Eric van Damme

Stability and Perfection of Nash Equilibria

Second, Revised and Enlarged Edition

With 105 Figures

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona Budapest

1 Introduction

In this chapter, we illustrate (by means of a series of examples) why the Nash equilibrium concept has to be refined and several refinements which have been proposed in the literature are introduced in an informal way. (No formal definitions are given in this chapter). First, in Sect. 1.1, it is motivated why the solution of a noncooperative game has to be a Nash equilibrium. In Sects. 1.2–1.4, we consider games in extensive form and discuss the following refinements of the Nash equilibrium concept: subgame perfect equilibria, sequential equilibria, perfect equilibria and stable equilibria. In Sects. 1.5 and 1.6, we consider refinements of the Nash equilibrium concept for normal form games, such as perfect equilibria, proper equilibria, persistent equilibria, essential equilibria and regular equilibria.

1.1 Informal Description of Games and Game Theory

In this section, an informal description of a (strategic) game and of Game Theory is given. For a thorough introduction to Game Theory, the reader is referred to Luce and Raiffa [1957], Owen [1968], Harsanyi [1977], Rosenmüller [1981], Shubik [1983], Moulin [1985] or Friedman [1986].[1]

Game Theory is a mathematical theory which deals with conflict situations. A conflict situation (game) is a situation in which two or more individuals (players) interact and thereby jointly determine the outcome. Each participating player can partially control the situation, but no player has full control. Each player has certain personal preferences over the set of possible outcomes and strives to obtain that outcome which is most profitable to him. It is assumed that these preferences can be described by a von Neumann-Morgenstern utility function, hence, that each player is characterized by a numerical function whose expected value he tries to maximize.

Game Theory is a *normative theory*: it aims to prescribe what each player in a game should do in order to promote his interests optimally, i.e. which strategy each player should play such that his partial influence on the situation benefits him most. Hence, the aim of Game Theory is to provide a solution (i.e. to give a characterization (definition) of "rational behavior") for every game.

The foundation of Game Theory was laid in an article by John von Neumann in 1928 (von Neumann [1928]), but the theory received widespread attention only after the publication of the fundamental book von Neumann and Morgenstern [1947]. In this book, the aim of Game Theory is described as follows:

(...) We wish to find the mathematically complete principles which define "rational behavior" for the participants in a social economy, and to derive from them the general characteristics of that behavior. And while the principles ought to be perfectly general – i.e., valid in all situations – we may be satisfied if we can find solutions, for the moment, only in some characteristic special cases.

von Neumann and Morgenstern [1947], p. 31

Traditionally, games have been divided into two classes: cooperative games and noncooperative games. In this book, we restrict ourselves to noncooperative games. A *noncooperative game* is a game in which there are no possibilities for communication, correlation or (pre)commitment, except for those that are explicitly allowed by the rules. (Hence, all relevant aspects should be captured by the rules of the game). A solution of such a game is a set of recommendations, which tell each player how to behave in every situation that may arise. This solution should be consistent, i.e. no player should have an incentive to deviate from his recommendation. Hence, a solution must be *self-enforcing*: As long as the others obey their recommendations, it should not be in my interest to deviate. In game theoretic terminology this means that the solution should be a *Nash equilibrium* (Nash [1950], [1951]), i.e. a strategy combination with the property that no layer can gain by unilaterally deviating from it. Let us illustrate this by means of the game of Fig. 1.1.1, which is the so called prisoners' dilemma game, probably the most discussed game of the literature.

The rows of the table represent the possible choices T and B for player 1, the columns represent the choices L and R of player 2. In each cell the upper left entry is the payoff to player 1, while the lower right entry is the payoff to player 2. The rules of the game are as follows: It is a one-shot game (each player has to make a choice just once), the players have to make their choices simultaneously and independently of each other, communication nor binding agreements are possible. (Hence, the picture tells the story.)

The most attractive strategy combination of the game of Fig. 1.1.1 is (T, L). However, a sensible theory cannot prescribe this strategy pair as the solution. Namely, suppose it is suggested to play (T, L). Then each player has an incentive to disobey his recommendation as long as he expects his opponent to obey it: By unilaterally deviating, player 1 obtains 11, which is more than the 10 that he gets if he obeys his recommendation. Hence, the solution (T, L) is inconsistent with the assumption that players try to maximize expected utility. The strategy pair (T, L)

	L	R
T	10 10	0 11
B	11 0	3 3

Fig. 1.1.1. Prisoners' dilemma

is not self-enforcing but self-destabilizing: both players have an incentive to deviate from it. In this game, only (B, R) has the property that no player can gain by unilaterally deviating, hence, only (B, R) is a Nash equilibrium. Therefore, Game Theory has to prescribe (B, R) as the solution in a noncooperative context.

Of course, if binding agreements would be possible, then (T, L) could be the solution: the players could just sign a binding contract that prohibits them from deviating. In this case, however, noncooperative modelling would require that one gives a complete specification of what is physically possible, hence, we would have an entirely different game and probably a different noncooperative solution (also see Harsanyi and Selten [1988], Chap. 1).

Note that in this specific example, communication does not bring anything: As long as commitment is impossible, the solution is (B, R). For general games, however, the solution might be different when communication is allowed. Namely, in the latter case the relevant solution concept is correlated equilibrium (Aumann [1974], for refinements see Myerson [1986a, 1986b]) and not Nash equilibrium. (Actually a correlated equilibrium is nothing but a Nash equilibrium of an extended game in which communication has been modelled in detail). In this book, we will stick to the purely noncooperative context, hence, we will not consider correlated equilibria.[2]

The discussion above clearly shows that the solution of a noncooperative game has to be a Nash equilibrium since every other strategy combination is self-destabilizing if binding agreements are not possible.[3] In general, however, a game may possess more than one Nash equilibrium and, therefore, the *core problem of noncooperative game theory* can be formulated as: given a game with more than one Nash equilibrium, which one of these should be chosen as the solution of the game? This core problem will not be solved in this monograph, but we will show that some Nash equilibria are better qualified to be chosen as the solution than others. Namely, we will show that not every Nash equilibrium has the property of being self-enforcing. The next 5 sections illustrate how such equilibria can arise and how game theoretists have tried to eliminate them. For convenience, Nash equilibria which fail to be self-enforcing will be called 'unreasonable' or 'nonsensible'.

1.2 Dynamic Programming

There are several ways in which a game can be described. One way is to summarize the rules of the game by indicating the choices available to each player, the information a player has when it is his turn to move, and the payoffs each player receives at the end of the game. A game described in this way is referred to as a *game in extensive form* (see Sect. 6.1). Usually, such a game is represented by a tree, following Kuhn [1953]. Another way of representing a game is by listing all the strategies (complete plans of action) each player has available together with the payoffs associated with the various strategy combinations. A game described in this way is called a *game in normal form* (see Sect. 2.1). In Sects. 1.2 – 1.4, we confine ourselves to games in extensive form. Normal form games will be considered in Sects. 1.5 and 1.6.

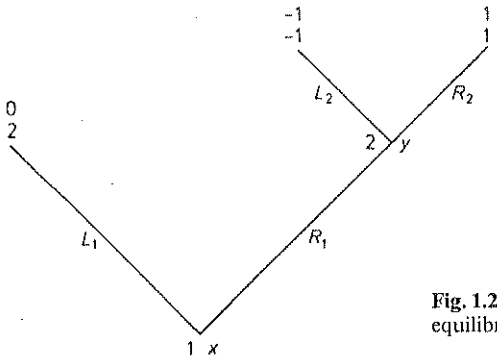

Fig. 1.2.1. An extensive form game with a Nash equilibrium which is not self-enforcing

As an example of a game in extensive form, consider the game of Fig. 1.2.1. The rules of this game are as follows. The game starts at the root x of the tree, where player 1 has to move. He can choose between L_1 and R_1. Player 2, who can choose between L_2 and R_2, has to move only after player 1 has chosen R_1. The payoffs to the players are represented at the endpoints of the tree, the upper number being the payoff to player 1. So, for example, if player 1 chooses L_1, then player 1 receives 0 and player 2 receives 2. The game is played just once.

The game of Fig. 1.2.1 possesses two (pure) Nash equilibria (or shortly equilibria), viz. (L_1, L_2) and (R_1, R_2). The equilibrium (L_1, L_2), however, is not self-enforcing. Namely, suppose the players have agreed to play* (L_1, L_2). If player 1 expects that player 2 will keep to the agreement, then indeed it is optimal for him to play L_1. But should player 1 expect that player 2 will keep to the agreement? The answer is no: since R_2 yields player 2 a higher payoff than L_2 does *if y is reached*, player 2 will play R_2 *if he actually has to make a choice*. Therefore, it is better for player 1 to play R_1 and so he will also violate the agreement and play R_1. So, although (L_1, L_2) is an equilibrium, it is not self-enforcing and, therefore, it is not qualified to be chosen as the solution of the game of Fig. 1.2.1.

Hence, the solution of this game is the equilibrium (R_1, R_2), which is indeed self-enforcing.

The equilibrium (L_1, L_2) of the game of Fig. 1.2.1 can be interpreted as a *threat equilibrium*: player 2 threatens player 1 that he will punish him by playing L_2 if he does not play L_1. Above we argued that this threat is not credible since player 2 will not execute it in the event: Facing the *fait accompli* that player 1 has chosen R_1 it is better for player 2 to play R_2. Note that here we use the basic feature of noncooperative games: no commitments are possible, except from those explicitly allowed by the rules of the game. Notice that the situation changes drastically if player 2 has the possibility to commit himself before the beginning of the game. In this case it is optimal for player 2 to commit himself to L_2, thereby forcing player 1 to play L_1.

To avoid misunderstandings, let us stress again that we do not think that commitments are not possible in conflict situations. We merely hold the view that,

* We will frequently use the phrase "suppose it has been agreed to play..." rather than the clumsier, but more accurate phrase "suppose the game theoretic recommendation is...".

if such commitments are possible, they should explicitly be incorporated in the model (cf. the above discussion of prisoners' dilemma). The great strategic importance of the possibility of committing oneself in games was first pointed out in Schelling [1960].

The game of Fig. 1.2.1 is an example of what is called an *extensive form game with perfect information*. A game is said to have perfect information if the following two conditions are satisfied:

there are no simultaneous moves, and (1.2.1)

at each decision point it is known which choices have previously been made. (1.2.2)

The argument used to exclude the equilibrium (L_1, L_2) in the game of Fig. 1.2.1 generalizes to all games with perfect information: Since in a noncooperative game there are no possibilities for commitment, once the decision point x is reached, the part of the game tree which does not come after x has become strategically irrelevant and, therefore, the decision at x should be based only on that part of the tree which comes after x. This implies that for games with perfect information only those equilibria which can be found by *dynamic programming* (Bellman [1957]), i.e. by inductively working backwards in the game tree, are sensible (i.e. self-enforcing) (cf. Kuhn [1953], Corollary 1).[4]

The game of Fig. 1.2.2 shows that this has the consequence that a sensible equilibrium may be *payoff dominated* by a non-sensible one. The unique equilibrium found by dynamic programming is $(L_1 r_1, R_2)$, i.e. player 1 plays L_1 at his first decision point, r_1 at his second, and player 2 plays R_2. Note that we require a strategy of player 1 to prescribe a choice at his second decision point also in the case in which this player chooses L_1 at his first decision point. The significance of this requirement will become clear in Sect. 1.4.[5] The equilibrium $(L_1 r_1, R_2)$ yields both players a payoff 1. Another equilibrium is $(R_1 l_1, L_2)$. This equilibrium yields both players a payoff 2. This one, however, is not sensible since player 1 cannot commit himself to playing l_1 at his second decision point: both players know that player 1 will play r_1 if this point is actually reached. Therefore, it is illusory of the players to think that they can get a payoff 2. If player 1 chooses R_1, he will end up with a payoff 0.

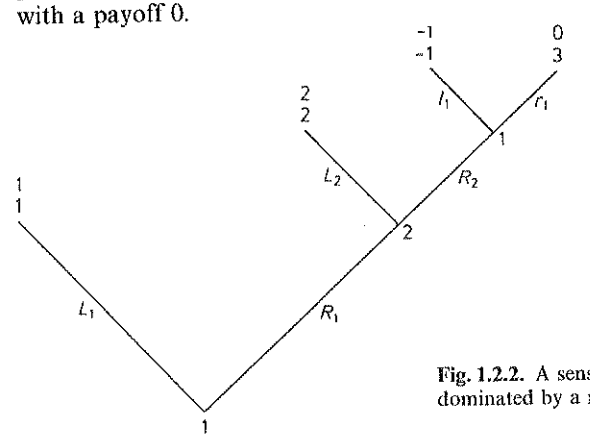

Fig. 1.2.2. A sensible equilibrium may be payoff dominated by a non-sensible one

1.3 Subgame Perfect Equilibria

For games without perfect information one cannot employ the straightforward dynamic programming approach which works so well for games with perfect information. In this section, we will illustrate a slightly more sophisticated dynamic programming approach to exclude equilibria that are not self-enforcing in games without perfect information.

As an example of a game without perfect information, consider the game of Fig. 1.3.1. In this game player 1 cannot discriminate between z and z' (i.e. he does not get to hear whether player 2 has played L_2 or R_2). This is denoted by a dotted line connecting z and z'. The set $\{z, z'\}$ is called an *information set* of player 1.

The straightforward dynamic programming approach fails in this example: in z player 1 should play l_1 and in z' he should play r_1. Hence, he faces a dilemma, since he does not know whether he is in z or in z'. For this game, the more sophisticated approach amounts to nothing else than going one step further backwards in the game tree. Namely, notice that the subtree starting at y constitutes a game of its own, called the *subgame* starting at y. An equilibrium of the game is sensible only if it prescribes an equilibrium also in this subgame. Namely, otherwise at least one player would have an incentive to deviate once the subgame is actually reached. It is easily seen that the subgame has only one equilibrium, viz. (r_1, R_2). Hence, player 1 should play r_1 at his information set $\{z, z'\}$ and player 2 should play R_2. Once this is established, it follows that player 1 should play R_1 at x. Hence, $(R_1 r_1, R_2)$ is the only sensible equilibrium of the game of Fig. 1.3.1. Notice that this is not the only equilibrium: $(L_1 l_1, L_2)$ is also an equilibrium of this game. This equilibrium, however, is not sensible since it involves the incredible threat of player 2 to play L_2.

It was first pointed out explicitly in Selten [1965] that the above argument is valid for every noncooperative game: Since commitments are not possible, behavior in a subgame can depend only on the subgame itself and, therefore, for an equilibrium to be sensible, it is necessary that this equilibrium induces an

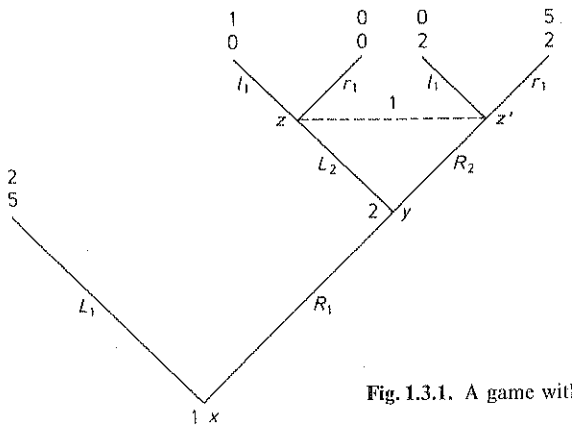

Fig. 1.3.1. A game with imperfect information

	L_2	M_2	R_2
L_1	10 10	0 11	0 0
M_1	11 0	3 3	1 0
R_1	0 0	0 1	0 0

Fig. 1.3.2. A normal form game Γ, which is a slight modification of the game of Fig. 1.1.1

equilibrium in every subgame. Equilibria which possess this property are called *subgame perfect equilibria*, following Selten [1975].

As a second illustration of the subgame perfectness concept, we will now consider a finite repetition of the modified prisoners' dilemma game of Fig. 1.3.2. For general results concerning subgame perfectness in repeated games, the reader is referred to Chap. 8.

Notice that the game Γ results from the game of Fig. 1.1.1 by adding for each player a dominated strategy. Also in Γ the strategy pair (L_1, L_2) is the most attractive one, but this pair is not an equilibrium. The unique equilibrium of Γ is (M_1, M_2).

Now consider the game $\Gamma(2)$, which consists of playing Γ twice in succession. In $\Gamma(2)$ each player tries to maximize the sum of the payoffs he receives at stage 1 and stage 2 and at the second stage each player gets to hear which choices have been made at the first stage.

At the second stage of $\Gamma(2)$ everything which has happened at the first stage had become strategically irrelevant and, therefore, the behavior at stage 2 can depend only on Γ. Hence, at stage 2 the players should play (M_1, M_2), the unique equilibrium of Γ. But, once this has been established, it follows that the players also should play (M_1, M_2) at the first stage. Hence, there is only one subgame perfect equilibrium of $\Gamma(2)$: (M_1, M_2) should be played at both stages.

However, $\Gamma(2)$ has a phletora of equilibria which are not subgame perfect. An example of such an equilibrium is the strategy combination (φ_1, φ_2), where φ_i $(i \in \{1, 2\})$ is given by (1.3.1):

$$\varphi_i \begin{cases} \text{at stage 1:} & \text{play } L_i \\ \text{at stage 2:} & \begin{cases} \text{play } M_i, \text{ if } (L_1, L_2) \text{ has been played at stage 1,} \\ \text{play } R_i, \text{ otherwise.} \end{cases} \end{cases} \tag{1.3.1}$$

In this equilibrium, each player threatens the other that he will be punished at the second stage if he does not cooperate at the first stage. If both players believe the threats, the "cooperative outcome" (L_1, L_2) will result at the first stage. This

equilibrium, however, is not sensible, since a player should not believe the other player's threat. If player 2 plays the strategy φ_2 of (1.3.1), then player 1, knowing that it is not optimal for player 2 to execute the threat, should play M_1 at the first stage.

In the last couple of years, many papers have been published in which the subgame perfectness concept is applied (some earlier applications are Selten [1965, 1973, 1977, 1978] and Ståhl [1972, 1977]). We will return to this concept in Chaps. 7 and 8. In Chap. 7, we will consider a class of games for which the subgame perfectness concept is very effective in excluding unreasonable equilibria. In Chap. 8, however, it will be shown that, for repeated games, the concept is not as restrictive as one might initially think.

1.4 Sequential Equilibria and Perfect Equilibria

It was first pointed out in Selten [1975] that a subgame perfect equilibrium may also prescribe irrational (non-maximizing) behavior at information sets which are not reached when the equilibrium is played. Consequently, a subgame perfect equilibrium need not be sensible. The 3-person game of Fig. 1.4.1, which is taken from Selten [1975], Sect. 6, can illustrate this fact.

Since there are no subgames in the game of Fig. 1.4.1, every equilibrium is subgame perfect (for the formal definition of a subgame see (6.1.16)). One equilibrium of this game is (L_1, R_2, R_3). However, this equilibrium is not sensible since player 2 will violate an agreement to play (L_1, R_2, R_3) in case his information set is actually reached. Namely, if player 2 plays L_2, then player 3 will not find out that the agreement is violated (he cannot discriminate between z and

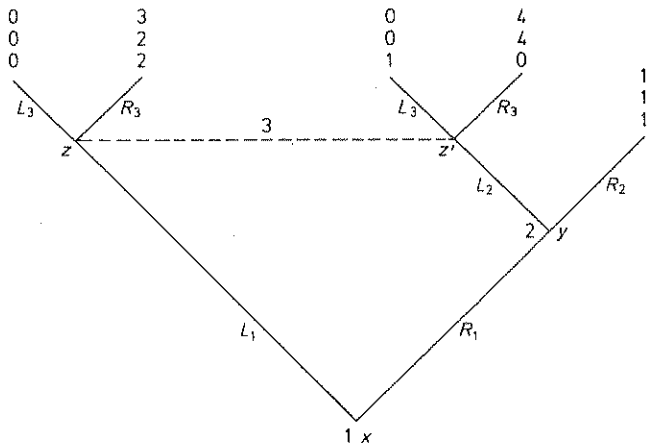

Fig. 1.4.1. A subgame perfect equilibrium need not to be sensible. The numbers at the endpoints of the tree represent the payoffs to the players; the upper number is the payoff to player 1, the second one is the payoff to player 2, etc.

z') and, therefore, this player will still play R_3. Hence, playing L_2 yields player 2 a payoff 4, which is more than R_2 yields and, therefore, this player will play L_2 if his information set is actually reached. Realizing this, player 1 will play R_1 (which yields him a payoff 4), rather than L_1 (which yields only 3). Hence, an agreement to play (L_1, R_2, R_3) is not self-enforcing and, therefore, the equilibrium (L_1, R_2, R_3) is not sensible. (It can be shown that any sensible equilibrium has player 1 playing R_1, player 2 playing R_2, and player 3 playing L_3 with a probability at least 3/4, see Selten [1975]).

The Nash equilibrium concept requires that each player chooses a strategy which maximizes his expected payoff, assuming that the other players will play in accordance with the equilibrium. The reason that the equilibrium (L_1, R_2, R_3) in the game of Fig. 1.4.1 is not sensible is the following: If (L_1, R_2, R_3) is played, the information set of player 2 is not reached and, therefore, the expected payoff of this player does not depend on his own strategy. This obviously implies that every strategy maximizes his expected payoff. However, since player 2 has to move only if the point y is actually reached, he should not let himself be guided by his *a priori* expected payoff, but by his *expected payoff after y*. The a priori expected payoff is based on the assumption that player 1 plays L_1, but, if y is reached this has shown to be wrong and player 2 should incorporate this in computing his expected payoff.

The discussion above shows that, for a subgame perfect equilibrium to be sensible, it is necessary that this equilibrium prescribes at each information set which is a singleton a choice which maximizes the expected payoff after that information set. Note that the restriction to singleton information sets is necessary to ensure that the expected payoff after the information set is well-defined. This restriction, however, has the consequence that not all subgame perfect equilibria which satisfy this additional condition are sensible. This is illustrated by means of the game of Fig. 1.4.2.

A subgame perfect equilibrium of this game which, moreover, satisfies the above condition is (A, R_2). This equilibrium is not sensible since it is always better for player 2 to play L_2 if his information set is reached. (Note that we can draw this conclusion without being able to compute the expected payoff of player 2 after

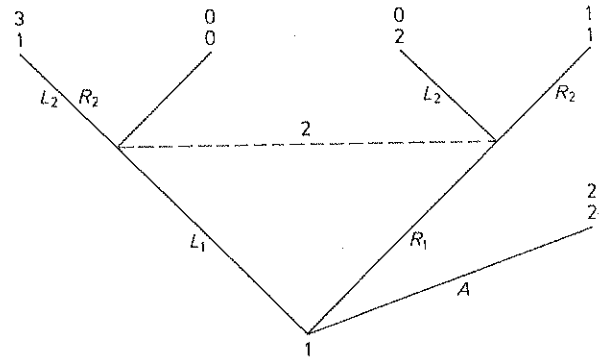

Fig. 1.4.2. An unreasonable subgame perfect equilibrium

his information set.) Realizing this, player 1 should play L_1 and therefore (L_1, L_2) is the only sensible equilibrium of the game of Fig. 1.4.2.

The examples in this section illustrate that a sensible (self-enforcing) equilibrium has to prescribe rational (maximizing) behavior at all information sets, including the ones which can be reached only after a deviation from the equilibrium. The problem, however, is: how should rational behavior at an information set with prior probability zero be defined. (Note that at such a set, if it is not a singleton, the conditional expected payoff is not well-defined.) In the literature two related solutions to this problem have been proposed, one in Selten [1975] (the concept of *perfect equilibria*) and one in Kreps and Wilson [1982a] (the concept of *sequential equilibria*). Let us first explain the concept of sequential equilibria.

The basic assumption underlying the sequential equilibrium concept is that the players are rational in the sense of Savage [1954], i.e. that a player who has to make a choice in the face of uncertainty will construct a personal probability for every event about which he is uncertain and that he will maximize expected utility with respect to these probabilities. To be more precise, suppose the players in an extensive form game have agreed to play an equilibrium φ and assume that a player nevertheless finds himself in an information set which could not be reached when φ is actually played. In this case, the player will try to reconstruct what has gone wrong, i.e. where a deviation from the equilibrium has occurred. In general, this player will not be able to reconstruct completely what has gone wrong and, therefore, he will not be able to tell in which point of his information set he actually is. However, he will represent his uncertainty by a posterior probability distribution on the nodes in this information set (his so called *beliefs* at the information set) and having constructed his beliefs, he will take a choice which maximizes his expected utility with respect to these beliefs, assuming that in the remainder of the game the players will play according to φ. A *sequential equilibrium* is then defined as an equilibrium φ which has the property that, if the players behave as indicated above, no player has an incentive to deviate from φ at any of his information sets. To be more precise: a strategy combination is a sequential equilibrium if there exist (consistent) beliefs such that each player's strategy prescribes at every information set a choice which is optimal with respect to these beliefs (see Def. 6.3.1).

In the game of Fig. 1.4.2 only the equilibrium (L_1, L_2) is a sequential equilibrium. No matter which beliefs player 2 has, it is always optimal for him to play L_2. Note that for an equilibrium to be sequential it is only necessary that it is optimal with respect to *some* beliefs, and that it does not have to be optimal with respect to *all* beliefs or even with respect to the most plausible ones. As we will see at the end of this section, this has the consequence that not every sequential equilibrium is sensible (also see Chaps. 6 and 10).

In Selten [1975] a somewhat different approach is followed to eliminate unreasonable subgame perfect equilibria. Selten assumes that there is always a small probability that a player will take a choice by *mistake*, which has the consequence that every choice will be taken with a positive probability. Therefore, in an extensive form game with mistakes (a so called *perturbed game*) every information set will be reached with a positive probability, which implies that an

equilibrium of such a game will prescribe rational behavior at every information set. The assumption that mistakes occur only with a very small probability leads Selten to define a *perfect equilibrium* as an equilibrium which can be obtained as a limit point of a sequence of equilibria of disturbed games in which the mistake probabilities go to zero. Hence, an equilibrium is perfect if each player's equilibrium strategy is not only optimal against the equilibrium strategies of his opponents, but if it is also optimal against *some* slight perturbations of these strategies (see Def. 6.4.2).

In the game of Fig. 1.4.2 only the equilibrium (L_1, L_2) is perfect. Namely, in a perturbed game associated with this game, player 1 will take the choices L_1 and R_1 with a positive probability (if only by mistake) and, therefore, the information set of player 2 will actually be reached, which forces player 2 to play L_2.

It can be proved that every game possesses at least one perfect equilibrium (Theorem 6.4.4) and that every perfect equilibrium is a sequential equilibrium (see Theorem 6.4.3). However, not every sequential equilibrium is perfect. To illustrate the difference between the two concepts, consider the following slight modification of the game of Fig. 1.4.2: Player 1 receives 3 if he plays A, all other payoffs remain as in Fig. 1.4.2. As before, one can see that player 2 has to play L_2. For player 1, both L_1 and A are best replies against L_2 and, therefore, in a sequential equilibrium player 1 can play any combination of L_1 and A. The only perfect equilibrium, however, is (A, L_2). The reason is that if player 1 plays A he is sure of getting 3, whereas if he plays L_1 he can expect only slightly less than 3 since player 2 with a small probability will make a mistake and play R_2.

In Kreps and Wilson [1982a] it is shown that there is not much difference between the solutions generated by the sequential equilibrium concept and the solutions generated by the perfectness concept. They proved that almost all sequential equilibria are perfect (Kreps and Wilson [1982a] Theorem 3; for a more exact formulation of this result, see Theorem 6.4.3). However, verifying whether a given equilibrium is sequential is easier than to verify whether it is perfect.

Two questions concerning the concepts of sequential and perfect equilibria remain to be answered:

(i) Don't we exclude any sensible equilibria by restricting ourselves to sequential (resp. perfect) equilibria?

(ii) Is every sequential (resp. perfect) equilibrium sensible?

In our view, the first question certainly has to be answered affirmatively for sequential equilibria: if an equilibrium is not sequential, then at least one player has an incentive to deviate from the equilibrium at some of his information sets and, therefore, this equilibrium is not self-enforcing.[6] Whether this question should be answered affirmatively for perfect equilibria depends on one's personal viewpoint of how seriously the possibility of mistakes should be taken.

The second question, however, has to be answered negatively: many perfect (and, hence, sequential) equilibria are not sensible. Loosely speaking this is caused by the fact that some sequential (resp. perfect) equilibria are sustained only by implausible beliefs (resp. implausible mistake probabilities). This may be illustrated by means of the game of Fig. 1.4.3.

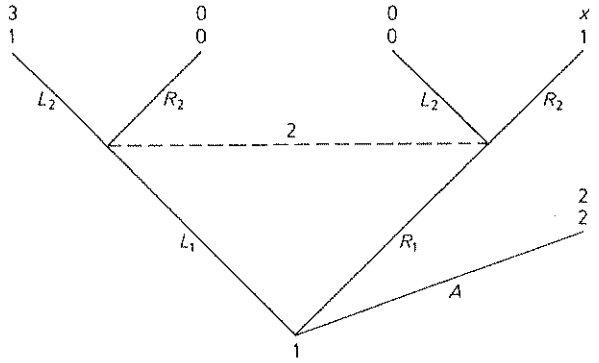

Fig. 1.4.3. Not every perfect (sequential) equilibrium is sensible

We claim that (A, R_2) is a perfect (hence, sequential) equilibrium of the game of Fig. 1.4.3 if $x \leqq 2$. Clearly, this is a Nash equilibrium: No player can gain by deviating unilaterally. Note that player 2 is not reached when A is played. Whenever 2 is reached, he could think that 1 has chosen R_1 and in this case playing R_2 is justified. Since such beliefs of player 2 are not excluded by the sequential equilibrium concept, we have that (A, R_2) is indeed sequential. This also is a perfect equilibrium: If player 2 expects the mistake R_1 to occur with a larger probability than the mistake L_1, then he should indeed play R_2. However, in our view, this equilibrium is not sensible. To support the equilibrium, player 2 should believe that player 1 has chosen R_1, but, if $x \leqq 0$, then R_1 is dominated by both L_1 and A. Hence, player 2 should believe that 1 has chosen his worst strategy and this does not seem reasonable. (If $x \leqq 0$, then (A, R_2) can be eliminated by invoking properness (see Sect. 1.5): R_1 is a more costly mistake than L_1 is, hence, according to properness the mistake L_1 should be considered much more likely.) If $0 < x \leqq 2$, then R_1 is no longer dominated by L_1, but it is still dominated by A. Note that L_1 is not dominated. Hence, if one accepts that a rational player will not choose a dominated strategy, then it follows that player 2 (whenever he is reached) should conclude that player 1 has chosen L_1. Consequently, 2 should choose L_2 and the only sensible equilibrium is (L_1, L_2).

The above example shows that there exist sequential equilibria that can be supported only by implausible beliefs and this raises the question whether one can devise general criteria to eliminate such equilibria. An affirmative answer to this question will be given in Chap. 10: Such equilibria can be eliminated by requiring that the equilibrium path should be stable against all small perturbations in the equilibrium strategies. (This stability criterion has first been proposed in Kohlberg and Mertens [1986]; it is easily seen that only (L_1, L_2) is stable in the game of Fig. 1.4.3.)

1.5 Perfect, Proper and Persistent Equilibria

If we have a game in which each player has to make a choice just once and if, moreover, the players make their choices simultaneously and independently of each other, then we speak of a *normal form game*. An example of such a game is the prisoners' dilemma game of Fig. 1.1.1. A normal form game can be considered as a special kind of extensive form game, but, on the other hand, with each extensive form game, one can associate a game in normal form (von Neumann and Morgenstern [1944], Kuhn [1953]). In the next two sections, it will be shown that also for normal form games it is necessary to refine the Nash equilibrium concept in order to obtain sensible solutions and in several examples the refinements which have been proposed for this class of games will be illustrated. These refinements will be of a slightly different kind than the ones we considered for games in extensive form. Namely, for extensive form games, the basic reason why one has to refine the equilibrium concept is that a Nash equilibrium may prescribe irrational behavior at unreached parts of the game tree. In a normal form game, however, every player has to make a choice, so that there are no unreached information sets. Yet, we will see that it is necessary to refine the equilibrium concept for normal form games, due to the fact that *an equilibrium* of such a game *need not be robust*. As an example of an equilibrium which is not robust consider the game of Fig. 1.5.1.

This game has two equilibria: the strategy combinations (L_1, L_2) and (R_1, R_2). In our view, the latter equilibrium is not a sensible one. This strategy combination satisfies Nash's equilibrium condition only since this condition presumes that each player will completely ignore all parts of the payoff matrix to which his opponent's strategy assigns zero probability. We feel, however, that a player should not ignore this information and that he, therefore, should play L. To be sure, if player 2 plays R_2, then player 1 cannot gain by playing L_1. However, by doing so, he cannot loose either and, as a matter of fact, if player 2 by mistake would play L_2, then player 1 is actually better off by playing L_1. Similarly, we have that player 2 can only gain by playing L_2. Therefore, if it is proposed to play (R_1, R_2), both players have an incentive to deviate. So, an agreement to play (R_1, R_2) is self-destabilizing and, therefore, this equilibrium is not sensible. The only sensible equilibrium of the game of Fig. 1.5.1 is the perfect equilibrium (L_1, L_2). If it is suggested to play this equilibrium, no player has an incentive whatever to disobey the recommendation.

	L_2	R_2
L_1	1 1	0 0
R_1	0 0	0 0

Fig. 1.5.1. The equilibrium (R_1, R_2) is not robust

	L_2	R_2
L_1	1 1	10 0
R_1	0 10	10 10

Fig. 1.5.2. A perfect equilibrium may be payoff dominated by a non-perfect one

	L_2	R_2	A_2
L_1	1 1	0 0	-1 -2
R_1	0 0	0 0	0 -2
A_1	-2 -1	-2 0	-2 -2

Fig. 1.5.3. Adding dominated strategies may enlarge the set of perfect equilibria

If one takes the possibility of the players making mistakes seriously, then, for normal form games, one can only consider perfect equilibria as being reasonable. If an equilibrium fails to be perfect it is unstable with respect to small perturbations of the equilibrium and, therefore, at least one player will have an incentive to deviate from it. By restricting oneself to perfect equilibria, however, one may eliminate equilibria with attractive payoffs, as is shown by the game of Fig. 1.5.2.

This game has two equilibria, viz. (L_1, L_2) and (R_1, R_2). The equilibrium (R_1, R_2) yields both players the highest payoff. The game of Fig. 1.5.2 has exactly the same structure as the game of Fig. 1.5.1 (each player can only gain by playing L) and, therefore, in this game, the equilibrium (R_1, R_2) is as unstable as it is in the game of Fig. 1.5.1. If the players expect mistakes to occur with a small probability, then no player can really expect a payoff 10: the only perfect equilibrium is (L_1, L_2).

It was first pointed out in Myerson [1978] that the perfectness concept does not eliminate all intuitively unreasonable equilibria. The game of Fig. 1.5.3, which is a slight modification of the example given by Myerson, can serve to demonstrate this. Notice that this game results from the game of Fig. 1.5.1 by adding for each player a strategy A. One might argue that, since A is strictly dominated by both L and R, this strategy is strategically irrelevant and that, therefore, the games of Fig. 1.5.1. and 1.5.3 have the same sets of reasonable equilibria. Hence, since (L_1, L_2) is the only reasonable equilibrium of the game of Fig. 1.5.1, this equilibrium is

also the unique reasonable equilibrium of the game of Fig. 1.5.3. However, the sets of perfect equilibria do not coincide for these games: in the game of Fig. 1.5.3 also the equilibrium (R_1, R_2) is perfect. Namely, if the players have agreed to play (R_1, R_2) and if each player expects that the mistake A will occur with a larger probability than the mistake L, then it is indeed optimal for each player to play R. Hence, adding strictly dominated strategies may change the set of perfect equilibria.

Myerson considers it to be an undesirable property of the perfectness concept that adding strictly dominated strategies may change the set of perfect equilibria and, therefore, he introduced a further refinement of the perfectness concept: the *proper equilibrium* (Myerson [1978], see Def. 2.3.1). The basic idea underlying the properness concept is that a player will make his mistakes in a more or less rational way, i.e. that he will make a more costly mistake with a much smaller probability than a less costly one, as a consequence of the fact that he will try much harder to prevent a more costly mistake.

According to the philosophy of the properness concept, in the game of Fig. 1.5.3, the players should not expect the mistake A to occur with a larger probability than the mistake L: since A is strictly dominated by L, each player will try harder to prevent the mistake A than he will try to prevent the mistake L and as a result A will occur with a smaller probability than L (in Myerson's view, the probability of A will even be of smaller order than the probability of L (cf. Def. 2.3.1)). If indeed the mistake L occurs with a larger probability than the mistake A, then each player will prefer to play L, hence, the equilibrium (R_1, R_2) is not proper. The only proper equilibrium of the game of Fig. 1.5.3 is (L_1, L_2): Once the players have agreed to play (L_1, L_2), no player has an incentive whatever to deviate from the equilibrium.

Myerson has shown that every normal form game possesses at least one proper equilibrium and that every proper equilibrium is perfect (Myerson [1978], see Theorem 2.3.3). A problem concerning this concept is that it is not clear whether the basic assumption underlying it (a more costly mistake is chosen with a probability which is of smaller order than the probability of a less costly one) can be justified. Myerson himself did not give a justification for this assumption. In Chaps. 4 and 5, we will investigate whether this assumption can be justified.

The game of Fig. 1.5.4 shows that not all proper equilibria possess the same degree of robustness. This game has several equilibria. Three of these are (L_1, L_2), (M_1, M_2) and (R_1, R_2). It is easily seen that the equilibrium (R_1, R_2) is not perfect: if mistakes might occur, each player will prefer M to R. The equilibria (L_1, L_2) and (M_1, M_2) are both perfect and even proper, but, the equilibrium (L_1, L_2) is much more robust than the equilibrium (M_1, M_2). Namely, if the players have agreed to play (L_1, L_2), then as long as mistakes occur with a probability smaller than $\frac{1}{2}$, each player is still willing to choose L. However, if the players have agreed to play (M_1, M_2), then each player is willing to keep to the agreement only if he expects that the mistake R will occur with a probability at least as big as the probability of the mistake L.

In Okada [1981a] a refinement of the perfectness concept, the *strictly perfect equilibrium*, is introduced which is based on the idea that a sensible equilibrium should be stable against *arbitrary slight perturbations* (see Def. 2.2.7). Obviously,

	L_2	M_2	R_2
L_1	2, 2	1, 1	0, 0
M_1	1, 1	1, 1	1, 1
R_1	0, 0	1, 1	1, 1

Fig. 1.5.4. Not all proper equilibria are equally robust

	L_2	M_2	R_2
L_1	1, 1	1, 0	0, 0
R_1	1, 1	0, 0	1, 0

Fig. 1.5.5. A game without strictly perfect equilibria

for the game of Fig. 1.5.4, the equilibrium (L_1, L_2) is strictly perfect whereas (M_1, M_2) is not. At first sight, it does not seem to be unreasonable to require that the solution of a game should be a strictly perfect equilibrium. The game of Fig. 1.5.5 shows that this cannot always be required, since there exist games without strictly perfect equilibria.

In the game of Fig. 1.5.5, every strategy pair in which player 2 plays L_2 is an equilibrium. None of these equilibria is strictly perfect: if player 1 expects that the mistake M_2 will occur more often than the mistake R_2, he should play L_1; if he expects this mistake to occur with a smaller probability, he should play R_1.

One might take the point of view that one really cannot hope for a general solution concept to be single-valued and that, in the game of Fig. 1.5.5, one should be willing to accept the set of all equilibria as the solution. This point of view is taken in Kohlberg and Mertens [1986]. In that paper, a set-valued analogon of the strict perfectness concept is proposed of which it is shown that it produces satisfactory answers in extensive form games as well as in games in normal form.[7] This stability concept of Kohlberg and Mertens will be considered in Chap. 10.

In the concluding example, we illustrate the concept of *persistent equilibria* that has been introduced in Kalai and Samet [1984]. This concept does not play a prominent role in the monograph, but it certainly deserves further study.[8]

The strategy pair in which both players choose $(\frac{1}{2}, \frac{1}{2})$ is an equilibrium of both games of Fig. 1.5.6. In fact, this is a strictly perfect equilibrium of both games. (It is easily seen that a completely mixed equilibrium is always strictly perfect, hence

	L_2	R_2
L_1	1, 1	0, 0
R_1	0, 0	1, 1

a

	L_2	R_2
L_1	1, 0	0, 1
R_1	0, 1	1, 0

b

Fig. 1.5.6. The mixed equilibrium is not persistent in **a**, but it is persistent in **b**

proper and perfect.) Yet, intuitively one would say that the mixed equilibrium of game b is more stable than the mixed equilibrium of game a. If there are small trembles in the equilibrium strategies, then it seems that any dynamic adjustment process[9] will lead to either (L_1, L_2) or (R_2, R_2) in game a. In game b, however, there are no equilibria with smaller support that threaten the interior equilibrium. (In fact, in this game the Brown/Robinson process converges to the interior equilibrium.) The point is that the interior equilibrium is "minimal" in b, but not in a and intuitively one feels that "nonminimal" equilibria are not that stable. It is this idea of minimality that is formalised by the concept of persistent equilibria (see Sect. 2.3): In game a only (L_1, L_2) and (R_1, R_2) are persistent, in b the completely mixed equilibrium is persistent.

1.6 Essential Equilibria and Regular Equilibria

In the previous section, we considered refinements of the Nash equilibrium concept which are based on the idea that a sensible equilibrium should be stable against slight perturbations of the equilibrium strategies. One could also argue that a sensible equilibrium should be stable against *slight perturbations in the payoffs* of the game. Namely, one can maintain that these payoffs can be determined only somewhat inaccurately. A refinement of the equilibrium concept, based on this idea is the *essential equilibrium* concept, introduced in Wu Wen-Tsün and Jiang Jia-He [1962]. An equilibrium φ of a game Γ is said to be essential if every game with payoffs near to Γ has an equilibrium near to φ. Intuitively, it will be clear that an essential equilibrium is very stable. This indeed will be proved in Chap 2, where we will, for instance, show that every essential equilibrium is strictly perfect. As a consequence, not every game possesses an essential equilibrium. Indeed the payoffs in the game of Fig. 1.5.5 can be perturbed in such a way that either L_1 or R_1 is the unique best reply against L_2 and, therefore, this game does not have an essential equilibrium. Hence, we cannot require that a sensible equilibrium should be essential. Moreover, even in games which have essential equilibria, it is not always true that an essential equilibrium should be preferred to a non-essential one. This is illustrated by the game of Fig. 1.6.1.

	L_2	M_2	R_2
L_1	1, 1	0, 0	0, 0
M_1	0, 0	2, 2	2, 2
R_1	0, 0	2, 2	2, 2

Fig. 1.6.1. An essential equilibrium is not necessarily preferable to a non-essential equilibrium

	L_2	R_2
L_1	2, 2	4, 1
R_1	4, 1	3, 3

Fig. 1.6.2. Instability of equilibria in mixed strategies

The unique essential equilibrium of this game is (L_1, L_2). However, M and R are just duplicates of each other and, once it has been agreed to play some combination of M and R, no player has an incentive to deviate. Hence, the essential equilibrium concept is too restrictive.

It should be noted that the set of all equilibria with payoff $(2, 2)$ is an essential set of equilibria and the reader might be inclined to think that the set-valued analogue of essentiality (this is called hyper-stability in Kohlberg and Mertens [1986]) will always lead to reasonable equilibria. This is not the case, however, since hyper-stable sets may include dominated strategies. (If one deletes R_2 in the game of Fig. 1.5.5 then only R_1 is dominated, but any hyper-stable set includes R_1.)

In this chapter, it was forcibly argued that the solution of a noncooperative game has to be self-enforcing and, therefore, a Nash equilibrium. In many examples we have seen that not all Nash equilibria are self-enforcing: there exist equilibria at which at least one player has an incentive to deviate. Now suppose we have an equilibrium at which no player can gain by deviating. Is this equilibrium necessarily self-enforcing? The answer seems to be no: although no player may have an incentive to deviate, it may be the case that no player has an incentive to play his equilibrium strategy either. This situation occurs for equilibria in mixed strategies as is illustrated by means of the game of Fig. 1.6.2.

This game has a unique equilibrium, which is in mixed strategies. The equilibrium strategy of player 1 is $(2/3L_1, 1/3R_1)$ and the equilibrium strategy of player 2 is $(1/3L_2, 2/3R_2)$. The equilibrium yields player 1 a payoff 10/3 and player 2 a payoff 5/3. This equilibrium seems unstable, since, if player 2 plays $(1/3L_2, 2/3R_2)$, then player 1 receives a payoff of 10/3, no matter what he does and, therefore, he can shift to any other strategy without penalty. So, what is his incentive to play his equilibrium strategy? The same remark applies to player 2: if player 1 plays his equilibrium strategy, player 2 receives 5/3 no matter what he does and, therefore, he can also shift to any strategy without penalty.

One could even argue (as is done in Aumann and Maschler [1972, Sect. 2]) that, in the game of Fig. 1.6.2, the players have an incentive to deviate from their equilibrium strategies. Namely, if the equilibrium is played, player 1 receives 10/3, which is just the *maximin value* of this game for player 1, i.e. the payoff which player 1 can *guarantee* himself. However, the equilibrium strategy of player 1 does not guarantee 10/3, it only yields 10/3 if player 2 plays his equilibrium strategy. In order to guarantee 10/3 player 1 should play his *maximin strategy*, which is $(1/3L_1, 2/3R_1)$. So, if player 1 knows that he cannot obtain more that 10/3, why should not he play his maximin strategy which guarantees 10/3? The same remark applies to player 2 and so he also could have an incentive to play his maximin strategy, rather than his equilibrium strategy.

Aumann and Maschler write that they do not know what to recommend in this situation (since the maximin strategies are not in equilibrium), but they prefer the maximin strategies (Aumann and Maschler [1972, Sect. 2]). In Harsanyi [1977] (especially in Sect. 7.7.) it is argued that the players should indeed play their maximin strategies (also see van Damme [1980a]), but Harsanyi has changed his position in favour of the equilibrium strategies (Harsanyi and Selten [1988], Chap. 1).

From the discussion above it will be clear that the instability of equilibria in mixed strategies poses a serious problem. This problem is serious indeed, since many games possess only equilibria in mixed strategies. In Harsanyi [1973a] it is shown that this instability is only a seeming instability. Harsanyi argues that a player can never know the payoffs (utilities) of another player exactly since these are subject to random disturbances, due to stochastic fluctuations in this player's mood or taste. Therefore, a conflict situation, rather than by an ordinary game, is more adequately modelled by a so called *disturbed game*, i.e.. a game in which each player, although knowing his own payoffs exactly, knows the payoffs of the other players only somewhat inexactly. Harsanyi shows that for such a disturbed game every equilibrium is essentially in pure strategies and is, therefore, stable (also see Theorem 5.4.2). Harsanyi also shows that almost every equilibrium of an ordinary normal form game (whether in pure or in mixed strategies) can be obtained as the limit of equilibria of disturbed games, in which the disturbances go to zero, i.e. in which each player's information about the other players' payoffs becomes better and better. Hence, for almost all equilibria in mixed strategies the instability disappears if we take account of the actual uncertainty each player has about the other players' payoffs. Upon a closer investigation (see Theorem 5.6.2) it turns out that the equilibria which are stable in this sense are the *regular equilibria*, which have been introduced in Harsanyi [1973b]. A regular equilibrium is defined as an equilibrium which has the property that the Jacobian of a certain mapping associated with the game evaluated at this equilibrium is nonsingular (see

Def. 2.5.1). These regular equilibria will play a prominent role in the monograph. It will be shown that regular equilibria possess all robustness properties one reasonably can expect equilibria to possess: they are perfect, proper and even strictly perfect (i.e. stable) and essential. (However, a regular equilibrium need not be persistent: All equilibria in the game of Fig. 1.5.5 are regular.)

Unfortunately not all normal form games possess regular equilibria, but it can be shown that for almost all normal form games all equilibria are indeed regular (Theorem 2.6.2). These results indicate that for generic normal form games there is actually little need to refine the Nash equilibrium concept.[10] For extensive form games, however, the situation is quite different as we will see in Chaps. 6 and 10. Almost any nontrivial extensive form game has equilibria that are 'unreasonable'.

Notes

1. Since the publication of the first edition some new good textbooks in game theory have appeared and a couple of others are about to appear. The following cover the whole spectrum from the elementary to the advanced level: Dixit and Nalebuff [1990], Fudenberg and Tirole [1991], Kreps [1990a, b], Myerson [1990] and Rasmusen [1989]. The important role of game theory in industrial organization is illustrated by Tirole [1988], while McMillan [1989] focuses on the use of game theory in management. An overview of some applications of game theory in macroeconomics is given in Persson and Tabelini [1990], and the strategic theory of international trade is surveyed in McMillan [1986]. Good introductory surveys of noncooperative game theory are Tirole [1988, Chapter 11] and Fudenberg and Tirole [1989]. Kohlberg [1989] provides an excellent overview of refinements of Nash equilibrium, i. e. on the topic of this book. A very nice overview of the history and development of game theory is given in Aumann [1987a]. Surveys of all important areas of game theory will appear in Aumann and Hart [1991]. The reader is also urged to consult Aumann [1987b], a survey that focuses on methodology, on the relation between game theory and the real world, and on the question 'what is game theory trying to accomplish?'.
2. Some recent papers dealing with communication in games of complete information are Farrell [1988], Myerson [1989] and Rabin [1990]. These papers incorporate the idea that language has a focal meaning. Farrell [1987], Farrell and Saloner [1988], Matthews [1989] and Matthews and Postlewaite [1989] illustrate the importance of preplay communication in specific settings. Cooper et al. [1989] report experimental results on the role of cheap talk in the battle of the sexes game. For papers dealing with communication in incomplete information games see note 18 to Chap. 10.
3. Bernheim [1984] and Pearce [1984] have argued that for a strategy combination to be 'reasonable' it is not necessary that it is a Nash equilibrium. They proposed the weaker notion of 'rationalizability' as a necessary condition for 'reasonableness'. Also see Aumann [1987c], Battigalli [1990], Bernheim [1986], Gul [1989] and Rubinstein and Wolinsky [1990].

4. The technique of dynamic programming (hence the concept of subgame perfect equilibrium) is based on the assumption of persistent rationality. Backward induction requires that, when a player has evidence that an opponent does not play according to the backward induction solution, he disregards this information and believes that from now on the opponent will always play as backward induction dictates. This assumption has been criticized in Basu [1988, 1990], Bicchieri [1989], Binmore [1987, 1988], Reny [1988a, 1988b] and Rosenthal [1981]. In these papers it is argued that in the game of Fig. 1.2.2 player 2 may justify playing L_2 with positive probability by the argument that since player 1 has played 'irrationally' at his first move he may also play 'irrationally' at his second move.
5. A strategy of player i is usually interpreted as a complete plan of action for this player. However, it also serves as a description of the opponents' expectations of player i's behavior. The strategy concept is scrutinized in Rubinstein [1988].
6. This statement needs qualification. As Kreps and Wilson already pointed out themselves, it is not completely clear that the consistency notion used in the definition of sequential equilibrium is the appropriate one. See the notes to Chap. 6 for further details.
7. The stability concept from Kohlberg and Mertens [1986] was not entirely satisfactory. A modification of the concept was proposed in Mertens [1980a].
8. A related concept of cyclically stable sets has been proposed in Gilboa and Matsui [1989].
9. Recently the questions of whether 'learning' will lead to a Nash equilibrium and of what kind of learning processes will lead to which type of equilibrium have received considerable attention. A sample of the papers in this area is Canning [1989, 1990], Fudenberg and Levine [1990], Fudenberg and Kreps [1988], Friedmann and Rosenthal [1986], Kalai and Lehrer [1990], Milgrom and Roberts [1989a, b] and Selten [1988a].
10. The most refined equilibrium notion that will be discussed in this book is the strict equilibrium, i.e. an equilibrium in pure strategies in which each player actually looses if he deviates unilaterally. Experiments reported in Van Huyck et al. [1988] suggest that not even all strict equilibria are self-enforcing. Consider the 2×2 bimatrix game Γ in which each player chooses between α and β: If both choose α both receive the payoff 1, if both choose β each receives the payoff 2, if choices don't match each receives 0. The results of Van Huyck et al. [1988] suggest that if it is recommended to the players to play (α, α), then even if one gives them all the arguments for why (α, α) is a good equilibrium, the players will (without having to communicate) succeed to jointly deviate to (β, β). Carlsson and Van Damme [1990] present a noncooperative theory that only allows (β, β) as the solution of Γ. In general this theory selects the risk dominant equilibrium (Harsanyi and Selten [1988]) as the unique solution of a 2×2 bimatrix game. Note that there may be a conflict between risk dominance and payoff dominance, see, for example, Aumann [1989] which also argues that not every strict Nash equilibrium is self-enforcing.