A General Theory of Equilibrium Selection in Games

John C. Harsanyi
and
Reinhard Selten

The MIT Press
Cambridge, Massachusetts
London, England

A General Theory of Equilibrium Selection in Games

John C. Harsanyi
and
Reinhard Selten

The MIT Press
Cambridge. Massachusetts
London, England

This book was set in Times New Roman by Asco Trade Typesetting Ltd. in Hong Kong, and printed and bound by Halliday Lithograph in the United States of America.

Library of Congress Cataloging-in-Publication Data

Harsanyi, John C.
A general theory of equilibrium selection in games.

Bibliography: p.
Includes index.
1. Equilibrium (Economics) 2. Cooperation. 3. Game theory. I. Selten, Reinhard. II. Title.
HB145.H38 1988 339.5 87-35700
ISBN 0-262-08173-3

Content

INTRODUCTION 131

DEFINITION OF THE TRACING PROCEDURE 137

COMPUTATIONALLY USEFUL FACTS ABOUT THE TRACING PROCEDURE 181

Foreword

The equilibrium concept of Nash is without doubt the single game-theoretic tool that is most often applied in economics; in recent years, especially, its use has increased dramatically. Together with this increased use has come a growing preoccupation with the philosophical and logical underpinnings of the concept. The current monumental work of John Harsanyi and Reinhard Selten, in the making for close to two decades, is a major contribution to this effort.

An equilibrium in a game is defined as an assignment to each player of a strategy that is optimal for him when the others use the strategies assigned to them. One of the oldest rationales for this concept, advanced already by von Neumann and Morgenstern (1944), is that any normative theory that advises players how to play games must pick an equilibrium in each game. A theory recommending anything other than an equilibrium would be self-defeating, in the sense that a player who believes that the others are following the theory will sometimes be motivated to deviate from it. Note that this holds only if the theory recommends a unique strategy for each player.

In general, a given game may have several equilibria. Yet uniqueness is crucial to the foregoing argument. Nash equilibrium makes sense only if each player knows which strategies the others are playing; if the equilibrium recommended by the theory is not unique, the players will not have this knowledge. Thus it is essential that for each game, the theory selects one unique equilibrium from the set of all Nash equilibria.

Of course the "theory" rationale makes sense only if all the players are advised by the same theory, and by no other theory, and they must be convinced that all will abide by the advice. This could happen if that theory alone were taught at the business (or law) schools that the players attended. An analogy is to industrial standardization, and to conventions such as driving on the right; indeed, such standards and conventions are illustrations of equilibrium selection.

In this book a coherent theory of equilibrium selection is constructed. The difficulties in constructing such a theory are formidable, as anybody reading this book will quickly realize. The major implication, like that of the first heavier-than-air flying machine, is that it can be done. The theory rationale for Nash equilibrium thus acquires a visible, demonstrated foundation.

The authors will probably be the first to acknowledge that their selection theory is not the only possible or reasonable one. Although the theory selects a unique equilibrium, as a theory it need not be unique. Every facet of the theory was carefully thought out; but as in any complex construction project, many decisions were made which, though far from arbitrary, could well have been made in some other way. During the fifteen or twenty years during which the theory was in the making, several of its aspects, both major and minor, were reconsidered and revised. No doubt, future streamlining and other improvements will be welcomed by the authors, and indeed, there is every chance that they themselves will participate in the process.

As a spin-off from demonstrating the feasibility of equilibrium selection, this book develops several new ideas that are important in their own right, quite independently of the selection problem. Prominent among these are the notions of risk dominance and the tracing procedure.

A consequence of the availability of a theory of equilibrium selection is the ability to implement what has been called the Nash program. A game is called cooperative if there is available a mechanism, such as a court, to enforce agreements. In a cooperative game any feasible outcome may be achieved if the players subscribe to the appropriate agreement. In the 1951 paper in which he defined equilibrium, Nash noted that by specifying and explicitly modeling the bargaining process by which agreements may be reached, one can view cooperative games as special instances of noncooperative games. Nash suggested that the originally given cooperative game be analyzed by means of one of the noncooperative games associated with it in this way.

One difficulty with this program is that even when the bargaining process is fully specified and completely modeled, the resulting noncooperative game often has many equilibria that are very different from each other; in this case the Nash program is not very informative. By selecting a particular one of the many equilibria appearing in such models, the Harsanyi-Selten theory removes this difficulty.

The authors have not contented themselves with a purely theoretical construction. They realize that the proof of the pudding is in the eating of it, that a game-theoretic concept cannot be judged solely on the basis of abstract considerations of plausibility but where it leads in applications. Chapters 6 through 9 of the book are devoted to applications, with emphasis on bargaining and multilateral trade.

In summary, the publication of this book constitutes a major event in game theory; it is likely to have an important influence on the discipline itself as well as on its applications to economic and political theory. The authors are to be congratulated for bringing a long and arduous task to a successful conclusion.

Robert Aumann
Jerusalem, Israel

Acknowledgments

Without mentioning names, we want to thank those many game theorists who in the long process of the development of our theory helped us to clarify our ideas by their critical remarks. We are also grateful to four anonymous readers for the MIT Press. We are indebted to Franz Weissing and Michael Mitzkewitz who thoroughly read various preliminary versions of our book and eliminated some of our mistakes. We also want to thank those who gave technical assistance, especially Mary Ann Huisman, who prepared the drawings, and Mary Macgregor, who did most of the computations for chapter 9.

A General Theory of Equilibrium Selection in Games

1 The Need for a New Solution Concept

1.1 Our Solution Concept

The purpose of this book is to propose a new solution concept, primarily defined for noncooperative games but applicable also to cooperative games, because every cooperative game can be remodeled as a bargaining game having the structure of a noncooperative game. For any noncooperative game, including noncooperative bargaining games, our theory always selects *one* equilibrium point as the solution. By reducing cooperative games to noncooperative bargaining games, our approach unifies the theories of cooperative and noncooperative games into one general theory.

1.2 Cooperative and Noncooperative Games

In contrast, in classical game theory, cooperative and noncooperative games are treated quite differently, and the distinction between these two game classes plays a very fundamental role. Nash (1950a, 1951), who first introduced this distinction, defined cooperative games as games that permit *both* free communication and enforceable agreements among the players, in contrast to noncooperative games, which permit *neither* communication nor enforceable agreements.

A binary distinction based on two simultaneous criteria is logically unsatisfactory, however. We cannot define one category as a class of all objects possessing both properties A and B and the other category as a class of all objects possessing neither property. If we do so, then one must ask what about objects having property A but not B, and objects having property B but not A?

It is preferable therefore to use a one-criterion distinction—to define cooperative games simply as those permitting enforceable agreements and noncooperative games as those not permitting them. Certainly, how much communication is allowed among the players is important in many cases, but this turns out to be a less fundamental issue. To illustrate the problem, consider the Prisoner's Dilemma game shown in figure 1.1. (For an explanation of the term "Prisoner's Dilemma," see Luce and Raiffa 1957, pp. 94–95.) In each cell of the payoff table the number in the upper left-hand corner is player 1's payoff, and that in the lower right-hand corner is player 2's. The rows of the table represent player 1's strategies C^* and N^* and the columns represent player 2's strategies C^{**} and N^{**}.

Because this game is completely symmetric between the two players, both

	C**	N**
C*	10 10	-10 11
N*	11 -10	1 1

Figure 1.1

players have positions of equal strength. Therefore it is natural to expect that they will agree on an outcome that yields them equal payoffs—by either choosing the strategy pair $C = (C^*, C^{**})$, which would yield the payoffs (10, 10), or the strategy pair $N = (N^*, N^{**})$, which would yield the payoffs (1, 1). If the game is played as a cooperative game (permitting enforceable agreements), then the players, assuming that they act rationally, will no doubt immediately agree to use the strategy pair C, since C will give them much higher payoffs than N would. Thus $C = (C^*, C^{**})$ may be called the *cooperative solution* of the game.

In contrast, if the game is played as a noncooperative game (i.e., if the players are unable to conclude enforceable agreements), then they cannot do any better than use the strategy pair $N = (N^*, N^{**})$, which may be called the *noncooperative solution.*

To establish this point, we will first show that if enforceable agreements are impracticable, then rational players cannot choose the strategy pair $C = (C^*, C^{**})$. Even if they did agree to use their C-strategies, they could not rationally expect each other to *keep* to this agreement, so any such agreement would be quite pointless. Suppose they were to make such an agreement and expect each other to keep it. Then player 1 would immediately have an incentive to violate this agreement by using strategy N^*, rather than C^*, because N^*, and not C^*, would be his best reply[1] to player 2's expected strategy C^{**}. Likewise player 2 would have an incentive to violate the agreement by using strategy N^{**}, rather than C^{**}, because N^{**}, and not C^{**}, would be his best reply to player 1's expected strategy C^*.

In a noncooperative game the strategy pair C cannot be chosen by rational players because it would be *self-destabilizing*: the fact that one player expects the other to abide by a C-strategy would give him a clear incentive to deviate from C. Our analysis also shows the mathematical reason why C has this undesirable property. The reason is that the two players' C-strategies are not best replies to each other. Rather, the best reply to C^{**} is N^*, and the best reply to C^* is N^{**}.

In contrast, the strategy pair $N = (N^*, N^{**})$ can be readily used by rational players in a noncooperative game because it is *self-stabilizing*: since N^* and N^{**} are mutually the best replies to each other, if the two players for any reason expect each other to use an N-strategy, then both of them will have a clear incentive to make this expectation come true by using N-strategies.

Clearly, in playing this game, the decisive question is whether the players can make enforceable agreements, and it makes little difference whether they are allowed to talk to each other. Even if they are free to talk and to negotiate an agreement, this fact will be of no real help if the agreement has little chance of being kept. An ability to negotiate agreements is useful only if the rules of the game make such agreements binding and enforceable. (In real life, agreements may be enforced externally by courts of law, government agencies, or pressure from public opinion; they may be enforced internally by the fact that the players are simply unwilling to violate agreements on moral grounds and know that this is the case.)

As Nash has already pointed out (1950a, 1951), similar considerations apply to all noncooperative games. Since in such games agreements are not enforceable, rational players will always choose a strategy combination that is *self-stabilizing* in the sense that the players will have some incentive to abide by a strategy combination (or at least will have no incentive not to do so) if they expect all *other* players to abide by it. Mathematically this means that they will always choose a strategy combination with the property that every player's strategy is a best reply to all other players' strategies. A strategy combination with this property is called an *equilibrium (point)*. Nash has also shown that every finite game[2] has at least one equilibrium point (in pure strategies or sometimes only in mixed strategies).

Nevertheless, the definitions of cooperative and noncooperative games are still in need of further clarification. As they stand, they may give the false impression that noncooperative games cannot be used for modeling game situations in which the players are able to make enforceable agreements (or to enter into other firm commitments,[3] e.g., irrevocable promises and threats). As we shall see in section 1.3, it is possible to incorporate self-commitment moves explicitly into the extensive form of a noncooperative game.

We propose therefore to rephrase our definitions as follows. A *non-*

cooperative game is a game modeled by making the assumption that the players are *unable* to make enforceable agreements (or other commitments), except insofar as the *extensive form of the game explicitly gives them an ability to do so*. In contrast, a *cooperative* game is a game modeled by making the assumption that the players are *able* to make enforceable agreements (and possibly other commitments) even if their ability to do so is not shown explicitly by the extensive form of the game.

1.3 Irrevocable Commitments within a Noncooperative Game

There are several ways of incorporating self-commitment moves into the extensive form of a game. For instance, we can define the payoffs in such a way that any violation of a commitment made by a player would carry heavy penalties, or we can add extra players to the game whose task is to punish violators. But the simplest method of doing it is this: At a suitable point of the game tree, we give the relevant player a choice between two moves, say, α and β, where α is interpreted as a commitment to do or not to do something at some later stage(s) of the game and β is interpreted as making no commitment. The commitment expressed by move α may be unconditional, or it may become operative only conditionally, subject to the occurrence of some future events. If the player chooses move β, then from that point the game will be governed by the remaining part of the original game tree, which we will call subtree T. But if he chooses move α, then from that point the game will be governed by a *modified* version of subtree T, to be called T'. T' will differ from T by having all branches *removed* that would correspond to moves violating the commitment that the player in question made when he chose move α (i.e., moves violating the commitment will simply not be available to this player).

It can of course happen that this removal of all commitment-violating moves will leave some of the players' information sets with one unique branch (one unique move), indicating that he no longer has a real choice at any of these information sets. Such information sets (and these unique branches) can always be omitted, since information sets permitting no real choice are irrelevant. This method can be easily generalized to cases where a player can choose not only between making and not making a specific commitment but rather among a number of alternative commitments.

For example, the extensive form of the game discussed in section 1.2 can be represented by the game tree in figure 1.2. The numbers 1 and 2 printed at the right of the two information-set symbols (the two ovals) indicate which player has a move at that particular information set.

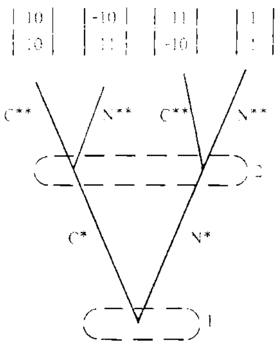

Figure 1.2

Now we can represent the players' ability to make an enforceable agreement about using their C-strategies as follows: At the beginning of the game, we give player 1 a choice between moves α^* and β^*, where α^* means "I commit myself to using strategy C^*, provided that player 2 will commit himself to using strategy C^{**}," while move β^* means "I make no commitment." In case player 1 has actually chosen move α^*, we now give player 2 a choice between moves α^{**} and β^{**}, where α^{**} means "Yes, I do commit myself to using strategy C^{**} as player 1 has suggested," and move β^{**} means "I make no commitment."

Now, we can distinguish three cases:

1. If player 1 chooses α^* while player 2 chooses α^{**}, then both players will be committed to using their C-strategies. Consequently the remaining part of the game will now be reduced to the subtree T_1, shown in figure 1.3. But, since each of the two information sets in T_1 has only one branch arising from it, we can omit both of these information sets as well as the two branches (C^* and C^{**}), which amounts to replacing the entire subtree T_1 by the payoff vector $\begin{vmatrix}10\\10\end{vmatrix}$ generated by it.

2. If player 1 chooses α^* while player 2 chooses β^{**}, then the two players will be under no commitment to restrict their freedom of action. Consequently the remaining part of the game will be governed by a subtree T_2 which is simply a copy of the original game tree.

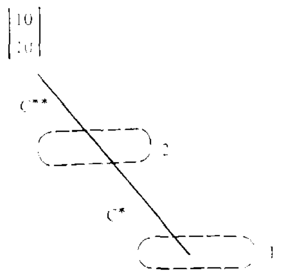

Figure 1.3

3. If player 1 chooses β^*, then once more the players will retain their freedom of action, and the remaining part of the game will be governed by a subtree T_3, which is again simply a copy of the original game tree.

Accordingly, the game tree of the enlarged game will be as shown in figure 1.4.

In the normal form of the enlarged game, we can characterize each player's strategies by three symbols. For example, the first symbol (α^* or β^* for player 1, and α^{**} or β^{**} for player 2) may be used to indicate the player's choice between commitment and no commitment, the second symbol (C^* or N^* for 1, and C^{**} or N^{**} for 2) may indicate the strategy that he would follow in subtree T_2, and the third symbol (C^* or N^*, or, alternatively, C^{**} or N^{**}) may indicate the strategy that he would follow in subtree T_3. Thus one possible strategy of player 1 would be $\alpha^*C^*N^*$. Obviously either player will have $2^3 = 8$ different pure strategies.

It is easy to verify that the enlarged game has only *one* perfect equilibrium point in pure strategies, $E_1 = (\alpha^*N^*N^*, \alpha^{**}N^{**}N^{**})$. In other words, if both players are able to commit themselves to C-strategies, it will be clearly in their interest to do so to obtain the payoffs (10, 10). At the same time the definition of E_1 contains two N^* and two N^{**} symbols. These indicate that each player would use his N-strategy if his opponent refused to commit himself to use his C-strategy. (This part of either player's strategy plan will of course not be implemented since the opponent will make the required commitment.)

Intuitively one can identify E_1 with the cooperative solution (C^*, C^{**}) of the original game. Thus we can say that by incorporating the commitment moves α^* and α^{**} (as well as no-commitment moves β^* and β^{**}) into the extensive form of the game, we have essentially turned the cooperative solution (C^*, C^{**}) into an equilibrium point—so as to make it an outcome achievable by rational players even if the game (or, rather, the enlarged version of the game) is played as a formally noncooperative game. Indeed, since E_1 is the *only* perfect equilibrium point of the enlarged game, we have turned E_1 into the only outcome consistent with rational behavior by both players. (For a more detailed analysis of the enlarged game, see section 1.14. As we will try to show in sections 1.9 and 1.10, only *perfect* equilibrium points are compatible with rational behavior by all of the players in a noncooperative game.)

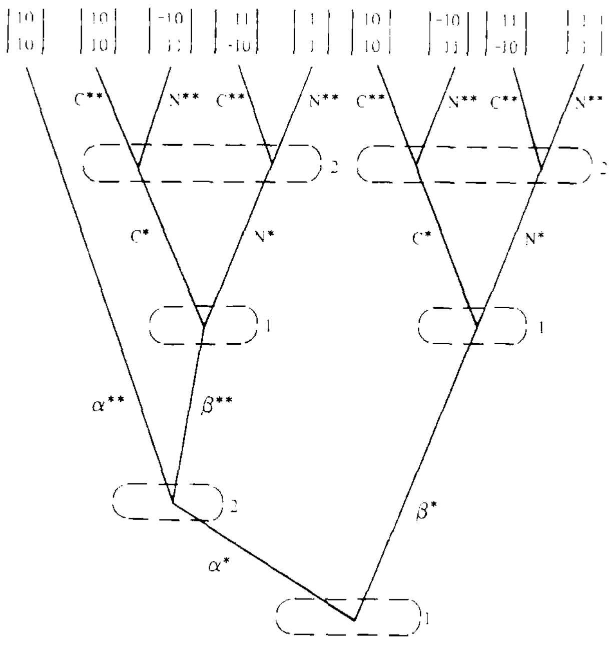

Figure 1.4

1.4 Limitations of the Classical Theory of Cooperative Games

The classical theory of *noncooperative* games is essentially a theory of one basic solution concept, that of equilibrium points. In contrast, the classical theory of *cooperative* games offers a rich variety of alternative solution concepts, namely the von Neumann-Morgenstern stable sets (1944), the Nash solution for two-person bargaining games (1950b, 1953), the Shapley value (1953), the core (Gillies 1959), the Aumann-Maschler bargaining sets (1964), among others.

Individually each of these solution concepts is of great theoretical interest. But as a group they fail to provide a clear, coherent theory of cooperative games. Indeed, most of the different solution concepts have very little logical connection and so cannot be interpreted as special cases of a general theory.

One may think that this fact is merely a conceptual limitation of classical game theory, which may be of some importance to the logician, methodologist, or philosopher but immaterial to the social scientist whose main interest lies in possible applications of game theory to economics, political science, and sociology. Yet, this conceptual limitation does in fact create major problems also in empirical applications.

First of all, although classical game theory offers a number of alternative solution concepts for cooperative games, it fails to provide a clear criterion as to *which* solution concept is to be employed in analyzing any real-life social situation. Nor does it give a clear answer to the obvious question of why so many different solution concepts are needed.

Many solution concepts generate some additional dimensions of indeterminacy. Even if the decision is made to analyze a given social situation in terms of some solution concept A, this will often fail to specify a well-defined outcome: it rather might tell us no more than that the outcome will be chosen from some (possibly very large) *set* S of "acceptable" outcomes: indeed, all it may tell us may be that the outcome will be a point lying in one of several alternative sets $S, S', S'', \ldots$, each equally consistent with the axioms of the chosen solution concept A.

An even more serious shortcoming of classical game theory is its failure to provide *any* usable solution concepts for some theoretically and empirically very important classes of cooperative (and of less than fully cooperative) games. These include:

1. Games *intermediate* between fully cooperative and fully noncooperative games. Examples are games where some types of agreements are enforceable while others are not; games where some groups of players are able to make enforceable agreements but others are not; and games where enforceable agreements can be concluded at some stages of the game but not at other stages.

2. Cooperative games with a *sequential* structure. (There is some overlap between cases 1 and 2.) These are games involving two or more successive stages and permitting agreements to be built up gradually in several consecutive steps. Unlike classical cooperative games, in which any agreement made is final, such sequential games might allow renegotiation and modification of earlier agreements at later stages of the game under specified conditions.

3. Cooperative games with *incomplete information.* (Since games with incomplete information, both cooperative and noncooperative, raise some special problems; we will discuss them at some length in section 1.5.)

All these difficulties are due to the fact that the classical theory of cooperative games systematically neglects any analysis of the *bargaining process* among the players, which is probably the most important activity in any cooperative game. This is done by describing this bargaining as "preplay negotiations" and by assuming that it takes part *before* the "game" is actually played, and that it is therefore not part of the "game" at all. This approach of course amounts to relinquishing any serious attempt to understand how the outcome of the game depends on the specifics of the bargaining process among players.

1.5 Games with Incomplete Information

One of the most serious deficiencies of classical game theory is its inability to deal with games involving incomplete information. We say that a game is one with *complete information* if all players know the nature of the game, in the sense of knowing the extensive form of the game (the game tree) or the normal form of the game (the payoff matrix).

A game with complete information can be a game with either *perfect* or *imperfect information.* In a game with perfect information the players know both the nature of the game and all previous moves (made by other players or by chance) at every stage of the game; in a game with imperfect information the players know the nature of the game but have less than full information about the earlier moves during the game.

In contrast, in a game with *incomplete* information, the players have less than full information about (1) the strategy possibilities and or (2) the payoff functions of the other players. Problem 2 may arise because the players may have limited information about:

1. the *physical consequences* to be produced by alternative strategy combinations,
2. the other players' *preference rankings* over these physical outcomes,
3. the other players' attitudes toward *risk taking*, or
4. some combination of these factors.

In addition the players may be ignorant about the amount of *information* that the other players have about any player's strategy possibilities and his payoff function.

Classical game theory cannot handle games with incomplete information at all (but does cover both games with *perfect* and with *imperfect* information as long as these have the nature of games with *complete* information). This obviously poses a very serious limitation since virtually all real-life game situations involve incomplete information. In particular, it very rarely happens that the participants of any real-life social situation have full information about each other's payoff functions. Uncertainty about the strategies available to the other players is also quite common.

We can, however, bring a game with incomplete information within the scope of game-theoretical analysis by using a probabilistic model to represent the incomplete information that the players have about various parameters of the game (Harsanyi 1967, 1968a, 1968b). In particular, the analysis of a game with incomplete information, *G*, can be reduced to analysis of a new game, *G**, involving suitably chosen random moves. We call *G** a *probabilistic model* for *G*. In this new game *G** the fact that (some or all of) the players have limited information about certain basic parameters of the game is mathematically represented by the assumption that these players have limited information about the outcomes of these *random moves*.

Formally, this probabilistic model game *G** will be a game with *complete* information. But it will be a game with *imperfect* information because of the players' having less than full information about the outcomes of the random moves occurring in the game. Thus our approach essentially amounts to reducing the analysis of a game with *incomplete* information, *G*, to the analysis of a game with *complete* (yet imperfect) information, *G**, which is fully accessible to the usual analytical tools of game theory.

By constructing suitable probabilistic models, we can produce games with any desired distribution of knowledge and ignorance among the players and can study how alternative informational assumptions will change the nature of the game. We can learn how a player can infer some pieces of information originally denied to him, by observing the moves of players who already possess this information, and also how a player can optimally convey information to other players or optimally withhold information in accordance with his own strategic interests. (We discuss the problem of optimally conveying information in chapter 9 where we analyze a two-person game with incomplete information on both sides. For the problem of optimally withholding information, see Aumann and Maschler's discussion, 1966, 1967, 1968, of infinitely repeated two-person zero-sum games under incomplete information, and also Stearns 1967; cf. Harsanyi 1977a.)

To be sure, the use of such probabilistic models provides only a partial solution for the problem of how to analyze games with incomplete information. For when a probabilistic-model game *G** is constructed for a game with incomplete information, *G*, there immediately arises the problem of what *solution concept* to use for this newly constructed game *G**.

If, in fact, the game *G* we start with is a *noncooperative* game with incomplete information, then this question has an easy answer. The probabilistic-model game *G** derived from *G* will also be a *noncooperative* game (though one with complete information), and *G** can be analyzed in terms of its equilibrium points; the concept of equilibrium points can be extended to games with incomplete information without difficulty (Harsanyi 1968a, pp. 320–329).

The situation is very different if the game *G* is a *cooperative* game with incomplete information. In this case the probabilistic-model game *G** derived from *G* will not admit of analysis in terms of *any* cooperative solution concept of conventional game theory. For example, the Nash solution for two-person bargaining games, which is an attractive solution concept for games with *complete* information, cannot be used for two-person bargaining games with *incomplete* information or for the probabilistic-model games derived from them. If we try to use the Nash solution for this purpose, we obtain completely nonsensical results (Harsanyi 1968a, pp. 329–334). Other classical cooperative solution concepts give equally unsatisfactory results when applied to incomplete information games. This lack of solution concepts applicable to games with incomplete information is another serious weakness of the classical theory of cooperative games.

1.6 Difficulties with the Concept of Equilibrium Points

Compared with the classical theory of cooperative games, the classical theory of noncooperative games presents a more satisfactory picture. It has more theoretical *unity* because it is based on one basic solution concept, that of equilibrium points. It is also a more complete theory because it tries to cover all aspects of a game and does not automatically exclude the players' bargaining moves from its analysis in the way the theory of cooperative games does. Furthermore the concept of equilibrium points—and therefore the classical theory of noncooperative games—can be easily extended to games with incomplete information.

Finally, the concept of equilibrium points is one of the very few game-theoretical solution concepts that has direct application to games, in *both* extensive and normal form. (This has many desirable consequences. One is that the classical theory of noncooperative games, unlike that of cooperative games, can easily handle games with sequential structure.)

Although the concept of equilibrium points has many strong points, it also has weaknesses, three of which are important to our discussion:

1. Almost every nontrivial game has many (sometimes infinitely many) different equilibrium points. Hence a theory that can only predict that the outcome of a noncooperative game is an equilibrium point—without specifying which equilibrium point it is—is an extremely weak and uninformative theory. This difficulty we call the *equilibrium selection problem.*

2. Any mixed-strategy equilibrium point is, or may appear to be, fundamentally unstable (see section 1.8), and therefore not a suitable solution of a game. This gives rise to what we call the *instability problem*: how are we to define a solution for a noncooperative game that has only mixed-strategy equilibrium points?

3. The third difficulty was pointed out by Reinhard Selten (1965, 1975): many equilibrium points require some or all of the players to use highly irrational strategies (see sections 1.9 and 1.10). He proposed to call such equilibrium points *imperfect* equilibrium points, to distinguish them from *perfect* equilibrium points, which involve no irrational strategies. The problem posed by games that contain imperfect equilibrium points we call the *imperfectness problem.*

1.7 The Equilibrium Selection Problem

Among the three problems posed by the concept of equilibrium points, the equilibrium selection problem is of particular importance. To illustrate the nature of this problem, we consider a very simple two-person bargaining game, where two players have to agree on how to divide $100; the money is lost to them if they cannot agree. (We will assume that both players have linear utility functions for money.) This game can be represented by the following bargaining model: Each player has to name a real number, representing his payoff demand. The numbers named by players 1 and 2 will be called x_1 and x_2, respectively. If $x_1 + x_2 \leq 100$ (if the two players' payoff demands are mutually compatible), then both will obtain their payoff demands, with $u_1 = x_1$ and $u_2 = x_2$. In contrast, if $x_1 + x_2 > 100$ (if their payoff demands are incompatible), they will receive zero payoffs $u_1 = u_2 = 0$ (as this will be taken to mean that they could not reach an agreement).

If the players are free to divide the $100 in all mathematically possible ways, this game will have infinitely many equilibrium points in pure strategies because all possible pairs (x_1, x_2) satisfying $x_1 + x_2 = 100$, where $x_1 \geq 0$ and $x_2 \geq 0$, will be equilibrium points. But even if we restrict the players to payoff demands representing integer numbers of dollars, the game will still have 101 equilibrium points, from (0, 100), (1, 99), . . . , to (100, 0). Clearly a theory telling us no more than that the outcome can be any one of these equilibrium points will not give us much useful information. We need a theory selecting one equilibrium point as the solution of the game. The purpose of our new solution concept is to provide a mathematical criterion that always selects one equilibrium point as the solution. In other words, with our one-point solution we attempt to overcome the equilibrium selection problem. (But, as we will try to show, our theory also overcomes the two other problems posed by the concept of equilibrium points—the instability problem and the imperfectness problem.)

1.8 The Instability Problem: A New Justification for Use of Mixed-Strategy Equilibrium Points

To illustrate the instability problem posed by games having only mixed-strategy equilibria, consider the game in figure 1.5. The only equilibrium in this game is in mixed strategies and has the form $E = (M, N)$, where $M = (\frac{1}{3}, \frac{2}{3})$ and $N = (\frac{4}{5}, \frac{1}{5})$ (i.e., player 1's equilibrium strategy M assigns the probabilities $\frac{1}{3}$ and $\frac{2}{3}$ to his two pure strategies A and B, respectively, while player 2's equilibrium strategy N assigns the probabilities $\frac{4}{5}$ and $\frac{1}{5}$ to his two pure strategies X and Y) To facilitate analysis of this game, we will add a new row, corresponding to M, and a new column, corresponding to N, to the payoff matrix (figure 1.6). As can be seen from this enlarged payoff matrix, if player 1 expects player 2 to use his equilibrium strategy N, then player 1 will have no real incentive to use his equilibrium strategy M. This is so because he will obtain the same payoff $u_1 = 36$, regardless of whether he uses his mixed equilibrium strategy M, either of his two pure strategies A and B, or any mixed strategy other than M. Likewise player 2 will have no real incentive to use his equilibrium strategy N, even if he expects player 1 to use the equilibrium strategy M. The reason is player 2 will obtain the same payoff $u_2 = 60$ regardless of whether he uses his equilibrium strategy N, either of his two pure strategies X and Y, or any mixed strategy other than N.

This is what we mean by saying that the equilibrium point $E = (M, N)$ is (seemingly) unstable: even if this does not provide an incentive for either player not to use his equilibrium strategy, it does not provide an incentive that would make it positively attractive for him to use his equilibrium strategy.

We now argue that the instability of such mixed-strategy equilibrium points is only apparent. Even if the players have as complete information about the payoff matrix of the game as they can possibly have, each player will always have some irreducible minimum of *uncertainty* about the other player's actual payoffs. For example, even though the payoff matrix shows player 2's payoff associated with the strategy pair (A, X) to be $H_2(A, X) = 30$, player 1 will never be able to exclude the possibility that at this very moment this payoff may be in fact $30 - \varepsilon$ or $30 + \varepsilon$, where ε is a small positive number. This is so because every person's utility function is subject to at least some, very small, unpredictable random fluctuations because of changes in his mood or perhaps a sudden urge to use one of his pure strategies in preference to his other pure strategy.

This means that a realistic model of any given game will not have fixed payoffs but rather *randomly fluctuating* payoffs, even though these fluctuations may be very small. Mathematical analysis shows that such a game will have no mixed-strategy equilibrium points.[4] Rather, all its equilibrium points will be in pure strategies, in the sense that neither player will ever intentionally randomize between his two pure strategies. Instead, he will always find that one of his two pure strategies will yield him a higher expected payoff, and this is the pure strategy that he will actually use.

At the same time it can be shown that the random fluctuations in the two players' payoffs will interact in such a way that player 1 will find strategy A to be more profitable than strategy B almost exactly one-third of the time, and B more profitable than A almost exactly two-thirds of the time. As a result, though he may make no attempt to randomize, he will use his two pure strategies almost exactly with the probabilities prescribed by his equilibrium strategy $M = (\frac{1}{3}, \frac{2}{3})$. By the same token, though player 2 may make no attempt to randomize, he will use his two pure strategies X and Y almost exactly with the probabilities prescribed by his equilibrium strategy $N = (\frac{4}{5}, \frac{1}{5})$. (For detailed discussion and for mathematical proofs, see Harsanyi 1973a.)

	X	Y
A	45 30	0 90
B	30 75	60 45

Figure 1.5

	X	Y	N
A	45 30	0 90	36 42
B	30 75	60 45	36 69
M	35 60	40 60	36 60

Figure 1.6