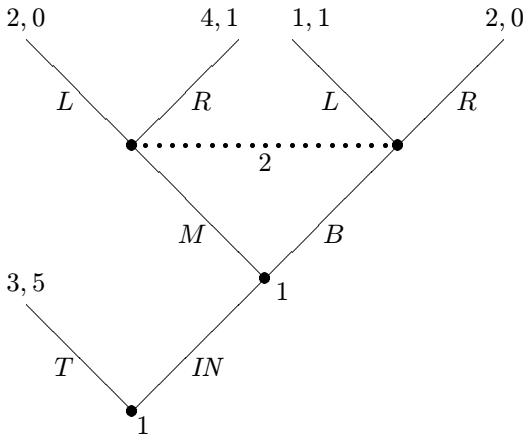


Figure 18

10.4. **Perfect Equilibrium.** The sequential equilibrium concept is closely related to a similar concept defined earlier by Selten (1975), the perfect equilibrium. Another closely related concept, which we shall argue is in some ways preferable, was

defined by van Damme (1984) and called the quasi-perfect equilibrium. Both of these concepts, like sequential equilibrium, are defined explicitly on the extensive form, and depend essentially on details of the extensive form. (They can, of course, be defined for a simultaneous move extensive form game, and to this extent can be thought of as defined for normal form games.) When defining these concepts we shall assume that the extensive form games satisfy perfect recall.

Myerson (1978) defined a normal form concept that he called proper equilibrium. This is a refinement of the concepts of Selten and van Damme when those concepts are applied to normal form games. Moreover there is a remarkable relation between the normal form concept of proper equilibrium and the extensive form concept quasi-perfect equilibrium.

Let us start by describing the definition of perfect equilibrium. The original idea of Selten was that however close to rational players were they would never be perfectly rational. There would always be some chance that a player would make a mistake. This idea may be implemented by approximating a candidate equilibrium strategy profile by a nearby completely mixed strategy profile and requiring that any of the deliberately chosen actions, that is, those given positive probability in the candidate strategy profile, be optimal, not only against the candidate strategy profile, but also against the nearby mixed strategy profile. If we are defining extensive form perfect equilibrium, a strategy is interpreted to mean a behavioral strategy and an action to mean an action at some information set. More formally, a profile of behavioral strategies b is a perfect equilibrium if there is a sequence of completely mixed behavioral strategy profiles $\{b^t\}$ such that at each information set and for each b^t, the behavior of b at the information set is optimal against b^t, that is, is optimal when behavior at all other information sets is given by b^t. If the definition is applied instead to the normal form of the game the resulting equilibrium is called a *normal form perfect equilibrium.*

Like sequential equilibrium, (extensive form) perfect equilibrium is an attempt to express the idea of backward induction. Any perfect equilibrium is a sequential equilibrium (and so is a subgame perfect equilibrium). Moreover the following result tells us that, except for exceptional games, the converse is also true.

Theorem 5 (Kreps and Wilson 1982, Blume and Zame 1994)**.** *For any extensive form, except for a closed set of payoffs of lower dimension than the set of all possible payoffs, the sets of sequential equilibrium strategy profiles and perfect equilibrium strategy profiles coincide.*

The concept of normal form perfect equilibrium, on the other hand, can be thought of as a strong form of admissibility. In fact for two player games the sets of normal form perfect and admissible equilibria coincide. In games with more players the sets may differ. However there is a sense in which even in these games normal form perfection seems to be a reasonable expression of admissibility. Mertens (1987) gives a definition of the admissible best reply correspondence that would lead to fixed points of this correspondence being normal form perfect equilibria, and argues that this definition corresponds "to the intuitive idea that would be expected from a concept of 'admissible best reply' in a framework of independent priors." (Mertens 1987, p. 15.)

Mertens (1991b) offers the following example in which the set of extensive form perfect equilibria and the set of admissible equilibria have an empty intersection. The game may be thought of in the following way. Two players agree about how a

certain social decision should be made. They have to decide who should make the decision and they do this by voting. If they agree on who should make the decision that player decides. If they each vote for the other then the good decision is taken automatically. If each votes for himself then a fair coin is tossed to decide who makes the decision. A player who makes the social decision is not told if this is so because the other player voted for him, or because the coin toss chose him. The extensive form of this game is given in Figure 19. The payoffs are such that each player prefers the good outcome to the bad outcome. However, if the bad outcome is chosen a player will feel slightly worse if he was the one who made the decision, than if it was the other player. We could associate, for example the payoff $(10, 10)$ to the outcome G, $(0, 1)$ to B_1, and $(1, 0)$ to B_2.

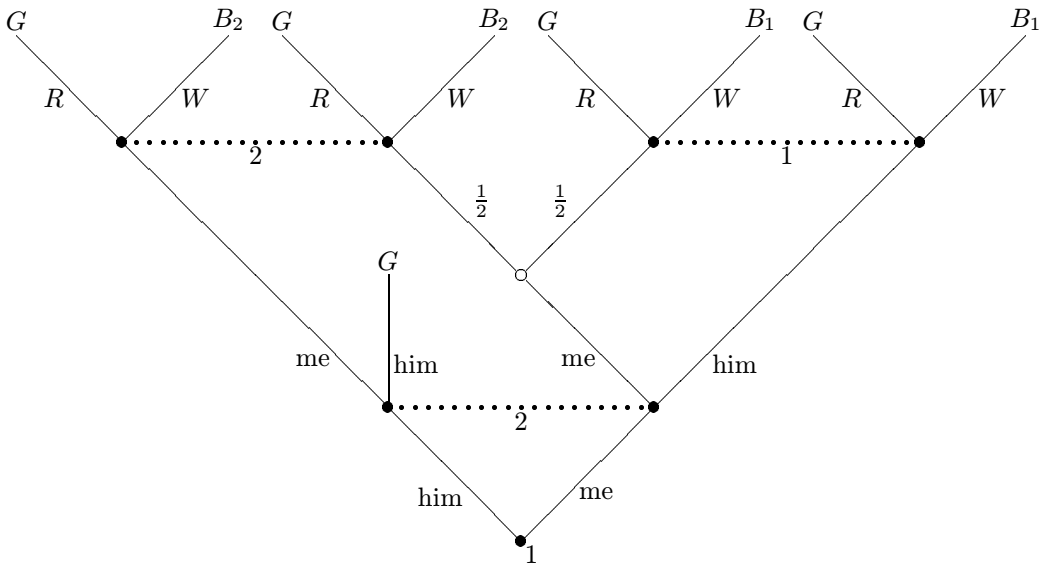


Figure 19

In this game the only admissible equilibrium has both players voting for themselves and taking the right choice if they make the social decision. However, any perfect equilibrium must involve at least one of the players voting for the other with certainty. At least one of the players must be at least as likely as the other to make a mistake in the second stage. And such a player, against such mistakes, does better to vote for the other.

10.5. **Perfect Equilibrium and Proper Equilibrium.** The definition of perfect equilibrium may be thought of as corresponding to the idea that players really do make mistakes, and that in fact it is not possible to think coherently about games in which there is no possibility of the players making mistakes. On the other hand one might think of the perturbations as instead encompassing the idea that the players should have a little strategic uncertainty, that is, they should not be completely

confident as to what the other players are going to do. In such a case a player should not be thought of as being uncertain about his own actions or planned actions. This is (one interpretation of) the idea behind van Damme's definition of quasi-perfect equilibrium.

That definition is largely the same as the definition of perfect equilibrium. The definitions differ only in that, instead of the limit strategy b being optimal at each information set against behavior given by b^t at *all* other information sets, it is required that b be optimal at all information sets against behavior at other information sets given by b for information sets that are owned by the same player who owns the information set in question, and by b^t for other information sets. The assumption of perfect recall guarantees that this requirement of optimality is well defined. That is, the player does not take account of his own "mistakes," except to the extent that they may make one of his information sets reached that otherwise would not be. This change in the definition leads to some attractive properties. Like perfect equilibria, quasi-perfect equilibria are sequential equilibrium strategies. But, unlike perfect equilibria, quasi-perfect equilibria are always normal form perfect, and thus admissible. Mertens (1991b) argues that quasi-perfect equilibrium is precisely the right mixture of admissibility and backward induction.

Also, as we remarked earlier, there is a relation between quasi-perfect equilibria and proper equilibria. A *proper equilibrium* (Myerson 1978) is defined to be a limit of ε-proper equilibria. An *ε-proper equilibrium* is a completely mixed strategy vector such that for each player if, given the strategies of the others, one strategy is strictly worse than another the first strategy is played with probability at most ε times the probability with which the second is played. In other words, more costly mistakes are made with lower frequency. Van Damme (1984) proved the following result. (Kohlberg and Mertens (1986) proved a similar result, replacing quasi-perfect with sequential.)

Theorem 6. *A proper equilibrium of a normal form game is quasi-perfect in any extensive form game having that normal form.*

van Damme actually states his theorem a little differently referring simply to a pair of games, one an extensive form game and the other the corresponding normal form. (He is also more explicit about the sense in which a quasi-perfect equilibrium, a behavioral strategy profile, *is* a proper equilibrium, a mixed strategy profile.) Thus he correctly states that the converse of his theorem is not true. There are such pairs of games and quasi-perfect equilibria of the extensive form that are in no sense equivalent to a proper equilibrium of the normal form. Kohlberg and Mertens (1986) state their theorem in the same form as we do, but refer to sequential equilibria rather than quasi-perfect equilibria. They too correctly state that the converse is not true. For any normal form game one could introduce dummy players, one for each profile of strategies having payoff one at that profile of strategies and zero otherwise. In any extensive form having that normal form the set of sequential equilibrium strategy profiles would be the same as the set of equilibrium strategy profiles originally.

However it is not immediately clear that the converse of the theorem as we have stated it is not true. Certainly we know of no example in the previous literature that shows it to be false. For example, van Damme (1991) adduces the game given in extensive form in Figure 20 and in normal form in Figure 20a to show that

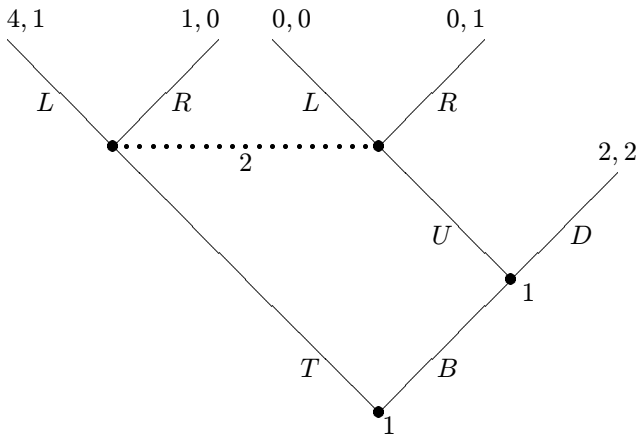


Figure 20

	L	R
T	$4,1$	$1,0$
BU	$0,0$	$0,1$
BD	$2,2$	$2,2$

Figure 20a

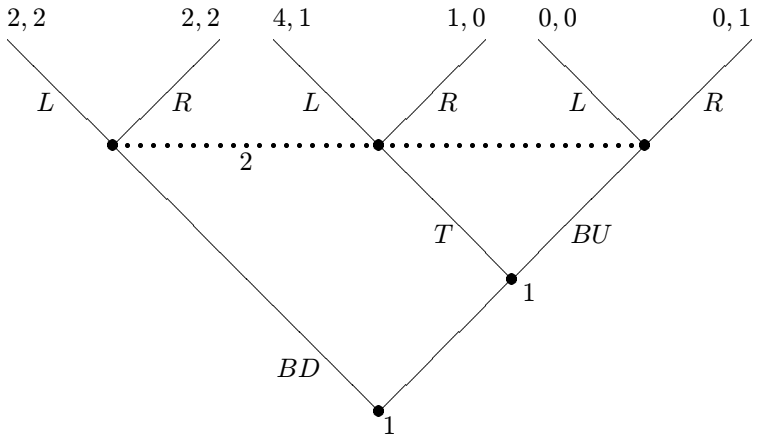


Figure 21

a quasi-perfect equilibrium may not be proper. The strategy (BD, R) is quasi-perfect, but not proper. Nevertheless there is a game—that of Figure 21—having the same normal form, up to duplication of strategies, in which that strategy is not quasi-perfect. Thus one might be tempted to conjecture, as we did in an earlier version of this chapter, that given a normal form game, any strategy vector that is quasi-perfect in any extensive form game having that normal form would be a proper equilibrium of the normal form game.

A fairly weak version of this conjecture is true. We fix not only the equilibrium under consideration but also the sequence of completely mixed strategies converging to it.

Theorem 7. *An equilibrium σ of a normal form game G is supported as a proper equilibrium by a sequence of completely mixed strategies $\{\sigma^k\}$ with limit σ if and only if $\{\sigma^k\}$ induces a quasi-perfect equilibrium in any extensive form game having the normal form G.*

Sketch of Proof. This is proved in Hillas (1996). A slightly weaker version of this theorem is proved in Mailath, Samuelson and Swinkels (1995), using quite different methods. The if direction is implied by van Damme's proof. (Though not quite by the result as he states it, since that leaves open the possibility that different extensive forms may require different supporting sequences.)

The other direction is quite straightforward. One first takes a subsequence such that the conditional probability on any subset of a player's strategy space converges. (This conditional probability is well defined since the strategy vectors in the sequence are assumed to be completely mixed.) Thus $\{\sigma^k\}$ defines, for any subset of a player's strategy space, a conditional probability on that subset. And so the sequence $\{\sigma^k\}$ partitions the strategy space S_n of each player into the sets $S_n^0, S_n^1, \ldots, S_n^J$ where S_n^j is the set of those strategies that receive positive probability conditional on one of the strategies in $S_n \backslash (\cup_{i<j} S_n^i)$ being played.

Now consider the following extensive form. The players move in order of their names, but without observing anything about the choices of those who chose earlier. (That is, they move essentially simultaneously.) Each player n first decides whether to play one of the strategies in S_n^0 or not. If he decides to do so then he chooses one of those strategies. If he decides not to then he decides whether to play one of the strategies in S_n^1, and so on.

The only behavioral strategy in such a game consistent with (the limiting behavior of) $\{\sigma^k\}$ has each player choosing at each opportunity to play one of the strategies in S_n^j rather than continuing with the process, and then choosing among those strategies according to the (strictly positive) limiting conditional (on S_n^j) probability distribution.

Now, if such a vector of strategies is a quasi-perfect equilibrium, then for sufficiently large k, for each player, every strategy in S_n^j is at least as good as any strategy in $S_n^{j'}$ for any $j' > j$ and is assigned probability arbitrarily greater than any such strategy. Thus $\{\sigma^k\}$ supports σ as a proper equilibrium. □

However, the conjecture as we originally stated it—that is, without any reference to the supporting sequences of completely mixed strategies—is not true. Consider the game of Figure 22. The equilibrium (A, V) is not proper. To see this we argue as follows. Given that Player 1 plays A, Player 2 strictly prefers W to Y and X to Y. Thus in any ε-proper equilibrium Y is played with at most ε times the

probability of W, and also at most ε times the probability of X. The fact that Y is less likely than W implies that Player 1 strictly prefers B to C, while the fact that Y is less likely than X implies that Player 1 strictly prefers B to D. Thus in an ε-proper equilibrium C and D are both played with at most ε times the probability of B. This in turn implies that Player 2 strictly prefers Z to V, and so there can be no ε-proper equilibrium in which V is played with probability close to 1. Thus (A, V) is not proper.

	V	W	X	Y	Z
A	$1,1$	$3,1$	$3,1$	$0,0$	$1,1$
B	$1,1$	$2,0$	$2,0$	$0,0$	$1,2$
C	$1,1$	$1,0$	$2,1$	$1,0$	$1,0$
D	$1,1$	$2,1$	$1,0$	$1,0$	$1,0$

Figure 22

Nevertheless, in any extensive form game having this normal form there are perfect—and quasi-perfect—equilibria equivalent to (A, V). The idea is straightforward and not at all delicate. In order to argue that (A, V) was not proper we needed to deduce, from the fact that Player 2 strictly prefers W to Y and X to Y that Y is played with much smaller probability than W, and much smaller probability than X. However there is no extensive representation of this game in which quasi-perfection has this implication. For example, consider an extensive form in which Player 2 first decides whether to play W and, if he decides not to, then decides between X and Y. Now if all we may assume is that Player 2 strictly prefers W to Y and X to Y then we cannot rule out that Player 2 strictly prefers X to W. And in that case it is consistent with the requirements of either quasi-perfectness or perfectness that the action W is taken with probability ε^2 and the action Y with probability ε. This results in the strategy Y being played with substantially greater probability than the strategy W. Something similar results for any other way of structuring Player 2's choice. See Hillas (1996) for greater detail.

10.6. **Uncertainties about the Game.** Having now defined normal form perfect equilibrium we are in a position to be a little more explicit about the work of Fudenberg, Kreps, and Levine (1988) and of Reny (1992a), which we mentioned in Section 10.1. These papers, though quite different in style, both show that if an out of equilibrium action can be taken to indicate that the player taking that action is not rational, or equivalently, that his payoffs are not as specified in the game, then any normal form perfect equilibrium is self-enforcing. Fudenberg, Kreps and Levine also show that if an out of equilibrium action can be taken to indicate that the game was not as originally described—so that others' payoffs may differ as well—then *any* strategic equilibrium is self-enforcing.

11. Forward Induction

In the previous section we examined the idea of backward induction, and also combinations of backward induction and admissibility. If we suppose that we have captured the implications of this idea, are there any further considerations that would further narrow the set of self-enforcing assessments?

Consider the game of Figure 23 (Kohlberg and Mertens 1982 and 1986). Here, every node can be viewed as a root of a subgame, so there is no further implication of "backward induction" beyond "subgame perfection". Since the outcome $(2, 5)$ (that is, the equilibrium where Player 1 plays T and the play of the subgame is (D, R)) is subgame perfect, it follows that backward induction cannot exclude it.

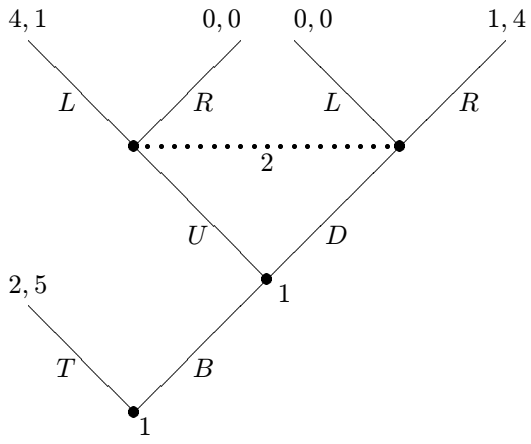


Figure 23

	L	R
T	$2, 5$	$2, 5$
BU	$4, 1$	$0, 0$
BD	$0, 0$	$1, 4$

Figure 23a

On the other hand, this outcome cannot possibly represent a self-enforcing assessment of the game. Indeed, in considering the contingency of Player 1's having played B, Player 2 must take it for granted that Player 1's subsequent move was U and not D (because a rational Player 1 will not plan to play B and then D, which can yield a maximum of 1, when he can play T for a sure 2). Thus the assessment of Player 1's second move must be U, and hence the assessment of Player 2's move must be L, which implies that the assessment of the whole game must be $(4, 1)$.

Notice that this argument does not depend on the order in which Players 1 and 2 move in the subgame. If the order was reversed so that Player 2 moved first in

the subgame the argument could be written almost exactly as it has been requiring only that "was U and not D" be changed to "will be U and not D."

Thus a self-enforcing assessment of the game must not only be consistent with deductions based on the opponents' rational behavior in the future (backward induction) but it must also be consistent with deductions based on the opponents' rational behavior in the past (forward induction).

A formal definition of forward induction has proved a little elusive. One aspect of the idea of forward induction is captured by the idea that a solution should not be sensitive to the deletion of a dominated strategy, as we discussed in Section 9. Other aspects depend explicitly on the idea of a solution as a set of equilibria. Kohlberg and Mertens (1986) give such a definition; and Cho and Kreps (1987) give a number of definitions in the context of signaling games, as do Banks and Sobel (1987). We shall leave this question for the moment and return to it in Sections 13.3, 13.5, and 13.6.

12. Ordinality and Other Invariances

In the following section we shall look again at identifying the self-enforcing solutions. In this section we examine in detail part of what will be involved in providing an answer, namely, the appropriate domain for the solution correspondence. That is, if we are asking simply what behavior (or assessment of the others' behavior) is self-enforcing, which aspects of the game are relevant and which are not?

We divide the discussion into two parts, the first takes the set of players (and indeed their names) as given, while the second is concerned with changes to the player set. The material in this section and the next is a restatement of some of the discussion in and of some of the results from Kohlberg and Mertens (1982, 1986) and Mertens (1987, 1989, 1991a, 1991b, 1991c, 1992).

Recall that we are asking only a rather narrow question: what solutions are self-enforcing. Or, equivalently, what outcomes are consistent with the individual incentives of the players (and the interaction of those incentives)? Thus if some change in the game does not change the individual incentives then it should not change the set of solutions that are self-enforcing. The following two subsections give some formal content to this informal intuition.

Before that, it is well to emphasize that we make no claim that a solution in any other sense, or the strategies that will actually be played, should satisfy the same invariances. This is a point that has been made in a number of places in the papers mentioned above,[1] but is worth repeating.

It is also worth reminding the reader at this point that we are taking the "traditional" point of view, that all uncertainty about the actual game being played has been explicitly included in the description of the game. Other analyses take different approaches. Harsanyi and Selten (1988) assume that with any game there are associated error probabilities for every move, and focus on the case in which all errors have the same small probability. Implicit, at least, in their analysis is a claim that one cannot think coherently, or at least cannot analyze, a game in which there is no possibility of mistakes. Others are more explicit. Aumann (1987b) argues that

[1] Perhaps most starkly in Mertens (1987, p. 6) where he states "There is here thus no argument whatsoever according to which the choice of equilibrium should depend only on this or that aspect of the game."

> to arrive at a strong form of rationality (one must assume that) irrationality cannot be ruled out, that the players ascribe irrationality to each other with small probability. True rationality requires 'noise'; it cannot grow on sterile ground, it cannot feed on itself only.

We differ. Moreover, if we are correct and it is possible to analyze sensibly games in which there are no possibilities of mistakes, then such an analysis is, in theory, more general. One can apply the analysis to a game in which all probabilities of error (or irrationality) are explicitly given.

12.1. **Ordinality.** Let us now consider a candidate solution of a game. This candidate solution would not be self-enforcing if some player, when he came to make his choices, would choose not to play as the solution recommends. That is, if given his beliefs about the choices of the others, the choice recommended for him was not rational. For fixed beliefs about the choices of the others this is essentially a single agent decision problem, and the concept of what is rational is relatively straight forward. There remains the question of whether all payoff maximizing choices, or just admissible ones, are rational, but for the moment we don't need to take a position on that issue.

Now consider two games in which, for some player, for any beliefs on the choices of the others the set of that player's rational choices is the same. Clearly if he had the same beliefs in both games then those recommendations that would be self-enforcing (rational) for him in one game would also be self-enforcing in the other.

Now suppose that this is true for all the players. If the beliefs of the players are consistent with the individual incentives in one of the games they will also be in the other, since given those beliefs the individual decision problems have, by assumption, the same solutions in the two games. Thus a combination of choices and beliefs about the others' choices will be self-enforcing in one game if and only if it is self-enforcing in the other. Moreover, the same argument means that if we were to relabel the strategies the conclusion would continue to hold. A solution would be self-enforcing in the one game if and only if its image under the relabeling was self-enforcing in the other.

Let us flesh this out and make it a little more formal. Previously we have been mainly thinking of the mixed strategy of a player as the beliefs of the other players about the pure action that the player will take. On the other hand there it is difficult to think how we might prevent a player from randomizing if he wanted to. (It is not even completely clear to us what this might mean.) Thus the choices available to a player include, not only the pure strategies that we explicitly give him, but also all randomizations that he might decide to use, that is, his mixed strategy space. A player's beliefs about the choices of the others will now be represented by a probability distribution over the product of the (mixed) strategy spaces of the other players. Since we are not using any of the structure of the mixed strategy spaces, we consider only probability distributions with finite support.

Let us consider an example to make things a little more concrete. Consider the games of Figure 24a and Figure 24b. The game of Figure 24a is the Stag Hunt game of Figure 1 that we discussed in Section 2.1. The game of Figure 24b is obtained from the game of Figure 24a by subtracting 7 from Player 1's payoff when Player 2 plays L and subtracting 7 from Player 2's payoff when Player 1 plays T. This clearly does not, for any assessment of what the other player will do, change

the rational responses of either player. Thus the games are ordinally equivalent, and the self-enforcing behaviours should be the same.

	L	R
T	$9,9$	$0,7$
B	$7,0$	$8,8$

Figure 24a

	L	R
T	$2,2$	$0,0$
B	$0,0$	$8,8$

Figure 24b

There remains the question of whether we should restrict the players' beliefs to be independent across the other players. We argued in Section 2 that in defining rationalizability the assumption of independence was not justified. That argument is perhaps not compelling in the analysis of equilibrium. And, whatever the arguments, the assumption of independence seems essential to the spirit of Nash equilibrium. In any case the analysis can proceed with or without the assumption.

For any such probability distribution the set of rational actions of the player is defined. Here there are essentially two choices: the set of that player's choices (that is, mixed strategies) that maximize his expected payoff, given that distribution; or the set of admissible choices that maximize his payoff. And if one assumes independence there is some issue of how exactly to define admissibility. Again, the analysis can go ahead, with some differences, for any of the assumptions.

For the rest of this section we focus on the case of independence and admissibility. Further, we assume a relatively strong form of admissibility. To simplify the statement of the requirement, the term belief will always mean an independent belief with finite support.

We say a player's choice is an *admissible best reply* to a certain belief p about the choices of the others if, for any other belief q of that player about the others, there are beliefs $\tilde{q}$ arbitrarily close to p and having support containing the support of q, such that the player's choice is expected payoff maximizing against $\tilde{q}$.

The statement of this requirement is a little complicated. The idea is that the choice should be optimal against some completely mixed strategies of the other players close to the actual strategy of the others. However, because of the ordinality requirement we do not wish to use the structure of the strategy spaces, so we make no distinction between completely mixed strategies and those that are not completely mixed. Nor do we retain any notion of one mixed strategy being close to another, except information that may be obtained simply from the preferences of the players. Thus the notion of "close" in the definition is that the sum over all the mixed strategies in the support of either p or $\tilde{q}$ of the difference between p and $\tilde{q}$ is small.

One now requires that if two games are such that a relabeling of the mixed strategy space of one gives the mixed strategy space of the other and that this relabeling transforms the best reply correspondence of the first game into the best reply correspondence of the second then the games are equivalent in terms of the individual incentives and the (self-enforcing) solutions of the one are the same as the (appropriately relabeled) solutions of the other.

Mertens (1987) shows that this rather abstract requirement has a very concrete implementation. Two games exhibit the same self-enforcing behavior if they have the same pure strategy sets and the same best replies for every vector of completely mixed strategies, this remains true for any relabeling of the pure strategies, and also for any addition or deletion of strategies that are equivalent to existing mixed strategies. That is, the question of what behavior is self-enforcing is captured completely by the reduced normal form, and by the best reply structure on that form. And it is only the best reply structure on the interior of the strategy space that is relevant.

If one dropped the assumption that beliefs should be independent one would require a little more, that the best replies to any correlated strategy of the others should be the same. And if one thought of the best replies as capturing rationality rather than admissible best replies, one would need equivalence of the best replies on the boundary of the strategy space as well.

12.2. **Changes in the Player Set.** We turn now to questions of changes in the player set. We start with the most obvious. While, as Mertens (1991c) points out, the names of the players may well be relevant for what equilibrium is actually played, it is clearly not relevant for the prior question of what behavior is self-enforcing. That is, the self-enforcing outcomes are anonymous.

Also it is sometimes the case that splitting a player into two agents does not matter. In particular the set of self-enforcing behaviors does not change if a player is split into two subsets of his agents, provided that in *any* play, whatever the strategies of the players at most one of the subsets of agents moves.

This requirement too follows simply from the idea of self-enforcing behavior. Given the player's beliefs about the choices of the others, if his choice is not rational then it must be not rational for at least one of the subsets, and conversely. Moreover the admissibility considerations are the same in the two cases. It is the actions of Nature and the other players that determines which of the agents actually plays.

Now, at least in the independent case, the beliefs of the others over the choices of the player that are consistent with the notion of self-enforcing behavior and the beliefs over the choices of the subsets of the agents that are consistent with the notion of self-enforcing behavior may differ. But only in the following way. In the case of the agents it is assumed that the beliefs of the others are the independent product of beliefs over each subset of the agents, while when the subsets of the agents are considered a single player no such assumption is made. And yet, this difference has no impact on what is rational behavior for the others. Which of the agents plays in the game is determined by the actions of the others. And so, what is rational for one of the other players depends on his beliefs about the choices of the rest of the other players and his beliefs about the choices of each agent of the player in question conditional on that agent actually playing in the game. And the restrictions on these conditional beliefs will be the same in the two cases.

Notice that this argument breaks down if both subsets of the agents might move in some play. In that case dependence in the beliefs may indeed have an impact. Further, the argument that the decision of the player is decomposable into the individual decisions of his agents in not correct.

We now come to the final two changes, which concern situations in which subsets of the players can be treated separately. The first, called the decomposition property, says that if the players of a game can be divided into two subsets and

that the players in each subset play distinct games with no interaction with the players in the other subset—that is, the payoffs of the one subset do not depend on the actions of the players in the other subset—then the self-enforcing behaviors in the game are simply the products of the self-enforcing behaviors in smaller games. Mertens (1989) puts this more succinctly as "if disjoint player sets play different games in different rooms, one can as well consider the compound game as the separate games." (p. 577)

Finally, we have the small worlds axiom, introduced in Mertens (1991a) and proved for stable sets in Mertens (1992). This says that if the payoffs of some subset of the players do not depend on the choices of the players outside this subset, then the self-enforcing behavior of the players in the subset is the same whether one considers the whole game or only the game between the "insiders."

These various invariance requirements can be thought of as either restrictions on the permissible solutions or as a definition of what constitutes a solution. That is, they define the domain of the solution correspondence. Also, the requirements become stronger in the presence of the other requirements. For example, the player splitting requirement means that the restriction to nonoverlapping player sets in the decomposition requirement can be dispensed with. (For details see Mertens (1989, p. 578).) And, ordinality means that in the decomposition and small worlds requirements we can replace the condition that the payoffs of the subset of players not change with the condition that their best replies not change, or even that their admissible best replies not change.

13. Strategic Stability

We come now to the concept of strategic stability introduced by Kohlberg and Mertens (1986) and reformulated by Mertens (1987, 1989, 1991a, 1991b, 1992) and Govindan and Mertens (1993). This work is essentially aimed at answering the same question as the concepts of Nash equilibrium and perfect equilibrium, namely, what is self-enforcing behavior in a game.

We argued in the previous section that the answer to such a question should depend only on a limited part of the information about the game. Namely, that the answer should be ordinal, and satisfy other invariances. To this extent, as long as one does not require admissibility, Nash equilibrium is a good answer. Nash equilibria are, in some sense, self-enforcing. However, as was pointed out by Selten (1965, 1975), this answer entails only a very weak form of rationality. Certainly, players choose payoff maximizing strategies. But these strategies may prescribe choices of strategies that are (weakly) dominated and actions at unreached information sets that seem clearly irrational.

Normal form perfect equilibria are also invariant in all the senses we require, and moreover are also clearly admissible, even in the rather strong sense we require. However, they do not always satisfy the backward induction properties we discussed in Section 9. Other solutions that satisfy the backward induction property do not satisfy the invariance properties. Sequential equilibrium is not invariant. Even changing a choice between three alternatives into two binary choices can change the set of sequential equilibria. Moreover, sequential equilibria may be inadmissible. Extensive form perfect equilibria too are not invariant. Moreover they too may be inadmissible. Recall that in Section 10.4 we examined a game (that of Figure 17) of Mertens (1991b) in which *all* of the extensive form perfect equilibria are

inadmissible. Quasi-perfect equilibria, defined by van Damme (1984), and proper equilibria, defined by Myerson (1978) are both admissible and satisfy backward induction. However, neither is ordinal—quasi-perfect equilibrium depends on the particular extensive form and proper equilibrium may change with the addition of mixtures as new strategies—and proper equilibrium does not satisfy the player splitting requirement.

We shall see in the next subsection that it is impossible to satisfy the backward induction property for any ordinal single valued solution concept. This leads us to consider set valued solutions and, in particular, to strategically stable sets of equilibria.

13.1. **The Requirements for Strategic Stability.** We are seeking an answer to the question: What are the self-enforcing behaviors in a game? As we indicated, the answer to this question should satisfy the various invariances we discussed in Section 12. We also require that the solution satisfy stronger forms of rationality than Nash equilibrium and normal form perfect equilibrium, the two "non correlated" equilibrium concepts that do satisfy those invariances. In particular, we want our solution concept to satisfy the admissibility condition we discussed in Section 9, and some form of the iterated dominance condition we also discussed in Section 9, the backward induction condition we discussed in Section 10, and the forward induction condition we discussed in Section 11. We also want our solution to give some answer for all games.

As we indicated above, it is impossible for a single valued solution concept to satisfy these conditions. In fact, two separate subsets of the conditions are inconsistent for such solutions. Admissibility and iterated dominance are inconsistent, as are backward induction and invariance.

	L	R
T	$3,2$	$2,2$
M	$1,1$	$0,0$
B	$0,0$	$1,1$

Figure 25

To see the inconsistency of admissibility and iterated dominance consider the game in Figure 25 (taken from Kohlberg and Mertens (1986, p. 1015)). Strategy B is dominated (in fact, strictly dominated) so by the iterated dominance condition the solution should not change if B is deleted. But in the resulting game admissibility implies that (T, L) is the unique solution. Similarly, M is dominated so we can delete M and then (T, R) is the unique solution.

To see the inconsistency of ordinality and backward induction consider the game of Figure 26 (taken from Kohlberg and Mertens (1986, p. 1018)). Nature moves at the node denoted by the circle going left with probability α and up with probability $1-\alpha$. Whatever the value of α the game has the reduced normal form of Figure 26a. (Notice that strategy Y is just a mixture of T and M.) Since the reduced normal form does not depend on α, ordinality implies that the solution of this game should

not depend on α. And yet, in the extensive game the unique sequential equilibrium has Player 2 playing L with probability $(4-3\alpha)/(8-4\alpha)$.)

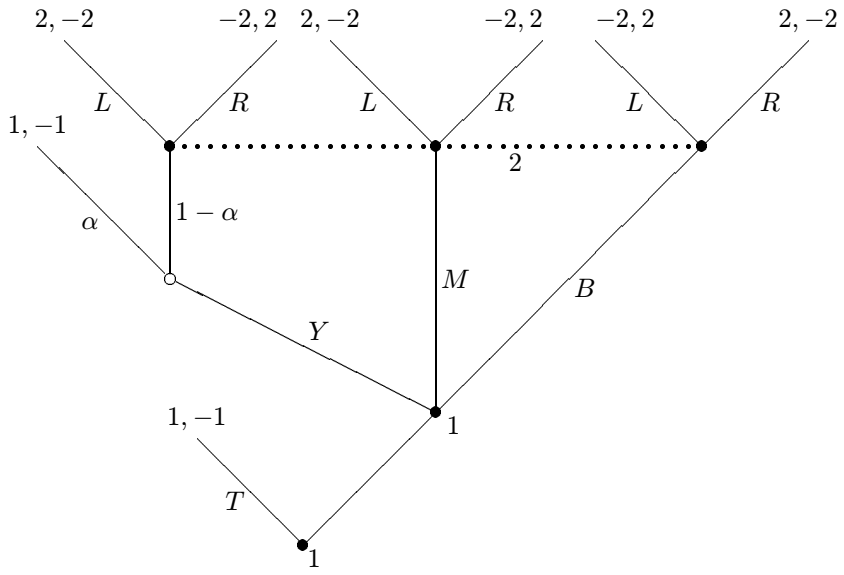


Figure 26

	L	R
T	$1,-1$	$1,-1$
M	$2,-2$	$-2,2$
B	$-2,2$	$2,-2$

Figure 26a

Thus it may be that elements of a solution satisfying the requirements we have discussed would be sets. However we would not want these sets to be too large. We are still thinking of each element of the solution as, in some sense, a single pattern of behavior. In generic extensive form games we might think of a single pattern of behavior as being associated with a single equilibrium outcome, while not specifying exactly the out of equilibrium behavior. One way to accomplish this is to consider only connected sets of equilibria. In the definition of Mertens discussed below the connectedness requirement is strengthened in a way that corresponds, informally, to the idea that the particular equilibrium should depend continuously on the "beliefs" of the players. Without a better understanding of exactly what it means for a set of equilibria to be the solution we cannot say much more. However some form of connectedness seems to be required.

In Section 13.4 we shall discuss a number of definitions of strategic stability that have appeared in the literature. All of these definitions are motivated by the

attempt to find a solution satisfying the requirements we have discussed. As we have just seen such a solution must of necessity be set valued, and we have claimed that such sets should be connected. For the moment, without discussing the details of the definition we shall use the term strategically stable sets of equilibria (or simply stable sets) to mean one of the members of such a solution.

13.2. **Comments on Sets of Equilibria as Solutions to Non-Cooperative Games.** We saw in the previous subsection that single valued solutions could not satisfy the requirements that were implied by a relatively strong form of the notion of self-enforcing. There are two potential responses to this observation. One is to abandon the strong conditions and view the concept of Nash equilibrium or normal form perfect equilibrium as the appropriate concept. The other is to abandon the notion that solutions to noncooperative games should be single valued.

It may seem that the first is more natural, that perhaps the results of Section 13.1 tell us that there is some inconsistency between the invariances we have argued for and the backward induction requirements as we have defined them. We take the opposite view. Our view is that these results tell us that there is something unwarranted in insisting on single valued solutions. Our own understanding of the issue is very incomplete and sketchy and this is reflected in the argument. Nevertheless the issue should be addressed and we feel it is worthwhile to state here what we *do* know, however incomplete.

Consider again the game of Figure 15. In this game the unique stable set contains all the strategy profiles in which Player 1 chooses T. The unique subgame perfect equilibrium has Player 2 choosing R. Recall our interpretation of mixed strategies as the assessment of others of how a player will play. In order to know his best choices Player 1 does not need to know anything about what Player 2 will choose, so there is no need to make Player 1's assessment precise. Moreover from Player 2's perspective one could argue for either L or R equally well on the basis of the best reply structure of the game. They do equally well against T while L is better against BU and R is better against BD. Neither BU nor BD are ever best replies for Player 1, so we cannot distinguish between them on the basis of the best reply correspondence.

What then of the backward induction argument? We have said earlier that backward induction relies on an assumption that a player is rational whatever he has done at previous information sets. This is perhaps a bit inaccurate. The conclusion here is that Player 1, being rational, will not choose to play B. The "assumption" that Player 2 at his decision node will treat Player 1 as rational is simply a test of the assumption that Player 1 will not play B. Once this conclusion is confirmed one cannot continue to maintain the assumption that following a choice of B by Player 1, Player 2 will continue to assume that Player 1 is rational. Indeed he cannot.

In this example Player 1's choices in the equilibrium tells us nothing about his assessment of Player 2's choices. If there is not something else—such as admissibility considerations, for example—to tie down his assessment then there is no basis to do so.

Consider the slight modification of this game of Figure 27. In this game it is not so immediately obvious that Player 1 will not choose B. However if we assume that Player 1 does choose B, believes that Player 2 is rational, and that Player 2 still believes that Player 1 is rational we quickly reach a contradiction. Thus it must not

be that Player 1, being rational and believing in the rationality of the other, chooses B. Again we can no longer maintain the assumption that Player 1 will be regarded as rational in the event that Player 2's information set is reached. Again both L and R are admissible for Player 2. In this case there is a difference between BU and BD for Player 1—BD is admissible while BU is not. However, having argued that a rational Player 1 will choose T there is no compelling reason to make any particular assumption concerning Player 2's assessment of the relative likelihoods of BU and BD. The remaining information we have to tie down Player 1's assessment of Player 2's choices is Player 1's equilibrium choice. Since Player 1 chooses T, it must be that his assessment of Player 2's choices made this a best response. And this is true for any assessment that puts weight no greater than a half on L. Again this set of strategies is precisely the unique strategically stable set.

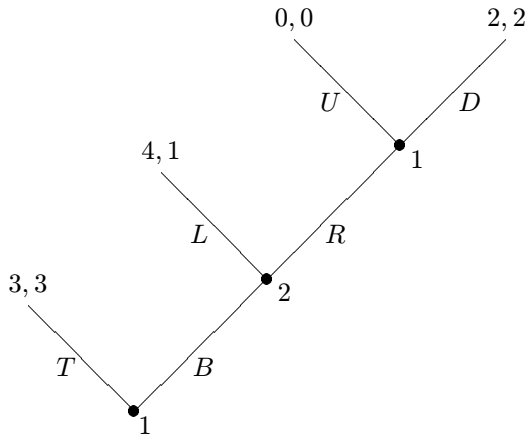


Figure 27

	L	R
T	$3,3$	$3,3$
BU	$4,1$	$0,0$
BD	$4,1$	$2,2$

Figure 27a

These examples point to the fact that there may be something unwarranted in specifying single strategy vectors as "equilibria." We have emphasized the interpretation of mixed strategy vectors as representing the uncertainty in one player's mind about what another player will do. In many games there is no need for a player to form a very precise assessment of what another player will do. Put another way, given the restrictions our ordinality arguments put on the "questions we

can ask a player" we cannot elicit his subjective probability on what the others will do. Rather we must be satisfied with less precise information.

As we said, this is at best suggestive. In fact, if there are more than two players there are real problems. In either of the examples discussed we could add a player with a choice between two strategies, his optimal choice varying in the stable set and not affecting the payoffs of the others. Here too we see a sense in which the set valued solution does represent the strategic uncertainty in the interaction among the players. However we do not see any way in which to make this formal, particularly since the notion of a "strategic dummy" is so elusive. For example, we could again amend the game by adding another game in which the third player interacts in an essential way with the other two and, say, having nature choose which of the two games will be played.

13.3. **Forward Induction.** We discussed in general terms in Section 11 the idea of forward induction; and discussed one aspect of this idea in Section 9. We are now in a position to be a little more explicit about another aspect of the idea.

The idea that the solution should not be sensitive to the (iterated) deletion of dominated strategies was that since it was known that players are never willing to choose such strategies it should not make a difference if such strategies were removed from the game. This suggests another aspect of forward induction. If, in the solution in question, there is some strategy of a player that the player would never be willing to choose then it should not make a difference if such a strategy were removed from the game.

In fact, like the deletion of a dominated strategy, it is clear that the deletion of a strategy that a player is never willing to play in the solution must change the situation. (It can, for example change undominated strategies into dominated ones, and vice versa.) So our intuition is that it should not matter much. This is reflected in the fact that various versions of this requirement have been stated by requiring, for example, that a stable set *contain* a stable set of the game obtained by deleting such a strategy, or by looking at equilibrium outcomes in generic extensive form games, without being explicit about what out of equilibrium behavior is associated with the solution.

There is one aspect of this requirement that has aroused some comment in the literature. It is that in deleting dominated strategies we delete weakly dominated strategies. In deleting strategies that are inferior in the solution we delete only strategies that are strictly dominated at all points of the solution. In the way we have motivated the definition this difference appears quite naturally. The weakly dominated strategy is deleted because the player would not in any case be willing to choose it. In order to be sure that the player would not anywhere in the solution be willing to choose the strategy it is necessary that the strategy be strictly dominated at all points of the solution. Another way of saying almost the same thing is that the idea that players should not play dominated strategies might be thought of as motivated by a requirement that the solution should not depend on the player being absolutely certain that the others play according to the solution. Again, the same requirement implies that only strategies that are dominated in some neighbourhood of the solution be deleted.

13.4. **The Definition of Strategic Stability.** There have been a number of definitions of strategic stability proposed, the first and most widely known that of Kohlberg and Mertens (1986). They consider Selten-type perturbations to the

normal form of a game and call a closed set of equilibria strategically stable if it is minimal with respect to the property that all small perturbations have equilibria close to the set. They explicitly viewed this definition as a first pass and were well aware of its shortcomings. In particular the sets so defined might fail to satisfy quite weak versions of backward induction. They give an example, due to Faruk Gul, of a game with such a stable set that does not yield the same outcome as the unique subgame perfect equilibrium. Moreover such sets are often not connected.

A minor modification of this definition somewhat expanding the notion of a perturbation remedies part of these deficiencies. Such a definition essentially expands the definition of a perturbation enough so that backward induction is satisfied while not so much as to violate admissibility, as the definition of fully stable sets by Kohlberg and Mertens does. This modification was known to Kohlberg and Mertens and was suggested in unpublished work of Phil Reny. It was independently discovered and called, following a suggestion of Mertens, fully stable sets of equilibria in Hillas (1990). While remedying some of the deficiencies of the original definition of Kohlberg and Mertens this definition does not satisfy at least one of the requirements given above. The definition defines different sets depending on whether two agents who never both move in a single play of a game are considered as one player or two.

Hillas (1990) also gave a different modification to the definition of strategic stability in terms of perturbations to the best reply correspondence. That definition satisfied the requirements originally proposed by Kohlberg and Mertens, though there is some problem with invariance and an error in the proof of the forward induction properties. This is discussed in Vermeulen, Potters, and Jansen (1994) and Hillas, Jansen, Potters, and Vermeulen (1996a). Related definitions are given in Vermeulen, Potters, and Jansen (1994) and Hillas, Jansen, Potters, and Vermeulen (1996b).

In a series of papers Mertens (1987, 1989, 1991a, and 1992) gave and developed another definition—or, more accurately, a family of definitions—of strategic stability that satisfies all of the requirements we discussed above. The definitions we referred to in the previous paragraph are, in our opinion, best considered as rough approximations to the definitions of Mertens.

Stable sets are defined by Mertens in the following way. One again takes as the space of perturbations the Selten-type perturbations to the normal form. The stable sets are the limits at zero of some connected part of the graph of the equilibrium correspondence above the interior of a small neighbourhood of zero in the space of perturbations. Apart from the minimality in the definition of Kohlberg and Mertens and the assumption that the part of the graph was connected in the definition of Mertens, the definition of Kohlberg and Mertens could be stated in exactly this way. One would simply require that the projection map from the part of the equilibrium correspondence to the space of perturbations be onto.

Mertens strengthens this requirement in a very natural way, requiring that the projection map not only be onto, be also be nontrivial in a stronger sense. The simplest form of this requirement is that the projection map not be homotopic to a map that is not onto, under a homotopy that left the projection map above the boundary unchanged. However such a definition would not satisfy the ordinality we discussed in Section 12.1. Thus Mertens gives a number of variants of the definition all involving coarser notions of maps being equivalent. That is, more maps are trivial (equivalent to a map that isn't onto) and so fewer sets are stable.

These definitions require that the projection map be nontrivial either in homology or in cohomology, with varying coefficient modules. Mertens shows that, with some restrictions on the coefficient modules, such definitions are ordinal.

One also sees in this formulation the similarities of the two aspects of forward induction we discussed in the previous section. In Mertens' work they appear as special cases of one property of stable sets. A stable set contains a stable set of the game obtained by deleting a strategy that is nowhere in the relevant part of the graph of the equilibrium correspondence played with more than the minimum required probability.

Govindan and Mertens (1993) give a definition of stability directly in terms of the best reply correspondence. One can think of the set of equilibria as the intersection of the graph of the best reply correspondence with the diagonal. To define stable sets one looks at the graph of the best reply correspondence in a neighbourhood of the diagonal. It is required that a map that takes points consisting of a strategy and a best reply to that strategy to the strategy space in the following way be essential. The points on the diagonal are projected straight onto the strategy space. Other points are taken to the best reply and then shifted by some constant times the difference between the original strategy vector and the best reply. As long as this constant is large enough the boundary of the neighbourhood will be taken outside the strategy space (at which point the function simply continues to take the boundary value). The form of the essentiality requirement in this paper is that the map be essential in Čech cohomology. Loosely, the requirement says that the intersection of the best reply correspondence with the diagonal should be essential. Govindan and Mertens show that this definition gives the same stable sets as the original definition of Mertens in terms of the graph of the equilibrium correspondence with the same form of essentiality.

13.5. **Strengthening Forward Induction.** In this section we discuss an apparently plausible strengthening of the forward induction requirement that we discussed in Section 11 that is not satisfied by the definitions of stability we have given. Van Damme (1989) suggests that the previous definitions of forward induction do not capture all that the intuition suggests.

Van Damme does not, in the published version of the paper, give a formal definition of forward induction but rather gives

> a (weak) property which in my opinion should be satisfied by any concept that is consistent with forward induction. ... The proposed requirement is that in generic 2-person games in which player i chooses between an outside option or to play a game Γ of which a unique (viable) equilibrium e^* yields this player more than the outside option, only the outcome in which i chooses Γ and e^* is played in Γ is plausible.

He gives the example in Figure 28 to show that strategic stability does not satisfy this requirement. In this game there are three components of normal form perfect equilibria, (BU, L), $\{(T, (q, 1-q, 0) \mid \frac{1}{3} \leq q \leq \frac{2}{3}\}$, and $\{(T, (0, q, 1-q) \mid \frac{1}{3} \leq q \leq \frac{2}{3}\}$. (Player 2's mixed strategy gives the probabilities of L, C, and R, in that order.) All of these components are stable (or contain stable sets) in any of the senses we have discussed. Since (U, L) is the unique equilibrium of the subgame that gives Player 1 more than his outside option the components in which Player 1 plays T clearly do not satisfy van Damme's requirement.

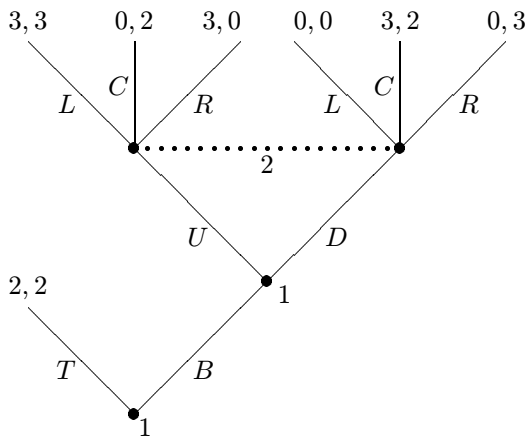


Figure 28

	L	C	R
T	$2, 2$	$2, 2$	$2, 2$
BU	$3, 3$	$0, 2$	$3, 0$
BD	$0, 0$	$3, 2$	$0, 3$

Figure 28a

It isn't clear if van Damme intends to address the question we ask, namely, what outcomes are self-enforcing. (Recall that he doesn't claim that outcomes not satisfying his condition are not self-enforcing, but rather that they are not plausible.) Nevertheless let us address that question. There does not seem, to us, to be a good argument that the outcome in which Player 1 plays T is not self-enforcing. There are two separate "patterns of behavior" by Player 2 that lead Player 1 to play T. To make matters concrete consider the patterns in which Player 2 plays L and C. Let us examine the hypothesis that Player 1 plays T because he is uncertain whether Player 2 will play L or C. If he is uncertain (in the right way) this is indeed the best he can do. Further at the boundary of the behavior of Player 2 that makes T the best for Player 1 there is an assessment of Player 2's choices that make BU best for Player 1 and another assessment of Player 2's choices that make BD best for Player 1. If Player 2 can be uncertain as to whether Player 1 has deviated to BU or to BD then indeed both strategies L and C can be rationalized for Player 2, as can mixtures between them.

To be very concrete, let us focus on the equilibrium $(T, (\frac{1}{3}, \frac{2}{3}, 0)$. In this equilibrium Player 1 is indifferent between T and BD and prefers both to BU. It seems quite consistent with the notion of self-enforcing that Player 2, seeing a deviation by Player 1, should not be convinced that Player 1 had played BU.

In a sense we have done nothing more than state that the stable sets in which Player 1 plays T satisfy the original statement of forward induction given by Kohlberg and Mertens (1986). The argument does however suggest to us that there is something missing in a claim that the stronger requirement of van Damme is a necessary implication of the idea of self-enforcing. To be much more explicit would require a better developed notion of what it means to have set-valued solutions, as we discussed in Section 13.2.

Before leaving this example, we shall address one additional point. Ritzberger (1994) defines a vector field, which he calls the Nash field, on the strategy space. This definition loosely amounts to moving in (one of) the direction(s) of increasing payoff. Using this structure he defines an index for isolated zeros of the field (that is, equilibria). He extends this to components of equilibria by taking the sum of indices of regular perturbations of the Nash field in a small neighbourhood of the component. He points out that in this game the equilibrium (BU, L) has index $+1$, and so since the sum of the indices is $+1$ the index of the set associated with Player 1 playing T must have index zero. However, as we have seen there are two separate stable sets in which Player 1 plays T. Now these two sets are connected by a further set of (nonperfect) equilibria. The methods used by Ritzberger seem to require us to define and index for this component in its entirety. As we said above both of the connected subsets of normal form perfect equilibria are stable in the sense of Mertens' reformulation. And it would seem that there might be some way to define an index so that these sets, considered separately, would have nonzero index, $+1$ and -1. Ritzberger comments that "the reformulation of Stable Sets (Mertens, 1989) eliminates the component with zero index" and speculates that "as a referee suggested, it seems likely that sets which are stable in the sense of Mertens will all be contained in components whose indices are non-zero." However his claim that neither of the sets are stable is in error and his conjecture is incorrect. It seems to us that the methods used by Ritzberger force us to treat two patterns of behavior, that are separately quite well behaved, together and that this leads, in some imprecise sense, to there indices canceling out. There seems something unwarranted in rejecting an outcome on this basis.

13.6. **Forward Induction and Backward Induction.** The ideas behind forward induction and backward induction are closely related. As we have seen backward induction involves an assumption that players assume, even if they see something unexpected, that the other players will choose rationally in the future. Forward induction involves an assumption that players assume, even if they see something unexpected, that the other players chose rationally in the past. However, a distinction between a choice made in the past whose outcome the player has not observed, and one to be made in the future goes very much against the spirit of the invariances that we argued for in Section 12.

In fact, in a much more formal sense there appears to be a relationship between backward and forward induction. In many examples—in fact, in all of the examples we have examined—a combination of the invariances we have discussed and backward induction gives the results of forward induction arguments and, in fact, the full strength of stability. Consider again the game of Figure 23 that we used to motivate the idea of forward induction. The game given in extensive form in Figure 29 and in normal form in Figure 29a has the same reduced normal form as the game of Figure 23. In that game the unique equilibrium of the subgame

is (BU, L). Thus the unique subgame perfect equilibrium of the game is (BU, L). Recall that this is the same result that we argued for in terms of forward induction in Section 11.

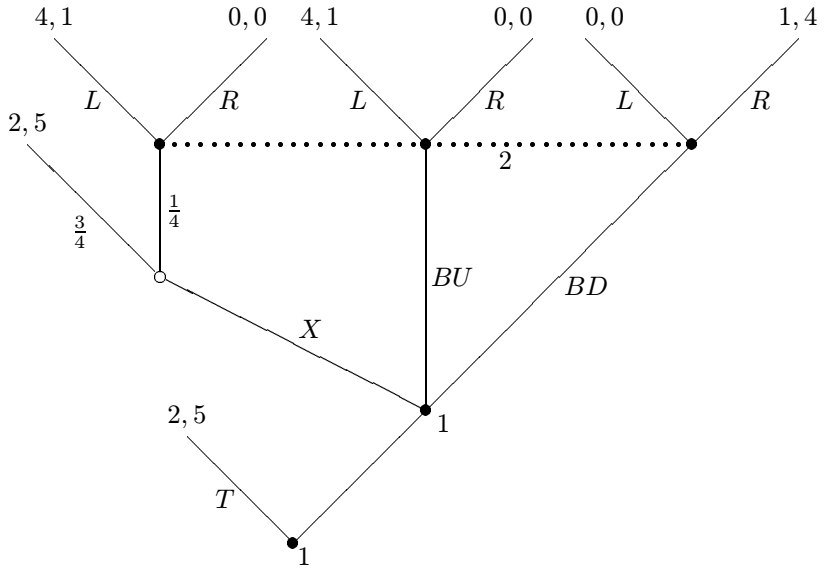


Figure 29

	L	R
T	$2, 5$	$2, 5$
BU	$4, 1$	$0, 0$
BD	$0, 0$	$1, 4$
X	$\frac{5}{2}, 4$	$\frac{3}{2}, \frac{15}{4}$

Figure 29a

This is perhaps not too surprising. After all this was a game originally presented to illustrate in the most compelling way possible the arguments in favor of forward induction. Thus the issues in this example are quite straightforward. Two rounds of the deletion of weakly dominated strategies gives the same result. What is a little surprising is that the same result occurs in cases in which the arguments for forward induction are less clear cut. A number of examples are examined in Hillas (1994) with the same result, including the examples presented by Cho and Kreps (1987) to argue against the strength of some definitions of forward induction and strategic stability.

We shall illustrate the point with one further example. We choose the particular example both because it is interesting in its own right and because in the example

the identification of the stable sets seems to be driven mostly by the definition of stability and not by an obvious forward induction logic. The example is from Mertens (1991b). It is a game with perfect information and a unique subgame perfect equilibrium—though the game is not generic. Moreover there are completely normal form arguments that support the unique subgame perfect equilibrium. In spite this the unique strategically stable set of equilibria consists of all the admissible equilibria, a set containing much besides the subgame perfect equilibrium. The extensive form of the game is given here in Figure 30 and the normal form in Figure 30a.

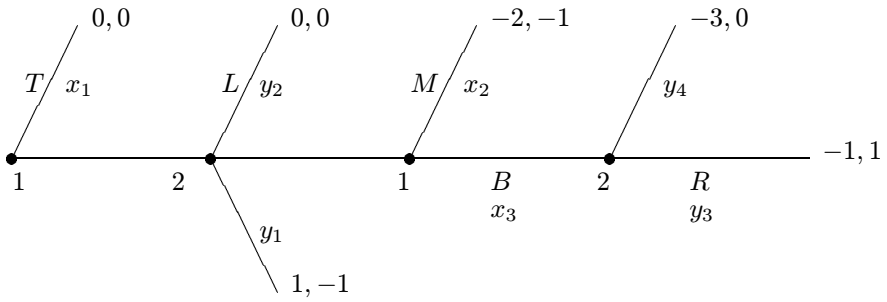


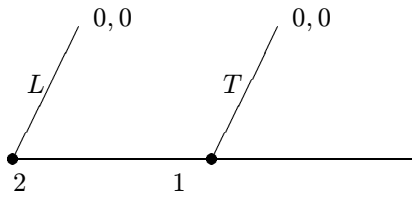
Figure 30

			L		R
		y_1	y_2	y_3	y_4
T	x_1	$0, 0$	$0, 0$	$0, 0$	$0, 0$
M	x_2	$1, -1$	$0, 0$	$-2, -1$	$-2, -1$
B	x_3	$1, -1$	$0, 0$	$-1, 1$	$-3, 0$

Figure 30a

The unique subgame perfect equilibrium of this game is (T, R). Thus this is also the unique proper equilibrium—though the uniqueness of the proper equilibrium does involve an identification of duplicate strategies that is not required for the uniqueness of the subgame perfect equilibrium—and so (T, R) is a sequential equilibrium of any extensive form game having this normal form. Nevertheless there are games having the same *reduced* normal form as this game for which (T, R) is not sequential, and moreover the outcome of any sequential equilibrium of the game is not the same as the outcome from (T, R).

Suppose that the mixtures $X = 0.9T + 0.1M$ and $Y = 0.86L + 0.1y_1 + 0.04y_4$ are added as new pure strategies. This results in the normal form game of Figure 31a. The extensive form game of Figure 31 has this normal form. It is straightforward to see that the unique equilibrium of the simultaneous move subgame is (M, Y) in which the payoff to Player I is strictly positive (0.02) and so in any subgame perfect equilibrium Player 1 chooses to play "in"—that is, M, or B, or X—at his first move. Moreover in the equilibrium of the subgame the payoff to Player 2 is strictly negative (-0.14) and so in any subgame perfect equilibrium Player 2 chooses to play L at his first move. Thus the unique subgame perfect equilibrium of this game is (M, L).



	y_1	R	y_4	Y
M	$1, -1$	$-2, -1$	$-2, -1$	$0.02, -0.14$
B	$1, -1$	$-1, 1$	$-3, 0$	$-0.02, -0.1$
X	$0.1, -0.1$	$-0.2, -0.1$	$-0.2, -0.1$	$0.002, -0.014$

Figure 31

	y_1	L	R	y_4	Y
T	$0, 0$	$0, 0$	$0, 0$	$0, 0$	$0, 0$
M	$1, -1$	$0, 0$	$-2, -1$	$-2, -1$	$0.02, -0.14$
B	$1, -1$	$0, 0$	$-1, 1$	$-3, 0$	$-0.02, -0.1$
X	$0.1, -0.1$	$0, 0$	$-0.2, -0.1$	$-0.2, -0.1$	$0.002, -0.014$

Figure 31a

This equilibrium is far from the unique backward induction equilibrium of the original game. Moreover its outcome is also different. Thus, if one accepts the arguments that the solution(s) of a game should depend only on the reduced normal form of the game and that any solution to a game should contain a "backward induction" equilibrium, then one is forced to the conclusion that in this game the solution must contain much besides the subgame perfect equilibrium of the original presentation of the game.

14. An Assessment of the Solutions

We offer here a few remarks on how the various solutions we have examined stand up. As we indicated in Section 12 we do not believe that the kinds of invariance we discussed there should be controversial. (We are aware that for some reason in some circles they are, but have every confidence that as the issues become better understood this will change.)

On the other hand we have less attachment to or confidence in our other maintained assumption, that of stochastic independence. We briefly discussed in the introduction our reasons for adopting that assumption in this chapter. Many of the issues we have discussed here in the context of strategic equilibrium have their analogs for correlated equilibrium. However the literature is less well developed—and our own understanding is even more incomplete.

The most straightforward, and in some ways weakest, requirement beyond that of equilibrium is admissibility. If one is willing to do without admissibility the unrefined strategic equilibrium seems a perfectly good self-enforcing outcome. Another way of thinking about this is given in the work of Fudenberg, Kreps, and Levine (1988). If the players are permitted to have an uncertainty about the game that is of the same order as their strategic uncertainty—that is, their uncertainty about the choices of the others given the actual game—then any strategic equilibrium appears self-enforcing in an otherwise strong manner.

There are a number of different ways that we could define admissibility. The standard way leads to the standard definition of admissible equilibrium, that is, a strategic equilibrium in undominated strategies. A stronger notion takes account of both the strategy being played and the assumption of stochastic independence. In Section 12.1 we gave a definition of an admissible best reply correspondence such that a fixed point of such a correspondence is a normal form perfect equilibrium. Again Fudenberg, Kreps, and Levine give us another way to think about this. If the players have no uncertainty about their own payoffs but have an uncertainty about the payoffs (or rationality) of the others that is of the same order as their strategic uncertainty then any normal form perfect equilibrium is self-enforcing. Reny's (1992a) work makes much the same point.

To go further and get some version of the backward induction properties that we discussed in Section 10 one needs to make stronger informational assumptions. We claimed in Section 10.1 that the idea of an equilibrium being robust to a little strategic uncertainty in the sense that the strategic uncertainty should be when possible of a higher order than the uncertainty about the game led us towards backward induction. At best this was suggestive. However as we saw in Section 13.6 it takes us much further. In combination with the invariances of Section 12 it seems to imply something close to the full strength of strategic stability. As yet we have seen this only in examples, but the examples include those that were constructed precisely to show the unreasonableness of the strength of strategic stability. We regard these examples to be, at worst, very suggestive.

Mertens (1989) described his attitude to the project that led to his work on strategic stability as follows:

> I should say from the outset that I started this whole series of investigations with substantial scepticism. I had (and still have) some instinctive liking for the bruto Nash equilibrium, or a modest refinement like admissible equilibria, or normal-form perfect equilibria. (pp. 582–583)

It is not entirely coincidental that we arrive at a similar set of preferred solutions. Our approaches have much in common and our work has been much influenced by Mertens.

15. Other Approaches

In this chapter we have examined one approach to thinking about the foundations of strategic equilibrium. There are, of course, other approaches. In this section we examine a small selection of these. We first mention very briefly two approaches and then examine two others in just a little more detail.

One approach that is quite different from that adopted here views, either explicitly or implicitly, the common assessment of the way in which the game will be played as resulting from a long series of observations of the way similar games have been played in the past. This kind of work looks at how players will learn from their observations and what the steady states of such learning processes will be. Such work includes Fudenberg and Kreps (1988, 1991a,b) and Fudenberg and Levine (1993a,b).

Somewhat more closely related to our approach is the work of Mailath, Samuelson, and Swinkels (1992, 1993, 1994) on the relation between extensive form structures and strategic form structures. This work shows that many of the considerations that are traditionally thought of as questions relating to the extensive form can be addressed quite explicitly in the strategic form. This work is at least consistent with our argument for reduced normal form invariance.

In the following two sections we briefly examine one example of an epistemic approach to equilibrium and a very small part of the literature on evolutionary approaches to equilibrium.

15.1. **Epistemic Conditions For Equilibrium.** There have been a number of papers relating the concept of strategic equilibrium to the beliefs or knowledge of the players. One of the most complete and successful is by Aumann and Brandenburger (1995). In the context of a formal theory of *interactive belief systems* they find sets of sufficient conditions on the beliefs of the players for strategic equilibrium. In keeping with the style of the rest of this paper we shall not develop the formal theory but shall rather state the results and sketch some of the proofs in an informal style. Of course, we can do this with any confidence only because we know the formal analysis of Aumann and Brandenburger.

There are two classes of results. One concerns the actions or the plans of the players and the other the conjectures of the players concerning the actions or plans of the other players. These correspond to the two interpretations of strategic equilibrium that we discussed in Section 3. The first class consists of a single result. It is quite simple; its intuition is clear; and its proof is, as Aumann and Brandenburger say, immediate.

Theorem 8. *Suppose that each player is rational and knows his own payoff function and the vector of strategies chosen by the others. Then these strategies form a strategic equilibrium.*

The results involving conjectures are somewhat more involved. The reader is warned that while the results we are presenting are precisely Aumann and Brandenburger's we are presenting them with a quite different emphasis. In fact, the central result we present next appears in their paper as a comment well after the

central arguments. They give central place to the two results that we give as corollaries (Theorems A and B in their paper).

Theorem 9. *Suppose that each player knows the game being played, knows that all the players are rational, and knows the conjectures of the other players. Suppose also that*

a) *the conjectures of any two players about the choices of a third are the same, and*
b) *the conjecture of any player about the choices of two other players are independent.*

Then the vector of conjectures about each player's choice forms a strategic equilibrium.

The proof is, again, fairly immediate. Suppositions a) and b) give immediately that the conjectures "are" a vector of mixed strategies, σ. Now consider the conjecture of player n' about player n. Player n' knows player n's conjecture, and by a) and b) it is σ. Also player n' knows that player n is rational. Thus his conjecture about player n's choices, that is, σ_n, should put weight only on those choices of player n that maximize n's payoff given n's conjecture, that is, given σ. Thus σ is a strategic equilibrium. □

Corollary 10. *Suppose that there are only two players, that each player knows the game being played, knows that both players are rational, and knows the conjecture of the other player. Then the pair of conjectures about each player's choice forms a strategic equilibrium.*

The proof is simply to observe that for games with only two players the conditions a) and b) are vacuous. □

Corollary 11. *Suppose that the players have a common prior that puts positive weight on the event that a particular game g is being played, that each player knows the game and the rationality of the others, and that it is common knowledge among the players that the vector of conjectures takes on a particular value φ. Then the conjectures of any two players about the choices of a third are the same, and the vector consisting of, for each player n, the conjectures of the others about player n's choice forms a strategic equilibrium.*

We shall not prove this result. It is clear that once it is proved that the conjectures of any two players about the choices of a third are the same and that the conjectures of any player about the choices of two other players are independent the rest of the result follows from Theorem 9. The proof that the conjectures of any two players about the choices of a third are the same is essentially an application of Aumann's (1976) "agreeing to disagree" result. The proof that the conjectures of any player about the choices of two other players are independent involves further work, for which we refer the reader to Aumann and Brandenburger's paper.

15.2. **Evolutionary Approaches.** There is, by now, a large literature on the game theoretic approach to evolutionary biology. This is described in detail in the chapter on "Game Theory and Evolutionary Biology," by Hammerstein and Selten (1994) in this Handbook. We shall not replicate that by going into that area at all here. There is however a smaller literature on models that while evolutionary

are explicitly not biological. Since these models are perhaps of some relevance to applications in economics and other social sciences we mention some of them here.

Jeroen Swinkels in a series of papers has shown that an evolutionary approach may lead to some of the same results as one based on strong assumptions about rationality. Swinkels (1992a, 1992b) considers a definition that while motivated by dynamic considerations is actually static. He modifies the definition of an evolutionarily stable strategy from the biological game theory literature in two ways. First, motivated by the idea of being in an economic environment rather than a biological one he restricts attention to entrant strategies that are best responses to the population that results from their entrance. Secondly, motivated by a desire to avoid, as much as possible, problems of nonexistence and by the idea that evolutionary considerations may not tie down out of equilibrium behavior, he considers sets of strategy profiles rather than single strategy profiles.

Swinkels (1992a) shows that, given some regularity conditions, sets defined in this way satisfy many of the properties we have discussed above. They contain a proper equilibrium and satisfy the forward induction property defined in Kohlberg and Mertens (1986). Swinkels (1992b) shows that, again given regularity conditions, a set that is stable against invasion by this kind of entrant must be stable against the kind of perturbations used by Kohlberg and Mertens to define hyperstability—and hence also those used to define full stability and stability—and those used to define stability in Hillas (1990). Swinkels (1993) shows that for some classes of dynamics, and given some regularity conditions, any set that is asymptotically stable is stable against the perturbations used by Kohlberg and Mertens to define hyperstability.

Nöldeke and Samuelson show that it is possible to use evolutionary models to imply even more. Theirs is a model of both learning and mutation. They find that the limiting distributions of the processes they study satisfy a version—in some senses even a stronger version—of the forward induction requirement, suggested by van Damme, that we discussed in Section 13.6.

References

Robert J. Aumann (1974): "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, 1, 67–96.

Robert J. Aumann (1976): "Agreeing to Disagree," *Annals of Statistics*, 4, 1236–1239.

Robert J. Aumann (1987a): "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1–18.

Robert J. Aumann (1987b): "Game Theory," in *The New Palgrave Dictionary of Economics*, edited by J. Eatwell, M. Milgate, and P. Newman, 460–482, W.W. Norton, New York.

Robert J. Aumann (1987c): "What is Game Theory Trying to Accomplish," in *Frontiers of Economics*, edited by Kenneth J. Arrow and S. Honkapohja, 28–100, Basil Blackwell, Cambridge.

Robert J. Aumann (1990): "Nash Equilibria Are Not Self-Enforcing," in *Economic Decision-Making: Games, Econometrics, and Optimization*, edited by J.J. Gabszewicz, J.-F. Richard, and A.L. Wolsey, 201–206, Elsevier, Amsterdam.

Robert J. Aumann (1992a): "Irrationality in Game Theory," in *Economic Analysis of Markets and Games*, edited by P. Dasgupta, D. Gale, O. Hart, and E. Maskin, 214–227, MIT Press, Cambridge, MA.

Robert J. Aumann (1992b): "Notes on Interactive Epistemology," unpublished.

Robert J. Aumann (1995): "Backward Induction and Common Knowledge of Rationality," *Games and Economic Behavior*, 8, 6–19.

Robert J. Aumann and Adam Brandenburger (1995): "Epistemic Conditions for Nash Equilibrium," *Econometrica*, 63, 1161–1180.

Robert J. Aumann and Michael Maschler (1972): "Some Thoughts on the Minimax Principle," *Management Science*, 18, 54–63.

Robert J. Aumann and Sylvain Sorin (1989): "Cooperation and Bounded Recall," *Games and Economic Behavior*, 1, 5–39.

Jeffrey Banks and Joel Sobel (1987): "Equilibrium Selection in Signaling Games," *Econometrica*, 55, 647–661.

Imre Bárány (1992): "Fair Distribution Protocols or How the Players Replace Fortune," *Mathematics of Operations Research*, 17, 327–340.

Kaushik Basu (1988): "Strategic Irrationality in Extensive Games," *Mathematical Social Sciences*, 15, 247–260.

Kaushik Basu (1990): "On the Non-Existence of Rationality Definition for Extensive Games," *International Journal of Game Theory*, 19, 33–44.

Pierpaolo Battigalli (1993): "Strategic Rationality Orderings and the Best Rationalization Principle," unpublished, Politecnico di Milano, forthcoming *Games and Economic Behavior*.

Pierpaolo Battigalli (1994): "Strategic Independence and Perfect Bayesian Equilibria," unpublished, Politecnico di Milano.

Pierpaolo Battigalli (1995): "On Rationalizability in Extensive Games," unpublished.

Richard Bellman (1957): *Dynamic Programming*, Princeton University Press, Princeton NJ.

Elchanan Ben-Porath (1995): "Rationality, Nash Equilibrium and Backwards Induction in Perfect Information Games," unpublished.

ELCHANAN BEN-PORATH AND EDDIE DEKEL (1992): "Signaling Future Actions and the Potential for Sacrifice," *Journal of Economic Theory*, 57, 36–51.

B. DOUGLAS BERNHEIM (1984): "Rationalizable Strategic Behavior," *Econometrica*, 52, 1007–1028.

B. DOUGLAS BERNHEIM (1986): "Axiomatic Characterizations of Rational Choice in Strategic Environments," *Scandinavian Journal of Economics*, 88(3), 473–488.

KEN BINMORE (1987): "Modeling Rational Players, Part I," *Journal of Economics and Philosophy*, 3, 179–214.

KEN BINMORE (1988): "Modeling Rational Players, Part II," *Journal of Economics and Philosophy*, 4, 9–55.

KEN BINMORE (1990): *Essays on the Foundations of Game Theory*, Basil Blackwell, Cambridge MA.

LAWRENCE BLUME, ADAM BRANDENBURGER, AND EDDIE DEKEL (1989): "An Overview of Lexicographic Choice under Uncertainty," *Annals of Operations Research*, 19, 231–246.

LAWRENCE BLUME, ADAM BRANDENBURGER, AND EDDIE DEKEL (1991a): "Lexicographic Probabilities and Choice Under Uncertainty," *Econometrica*, 59, 61–79.

LAWRENCE BLUME, ADAM BRANDENBURGER, AND EDDIE DEKEL (1991b): "Lexicographic Probabilities and Equilibrium Refinements," *Econometrica*, 59, 81–98.

LAWRENCE E. BLUME AND WILLIAM R. ZAME (1994): "The Algebraic Geometry of Perfect and Sequential Equilibrium," *Econometrica*, 62, 783–794.

TILMAN BÖRGERS (1989): "Bayesian Optimization and Dominance in Normal Form Games," unpublished, University of Basel.

TILMAN BÖRGERS (1991): "On the Definition of Rationalizability in Extensive Games," Discussion Paper 91–22, University College, London.

TILMAN BÖRGERS (1994): "Weak Dominance and Approximate Common Knowledge," *Journal of Economic Theory*, 64, 265–276.

ADAM BRANDENBURGER (1992a): "Knowledge and Equilibrium in Games," *Journal of Economic Perspectives*, 6(4), 83–102.

ADAM BRANDENBURGER (1992b): "Lexicographic Probabilities and Iterated Admissibility," in *Economic Analysis of Markets and Games*, edited by Partha Dasgupta, Douglas Gale, Oliver Hart, and Eric Maskin, MIT Press, Cambridge, MA.

ADAM BRANDENBURGER AND EDDIE DEKEL (1987): "Rationalizability and Correlated Equilibrium," *Econometrica*, 55, 1391–1402.

IN-KOO CHO (1987): "A Refinement of Sequential Equilibrium," *Econometrica*, 55, 1367–1389.

IN-KOO CHO AND DAVID M. KREPS (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102, 179–221.

IN-KOO CHO AND JOEL SOBEL (1990): "Strategic Stability and Uniqueness in Signaling Games," *Journal of Economic Theory*, 50, 381–413.

AUGUSTIN A. COURNOT (1838): *Recherches sur les principes mathématiques de la théorie des richesse*, M. Riviere, Paris, Translated in *Researches into the Mathematical Principles of Wealth,* A. M. Kelly, New York, 1960.

NORMAN DALKEY (1953): "Equivalence of Information Patterns and Essentially Determinate Games," in *Contributions to the Theory of Games, Vol. 2*, edited

by H. W. Kuhn and A. W. Tucker, 127–143, Princeton University Press, Princeton, NJ.

Gerard Debreu (1952): "A Social Equilibrium Existence Theorem," *Proceedings of the National Academy of Sciences*, 38, 886–893.

Eddie Dekel and Drew Fudenberg (1990): "Rational Behavior with Payoff Uncertainty," *Journal of Economic Theory*, 52, 243–267.

Amrita Dhillon and Jean-François Mertens (1991): "Perfect Correlated Equilibria," Discussion Paper DP–91–24, Institute for Decision Sciences, SUNY at Stony Brook.

Susan Elmes and Philip J. Reny (1994): "On the Strategic Equivalence of Extensive Form Games," *Journal of Economic Theory*, 62, 1–23.

Ky Fan (1952): "Fixed-Point and Minimax Theorems in Locally Convex Linear Spaces," *Proceedings of the National Academy of Sciences*, 38, 121–126.

Peter C. Fishburn (1994): "Utility and Subjective Probability," in *Handbook of Game Theory with Economic Applications, Vol. 2*, edited by Robert J. Aumann and Sergiu Hart, Elsevier, Amsterdam.

Françoise Forges (1990): "Universal Mechanisms," *Econometrica*, 58, 1341–1364.

Drew Fudenberg and David M. Kreps (1988): "A Theory of Learning, Experimentation, and Equilibrium in Games," unpublished.

Drew Fudenberg and David M. Kreps (1991a): "Learning and Equilibrium in Games, Part I: Strategic-Form Games," unpublished.

Drew Fudenberg and David M. Kreps (1991b): "Learning Mixed Equilibria," *Games and Economic Behavior*, 5, 320–367.

Drew Fudenberg, David M. Kreps, and David K. Levine (1988): "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory*, 44, 351–380.

Drew Fudenberg and David K. Levine (1993a): "Self-Confirming Equilibrium," *Econometrica*, 61, 523–545.

Drew Fudenberg and David K. Levine (1993b): "Steady State Learning and Nash Equilibrium," *Econometrica*, 61, 547–573.

John Geanakoplos (1992): "Common Knowledge," *Journal of Economic Perspectives*, 6(4), 53–82.

John Geanakoplos (1994): "Common Knowledge," in *Handbook of Game Theory with Economic Applications, Vol. 2*, edited by Robert J. Aumann and Sergiu Hart, Elsevier, Amsterdam.

I. L. Glicksberg (1952): "A Further Generalization of the Kakutani Fixed-Point Theorem with Applications to Nash Equilibrium Points," *Proceedings of the American Mathematical Society*, 3, 170–174.

Srihari Govindan (1992a): "A Backward Induction Property of Stable Equilibria," unpublished.

Srihari Govindan (1992b): "Every Stable Set Contains a Fully Stable Set," unpublished.

Srihari Govindan (1992c): "Stability and the Chain Store Paradox," CORE Discussion Paper 9204.

Srihari Govindan and Andrew McLennan (1995): "A Game Form with Infinitely Many Equilibria on an Open Set of Payoffs," unpublished.

Srihari Govindan and Jean-François Mertens (1993): "An Equivalent Definition of Stable Equilibria," unpublished.

SRIHARI GOVINDAN AND ROBERT WILSON (1995a): "Invariance of the Degree of Nash Equilibria," unpublished.

SRIHARI GOVINDAN AND ROBERT WILSON (1995b): "A Sufficient Condition for Invariance of Essential Components," to appear in *Duke Mathematical Journal,* 1996.

SANFORD GROSSMAN AND MOTTY PERRY (1986): "Perfect Sequential Equilibrium," *Journal of Economic Theory*, 39, 97–119.

PETER HAMMERSTEIN AND REINHARD SELTEN (1994): "Game Theory and Evolutionary Biology," in *Handbook of Game Theory with Economic Applications, Vol. 2*, edited by Robert J. Aumann and Sergiu Hart, Elsevier, Amsterdam.

PETER J. HAMMOND (1993): "Aspects of Rationalizable Behavior," in *Frontiers of Game Theory*, edited by Ken Binmore, Alan Kirman, and Piero Tani, MIT Press, Cambridge MA.

JOHN C. HARSANYI (1967-1968): "Games with Incomplete Information Played by 'Bayesian' Players, I–III," *Management Science*, 14, 159–182, 320–334, 486–502.

JOHN C. HARSANYI (1973): "Games with Randomly Distributed Payoffs: A New Rationale for Mixed Strategy Equilibrium Points," *International Journal of Game Theory*, 2, 1–23.

JOHN C. HARSANYI (1977): *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press, Cambridge.

JOHN C. HARSANYI AND REINHARD SELTEN (1988): *A General Theory of Equilibrium Selection in Games*, MIT Press, Cambridge MA.

SERGIU HART (1992): "Games in Extensive and Strategic Form," in *Handbook of Game Theory with Economic Applications, Vol. 1*, edited by Robert J. Aumann and Sergiu Hart, Elsevier, Amsterdam.

SERGIU HART AND DAVID SCHMEIDLER (1989): "Existence of Correlated Equilibrium," *Mathematics of Operations Research*, 14, 18–25.

JOHN HILLAS (1990): "On the Definition of the Strategic Stability of Equilibria," *Econometrica*, 58, 1365–1390.

JOHN HILLAS (1994): "How Much of 'Forward Induction' Is Implied by 'Backward Induction' and 'Ordinality'?" unpublished, Institute for Decision Sciences, SUNY at Stony Brook.

JOHN HILLAS (1996): "On the Relation Between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games," unpublished.

JOHN HILLAS, MATHIJS JANSEN, JOS POTTERS, AND DRIES VERMEULEN (1996a): "On Stable Equilibria: Errors and Corrections," unpublished.

JOHN HILLAS, MATHIJS JANSEN, JOS POTTERS, AND DRIES VERMEULEN (1996b): "On the Relation Between Some Definitions of Strategic Stability," unpublished.

JOHN HILLAS, DRIES VERMEULEN, AND MATHIJS JANSEN (1996c): "On the Finiteness of Stable Sets," unpublished.

M. J. M. JANSEN, A. P. JURG, AND P. E. M. BORM (1994): "On Strictly Perfect Sets," *Games and Economic Behavior*, 6, 400–415.

SHIZUO KAKUTANI (1941): "A Generalization of Brouwer's Fixed-Point Theorem," *Duke Mathematics Journal*, 8, 457–459.

EHUD KALAI AND DOV SAMET (1984): "Persistent Equilibria in Strategic Games," *International Journal of Game Theory*, 13, 129–144.

ELON KOHLBERG (1990): "Refinement of Nash Equilibrium: The Main Ideas," in *Game Theory and Applications*, edited by Tatsuro Ichiishi, Abraham Neyman, and Yair Tauman, Academic Press, San Diego CA.

ELON KOHLBERG AND JEAN-FRANÇOIS MERTENS (1982): "On the Strategic Stability of Equilibria," CORE Discussion Paper 8248, Université Catholique de Louvain, Louvain-la- Neuve, Belgium.

ELON KOHLBERG AND JEAN-FRANÇOIS MERTENS (1986): "On the Strategic Stability of Equilibria," *Econometrica*, 54, 1003–1038.

ELON KOHLBERG AND PHILIP RENY (1992): "On the Rationale for Perfect Equilibrium," Harvard Business School Working Paper 92–011.

ELON KOHLBERG AND PHILIP RENY (1994): "An Interpretation of Consistent Assessments," unpublished.

DAVID M. KREPS (1986): "Out of Equilibrium Beliefs and Out of Equilibrum Behavior," unpublished.

DAVID M. KREPS (1987): "Nash Equilibrium," in *The New Palgrave Dictionary of Economics*, edited by J. Eatwell, M. Milgate, and P. Newman, W.W. Norton, New York.

DAVID M. KREPS (1990): *Game Theory and Economic Modeling*, Oxford University Press, New York, NY.

DAVID M. KREPS, PAUL MILGROM, JOHN ROBERTS, AND ROBERT WILSON (1982): "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory*, 27, 245–252.

DAVID M. KREPS AND GAREY RAMEY (1987): "Structural Consistency, Consistency, and Sequential Rationality," *Econometrica*, 55, 1331–1348.

DAVID M. KREPS AND ROBERT WILSON (1982): "Sequential Equilibria," *Econometrica*, 50, 863–894.

HAROLD W. KUHN (1953): "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games, Vol. 2*, 193–216, Princeton University Press, Princeton NJ.

R. DUNCAN LUCE AND HOWARD RAIFFA (1957): *Games and Decisions: Introduction and Critical Survey*, John Wiley & Sons, New York.

GEORGE J. MAILATH, LARRY SAMUELSON, AND JEROEN M. SWINKELS (1993): "Extensive Form Reasoning in Normal Form Games," *Econometrica*, 61, 273–302.

GEORGE J. MAILATH, LARRY SAMUELSON, AND JEROEN M. SWINKELS (1994): "Normal Form Structures in Extensive Form Games," *Journal of Economic Theory*, 64, 325–371.

GEORGE J. MAILATH, LARRY SAMUELSON, AND JEROEN M. SWINKELS (1995): "How Proper is Sequential Equilibrium?" unpublished, University of Pennsylvania.

JOHN MAYNARD SMITH AND GEORGE R. PRICE (1973): "The Logic of Animal Conflict," *Nature*, 246, 15–18.

ANDREW MCLENNAN (1985a): "Justifiable Belieefs in Sequential Equilibrium," *Econometrica*, 53, 889–904.

ANDREW MCLENNAN (1985b): "Subform Perfection," unpublished, Cornell University.

ANDREW MCLENNAN (1995): "Invariance of Essential Sets of Nash Equilibria," unpublished.

JEAN-FRANÇOIS MERTENS (1987): "Ordinality in Non Cooperative Games," CORE Discussion Paper 8728, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

JEAN-FRANÇOIS MERTENS (1989): "Stable Equilibria—A Reformulation, Part I: Definition and Basic Properties," *Mathematics of Operations Research*, 14, 575–624.

JEAN-FRANÇOIS MERTENS (1991a): "Equilibrium and Rationality: Context and History Dependence," in *Issues in Contemporary Economics, Vol. 1: Markets and Welfare, Proceedings of the Ninth World Congress of the International Economic Association, I.E.A. Conference Vol. 98*, edited by Kenneth J. Arrow, 198–211, MacMillan Co., New York.

JEAN-FRANÇOIS MERTENS (1991b): "Stable Equilibria—A Reformulation, Part II: Discussion of the Definition and Further Results," *Mathematics of Operations Research*, 16, 694–753.

JEAN-FRANÇOIS MERTENS (1992): "The Small Worlds Axiom for Stable Equilibria," *Games and Economic Behavior*, 4, 553–564.

JEAN-FRANÇOIS MERTENS (1995): "Two Examples of Strategic Equilibria," *Games and Economic Behavior*, 8, 378–388.

JEAN-FRANÇOIS MERTENS AND SHMUEL ZAMIR (1985): "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 14, 1–29.

HERVÉ MOULIN (1979): "Dominance Solvable Voting Schemes," *Econometrica*, 47, 1337–1351.

HERVÉ MOULIN AND J.-P. VIAL (1978): "Strategically Zero-Sum Games: The Class of Games Whose Completely Mixed Equilibria Cannot Be Improved Upon," *International Journal of Game Theory*, 7(3/4), 1337–1351.

ROGER MYERSON (1978): "Refinement of the Nash Equilibrium Concept," *International Journal of Game Theory*, 7, 73–80.

ROGER MYERSON (1986a): "Acceptable and Predominant Correlated Equilibria," *International Journal of Game Theory*, 15, 133–154.

ROGER MYERSON (1986b): "Multistage Games with Communication," *Econometrica*, 54, 323–358.

ROGER MYERSON (1994): "Communication, Correlated Equilibria, and Incentive Compatibility," in *Handbook of Game Theory with Economic Applications, Vol. 2*, edited by Robert J. Aumann and Sergiu Hart, 827–847, Elsevier, Amsterdam.

JOHN NASH (1950): "Equilibrium Points in N-Person Games," *Proceedings of the National Academy of Sciences*, 36, 48–49.

JOHN NASH (1951): "Non-Cooperative Games," *Annals of Mathematics*, 54, 286–295.

ROBERT F. NAU AND KEVIN F. MCCARDLE (1990): "Coherent Behavior in Noncooperative Games," *Journal of Economic Theory*, 50, 424–444.

ABRAHAM NEYMAN (1992): "Bounded Common Knowledge Justifies Cooperation in the Finitely Repeated Prisoners' Dilemma," unpublished.

GEORG NÖLDEKE AND LARRY SAMUELSON (1993): "An Evolutionary Analysis of Backward and Forward Induction," *Games and Economic Behavior*, 5, 425–454.

AKIRA OKADA (1981): "On Stability of Perfect Equilibrium Points," *International Journal of Game Theory*, 10, 67–73.

MARTIN OSBORNE (1990): "Signaling, Forward Induction, and Stability in Finitely Repeated Games," *Journal of Economic Theory*, 50, 22–36.

DAVID G. PEARCE (1984): "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52, 1029–1050.

PHILIP J. RENY (1992a): "Backward Induction, Normal Form Perfection and Explicable Equilibria," *Econometrica*, 60, 627–649.

PHILIP J. RENY (1992b): "Rationality in Extensive Form Games," *Journal of Economic Perspectives*, 6(4), 103–118.

PHILIP J. RENY (1993): "Common Belief and the Theory of Games with Perfect Information," *Journal of Economic Theory*, 59, 257–274.

KLAUS RITZBERGER (1994): "The Theory of Normal Form Games from the Differentiable Viewpoint," *International Journal of Game Theory*, 23, 207–236.

ROBERT ROSENTHAL (1981): "Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox," *Journal of Economic Theory*, 25, 92–100.

ARIEL RUBINSTEIN (1989): "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge'," *American Economic Review*, 79, 385–391.

ARIEL RUBINSTEIN (1991): "Comments on the Interpretation of Game Theory," *Econometrica*, 59, 909–924.

DOV SAMET (1993): "Hypothetical Knowledge and Games with Perfect Information," unpublished, Tel-Aviv University.

LARRY SAMUELSON (1992): "Dominated Strategies and Common Knowledge," *Games and Economic Behavior*, 4, 284–313.

REINHARD SELTEN (1965): "Speiltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit," *Zeitschrift fur die gesamte Staatswissenschaft*, 121, 301–324, 667–689.

REINHARD SELTEN (1975): "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4, 25–55.

REINHARD SELTEN (1978): "The Chain-Store Paradox," *Theory and Decision*, 9, 127–159.

SYLVAIN SORIN (1992): "Information and Rationality: Some Comments," *Annales D'economie et de Statistique*, 25/26, 315–325.

ROBERT STALNAKER (1994): "On the Evaluation of Solution Concepts," *Theory and Decision*, 37, 49–73.

JEROEN M. SWINKELS (1992a): "Evolutionary Stability with Equilibrium Entrants," *Journal of Economic Theory*, 57, 306–332.

JEROEN M. SWINKELS (1992b): "Stability and Evolutionary Stability: from Maynard Smith to Kohlberg and Mertens," *Journal of Economic Theory*, 57, 333–342.

JEROEN M. SWINKELS (1993): "Adjustment Dynamics and Rational Play in Games," *Games and Economic Behavior*, 5, 455–484.

JEROEN M. SWINKELS (1994): "Independence for Conditional Probability Systems," unpublished, Northwestern University.

TOMMY CHIN-CHIU TAN AND SÉRGIO RIBEIRO DA COSTA WERLANG (1988): "The Bayesian Foundations of Solution Concepts of Games," *Journal of Economic Theory*, 45, 370–391.

F. B. THOMPSON (1952): "Equivalence of Games in Extensive Form," RM 759, The Rand Corporation.

ERIC VAN DAMME (1984): "A Relation Between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games," *International Journal of Game Theory*, 13, 1–13.

ERIC VAN DAMME (1989): "Stable Equilibria and Forward Induction," *Journal of Economic Theory*, 48, 476–496.

ERIC VAN DAMME (1991): *Stability and Perfection of Nash Equilibria*, Springer-Verlag, Berlin, 2nd edition.

ERIC VAN DAMME (1992): "Refinement of Nash Equilibrium," in *Advnaces in Economic Theory, Volume 1*, edited by J. J. Laffont, Cambridge University Press, Cambridge.

ERIC VAN DAMME (1994): "Strategic Equilibrium," in *Handbook of Game Theory with Economic Applications, Vol. 3*, edited by Robert J. Aumann and Sergiu Hart, Elsevier, Amsterdam.

A. J. VERMEULEN AND M. J. M. JANSEN (1995): "Are Strictly Perfect Equilibria Proper? A counterexample," Report 9503, Department of Mathematics, University of Nijmegen.

A. J. VERMEULEN, J. A. M. POTTERS, AND M. J. M. JANSEN (1996): "On Quasi-Stable Sets," *International Journal of Game Theory*, 25, 34–49.

A.J. VERMEULEN AND M.J.M. JANSEN (1994): "On the Invariance of Solutions of Finite Games," unpublished.

A.J. VERMEULEN, M.J.M. JANSEN, AND J.A.M. POTTERS (1994a): "On a Method to Make Solution Concepts Invariant," unpublished.

A.J. VERMEULEN, J.A.M. POTTERS, AND M.J.M. JANSEN (1994b): "On Stable Sets of Equilibria," unpublished.

JOHN VON NEUMANN AND OSCAR MORGENSTERN (1944): *Theory of Games and Economic Behavior*, Princeton University Press, Princeton NJ.

WU WEN-TSÜN AND JAING JIA-HE (1962): "Essential Equilibrium Points of n-Person Non-Cooperative Games," *Scientia Sinica*, 10, 1307–1322.

ROBERT WILSON (1992): "Computing Simply Stable Equilibria," *Econometrica*, 60, 1039–1070.

ROBERT WILSON (1995): "Admissibility and Stability," unpublished, Stanford University.

E. ZERMELO (1913): "Über eine Anwendungen der Mengenlehre auf die Theorie der Schachspiels," in *Proceedings of the International Fifth Congress of Mathematicians, Cambridge, 1912, Vol. II*, 501–504, Cambridge, Cambridge University Press.