Preface

The theory presented in this book is for the most part quite new. The most primitive ideas of mathematical game theory—the extensive and normal forms, and the equilibrium point—are developed in a new way so as to build a positive theory applicable to the "real world."

The history of these developments is as follows. In 1965, Russell L. Ackoff invited the author to join the Management Science Center, University of Pennsylvania, which at that time had an unclassified contract with the United States Arms Control and Disarmament Agency to develop methods of analyzing the escalation and de-escalation of conflict. The principal investigator was Ackoff himself. The sponsor at A.C.D.A. was Thomas L. Saaty, whose emphasis on the need to develop analytic methods for the understanding of real-world conflict problems had inspired the agency to support research in this area (see Saaty 1964).

The author was asked to work on the A.C.D.A. contract, and it was in this encouraging milieu that metagame theory was born. One can safely say that had it not been for the efforts of Saaty and Ackoff the work presented here would not now exist. Saaty in particular proved an ideal sponsor.

Others have contributed this and other kinds of help. Anatol Rapoport became keenly interested in the ideas at an early stage, arranging for their early publication in the *General Systems Yearbook* (Howard 1966, 1970). Walter Isard arranged for the publication of this book and in other ways disinterestedly "pushed" the approach. Meanwhile, the interest shown by Herbert Scoville, Jr., then at the U.S. Arms Control Agency, encouraged experimental applications to realistic problems as in the case of the Vietnam analysis described in Section 5.3. P. J. Long was among other members of the Agency who (actively persuaded by Dr. Saaty) cooperated in these applications. On a technical level, André Ducamp, Jeffrey Smith, and John Hall worked with me at the Management Science Center for various periods during the growth of these ideas and contributed much during many hours of ardent discussion. Finally, in 1969, the National Research Council of Canada contributed research funds

and H. K. Kesavan found a place in the Department of Systems Design at the University of Waterloo for the continuing development of metagame theory and the writing of the final version of this book. An earlier version was accepted as an external Ph.D. thesis by the University of London (1968), and I am grateful to Drs. George Morton and Ailsa Land of the London School of Economics, who supervised my Ph.D. work, for their interest, inevitably "long-distance" though it was.

I shall now attempt to elucidate the connections between this work and the work of other game theorists. What follows will therefore assume a knowledge both of their work and the contents of this book; but those who know only the work of other game theorists may nevertheless profit by skimming through this first.

EXTENSIVE AND NORMAL FORMS. We build the extensive form in much the same way as other game theorists (which is essentially the way laid down by von Neumann) except for one difference. Preferences, or utilities, are not introduced at this stage.

The game tree is merely a structure in which particular strategies chosen by the players and by "chance" lead to particular *time-paths of the system*, or "scenarios." Hence when the normal form is constructed, a particular strategy n-tuple—choice of one strategy by each player—leads to a probability distribution over various possible time-paths, or a probabilistic spectrum of "scenarios." Preferences of a *nonquantitative* kind are now introduced; that is, for each player a preference *relation* of a general kind is assumed over the various strategy n-tuples (now called "outcomes"). The reason for this different treatment is to carry through a nonquantitative approach to preferences. We end up with a normal form differing from the usual one only in that it is more general: general preference relations take the place of numerical utility functions.

EQUILIBRIUM POINTS AND RATIONAL OUTCOMES. Within this framework, equilibrium points are defined as usual. Much use is made, however, of the concept of a "rational outcome for player i" (a point at which player i is optimizing his preferences). This is done to bring out a point that is often overlooked—that the arguments for "stability" of an equilibrium point actually do not depend on players knowing

each other's preferences. This is mathematically trivial but carries with it a considerable reinterpretation of applied game theory.

MIXED STRATEGIES. Most game theorists, after some examination of the general case, proceed to further consideration of only a certain type of normal-form game—that which is a *mixed extension* of (another) normal-form game. We do not specialize in this way. Of course, we consider mixed extensions, these being seen as a particular kind of normal-form game in which each player has an infinite set of strategies (his so-called mixed strategies) and a numerical utility function. But we do not consider only these—we consider all normal forms, even ones in which players have a finite number of strategies and their preferences need not even be ordinal. We do not take advantage of the special structure possessed by mixed extensions.

CORRELATED MIXED STRATEGIES. Since we consider all normal forms, we consider also games in which players may *correlate* their mixed strategies (Luce and Raiffa 1957, p. 116) provided that this situation is first represented in normal form—a prior step omitted by most authors. For instance, a normal-form representation might allow each player i to choose a strategy consisting of a 3-tuple

$$(C, \sigma_C, \sigma_i),$$

in which C is his choice of a *coalition*, σ_C his choice of a *coalition correlated mixed strategy*, and σ_i his choice of an *individual mixed strategy*. The choice σ_C is carried out if and only if each player in C chooses C and also σ_C; otherwise, the choice σ_i is carried out. Numerical payoff functions are defined in the obvious way.

In order to refer to this normal form later on, we may call it a *correlated mixed extension*. We remark that unless some normal-form representation of a game is given, the game cannot be run experimentally. The universality of the normal form (or of the extensive form from which a normal form can be derived) consists in the fact that giving any well-defined set of rules for actually playing a game amounts to specifying the game in this form. Of course, in experiments the players' final choices may or may not be preceded by a period of communication—the normal form says nothing about this.

Unfortunately the normal form suggested here—the correlated mixed extension—does more than allow correlated mixed strategies. It also allows the individual players to choose "threat" strategies similar to (though not the same as) the threat strategies defined by Nash (1953) in his "bargaining solution" to a game. These are the individual strategies σ_i that are implemented in case the coalition fails to agree. They can be used as "threats" against the other members of the coalition. It does not seem possible to avoid adding this feature if we wish to construct a normal form allowing correlated mixed strategies. This is unfortunate because, if we consider a game without correlated mixed strategies, the metagame approach makes "threat" strategies in the normal form superfluous; whatever can be achieved by them can be achieved more naturally by using metagame "policies"; and if despite this we allow "threats" to appear in the normal form, we obliterate a very meaningful distinction between "basic" equilibria (equilibria in the original game) and "nonbasic" ones (which are equilibria only by a process of derivation from some metagame).

TRANSFERABLE UTILITIES. Our approach also covers the case of numerical utilities freely transferable between players, provided again that this is represented in normal form. This can be done by allowing each player, in addition to making his "ordinary" strategy choice (pure or mixed or correlated as above) to choose the proportions in which he will distribute his utility payoff among all the players.

α- AND β-EFFECTIVENESS. Generally speaking, the notion of *α-effectiveness* (Aumann 1961) corresponds to the metagame notion of *general metarationality*, while *β-effectiveness* corresponds to *symmetric metarationality*. To make this precise, however, we must first note that α- and β-effectiveness are defined (Aumann 1961) only for a limited class of games—those which are correlated mixed extensions.

But if in a correlated mixed extension with numerical payoff functions M_i we let "s_K" stand for a joint strategy of the coalition K and "N" for the set of all players, the coalition C is said to be α-effective for the payoff vector x if

$$\exists s_C \, \forall s_{N-C} : \forall (i \in C) : M_i(s_C, s_{N-C}) \geqslant x_i,$$

and it is β-effective for x if

$$\forall s_{N-C}\, \exists s_C : \forall (i \in C) : M_i(s_C, s_{N-C}) \geqslant x_i.$$

On the other hand, we find that an outcome $\bar{s}$ (a strategy n-tuple) fails to be general metarational for C if

$$\exists s_C\, \forall s_{N-C} : \forall (i \in C) : M_i(s_C, s_{N-C}) > M_i\bar{s},$$

and it fails to be symmetric metarational for C if

$$\forall s_{N-C}\, \exists s_C : \forall (i \in C) : M_i(s_C, s_{N-C}) > M_i\bar{s}.$$

The similarity can be seen. The real difference is that between strict and nonstrict inequalities. A coalition C is α-effective for x if C can guarantee each of its members i at least x_i; an outcome yielding the payoff vector x fails to be metarational for C if C can guarantee each of its members i more than x_i. A coalition C fails to be β-effective for x if $N-C$ can guarantee that some member i of C will receive less than x_i; an outcome yielding x is symmetric metarational for C if $N-C$ can guarantee that some member i of C will receive no more than x_i.

α- AND β-DOMINATION. Based on the above two concepts of effectiveness, Aumann (1961) forms two concepts of *domination*, saying that a payoff vector x *α-dominates* (respectively *β-dominates*) y *through* the coalition C if $x_i > y_i$ for all $i \in C$, and C is α-(respectively β-)effective for x.

The result, at least in "well-behaved" games, is that the set of general (respectively symmetric) metarational outcomes for C is precisely the set of outcomes yielding payoff vectors α-(respectively β-) *undominated* through C.

Finally, the "α-core" and "β-core" being defined as the sets of payoff vectors not α-(respectively not β-)dominated through any coalition C, we find that the sets of outcomes that are general (respectively symmetric) metarational for all coalitions are just those that yield payoffs in the α- and β-core. Hence in this book we have called these sets (though they are sets of strategy n-tuples, not payoff vectors) the α-core and β-core, respectively.

We remark again that general and symmetric metarationality are defined for far wider classes of games than correlated mixed exten-

sions, being defined for all normal-form games, even ones with "general" preferences that are not even ordinal. But corresponding generalizations of the notions of effectiveness and domination present no difficulty. With general preferences, one can no longer speak of numerical payoff vectors. The simplest way around this seems to be to define outcomes as *strategy n-tuples*, as we do in this book, rather than as *payoff n-tuples*. We may then form general preference relations over the set of outcomes (strategy n-tuples). The alternative chosen by Peleg (1966), which is to define a function leading from strategy n-tuples to a set of nonnumerical "payoffs," over which, finally, preference relations are formed, seems clumsier, introducing as it does an unnecessary concept in between strategy n-tuples and preferences.

GAMES IN CHARACTERISTIC FUNCTION FORM. If a game is given in characteristic function form, we must put it in normal form before we can apply metagame theory. That this can be done is well known. It is also well known that if a game is transformed from normal form to characteristic function form and then back again, much detail is lost. Accordingly, many distinctions made in metagame theory are obliterated. The sets of general and symmetric metarational outcomes remain the same, but there is no longer any distinction between different types of metarationality within these broad categories.

These remarks apply not only to the von Neumann–Morgenstern characteristic function form (derived from a numerical correlated mixed extension with transferable utilities), in which symmetric and general metarationality coincide, but also to the corresponding form for games without side payments in Stearns (1964) and Aumann and Peleg (1960) and to the game in partition function form (Thrall and Lucas 1963).

BARGAINING SOLUTIONS. Following Nash (1950, 1953), Harsanyi (1959, 1963, 1966) has developed a "bargaining solution." Aumann and Maschler (1964) have investigated the "bargaining set." Maschler and Peleg (1966) have developed the idea of the "kernel"; and we know of the "Shapley value" (Shapley 1953). These concepts, unlike the ones discussed so far, presuppose players who know each other's

preferences. But a metagame approach to such players is at present very underdeveloped; the last chapter of this book merely sets out some simple notions on the subject. Hence no real connection yet exists between metagame theory and these approaches.

GENERAL DISCUSSION. There are three main points to be made about the approach taken in this book.

a. The approach is *positive*. It is neither purely formal, nor is it normative. This means that assertions about the behavior of "players" in a "game" are to be interpreted as empirical statements and their consequences tested under controlled conditions if possible. The assertions are to be rejected if their consequences fail to pass such tests. It also means that to us a "game" is not only a mathematical object but also and simultaneously an *experimental* object. Hence our insistence on the normal-form (or extensive-form) representation, without which it is not clear under what experimental conditions any assertions are to be tested.

b. The approach is *nonquantitative*. Numbers are not used. Our motive is not only mathematical generality but also real-world applicability. Numerical utilities, even if they are well founded theoretically, cannot be reliably estimated in the real world.

The result, mathematically, is an extremely general approach—so much so that mathematically this is really a book about abstract set theory à la Cantor. Fraenkel (1953) is a good book to read to see how the subject appears from this viewpoint. But for this generality we pay a price. Cold winds blow through unstructured sets! Existence theorems in particular are hard to find. Thus the papers (e.g., Aumann 1961) to which we have referred earlier, and concerning which we have had to remark that the corresponding metagame concepts are more general, usually contain delicate and interesting existence theorems that we lose entirely.

c. The approach is *based on the metagame tree*. This is a separate point from the two preceding ones. We could have constructed a *formal* (nonempirical) or a *normative* theory based on the metagame tree; moreover, this theory could have been *quantitative*. Instead, we have a positive, nonquantitative metagame approach.

The metagame idea is that to analyze a game we should analyze

the n metagames based on it. This was first recommended by von Neumann and Morgenstern (1953, section 14.2) in the context of the two-person zero-sum game. However, they failed to follow it through. They did not see that the recommendation is recursive, and hence they did not analyze the metagames based on the metagames (the "minorant" and "majorant" games) that they did analyze. Clearly, to follow the idea through we must analyze the whole infinite tree of metagames. Also, they did not extend the idea beyond two-person zero-sum games.

Thus von Neumann and Morgenstern may be said to have originated the metagame approach. In other respects also our approach is based on theirs. Thus, the von Neumann–Morgenstern approach is, in contrast to later approaches, thoroughly positive. True, they did not experiment, but their approach was neither normative nor purely formal. In drawing an analogy with the development of physical theory, they compare themselves (1953, section 1) to theoretical, in contrast to experimental, physicists. Second, von Neumann and Morgenstern looked forward (1953, section 66) to a nonquantitative generalization of game theory. This suggestion has not of course been so thoroughly neglected as the metagame suggestion; our work is merely the least quantitative of all that has been done so far (see, for example, Peleg 1966).

Finally, von Neumann and Morgenstern (1953, section 66.4) look forward to a unification of their rather separate theories of the two-person zero-sum game and the n-person variable-sum game. Such a unification has probably been delayed by the distinction introduced later between "noncooperative" and "cooperative" theories of games —a distinction that clearly has its genesis in von Neumann and Morgenstern's two approaches. Their plea for unity is, however, answered by metagame theory. There is no need in metagame theory for different approaches or different "solution concepts." We have a unified treatment applied to all games and based on

1. The normal form.
2. The concept of the equilibrium point as an intersection of rational outcomes.
3. The metagame tree.

Within this framework we find that the outcomes *undominated through C* (which in the von Neumann–Morgenstern treatment are the outcomes at which C receives at least the characteristic function value $v(C)$) are *derived.* They are just the outcomes metarational for C from *some* metagame in the infinite tree. Hence this set of outcomes —and similarly the sets of α- and β-undominated outcomes—is inevitably singled out by our approach.

We start, in other words, by looking for rational outcomes: an approach that is definitely "two-person zero-sum" and "non-cooperative." We apply this to the metagame tree. And we obtain the "many-person variable sum" and "cooperative" concept of the set of outcomes undominated through C. Meanwhile, what has happened to the noncooperative solution? Has it disappeared? Not at all. The outcomes rational for C (corresponding to the noncooperative solution) are those which are metarational for C from *every* metagame in the infinite tree. The distinction previously embodied in separate theories is now simply that between different classes of metarational outcomes in the metagame tree.

Traditionally, game theorists have used an arbitrary and ad hoc procedure that consists in first proposing a "solution concept" and then investigating its properties. Because this in effect means that the disunity of the field is continually increased by the laying down of new sets of basic definitions, it has been abandoned. We lay down no new basic definitions except the definition of a metagame. With this one exception, our definitions are not proposed as basic concepts but as tools with which to explore the structure of the metagame tree, a distinction that, though it may be somewhat cloudy, embodies an essential difference. New tools with which to explore a given structure may create unity; the continual creation of new structures to be explored has the opposite effect.

THE EXISTENTIALIST AXIOM, THE FREE WILL ARGUMENT, AND THE AXIOM OF CHOICE. As we have said, the one new basic definition introduced is that of a metagame; and we have therefore gone to some trouble to interpret and justify this concept. In so doing we have used two arguments that may interest philosophers and one that may interest students of the foundations of mathematics. In Section 3.3

we argue that if a person comes to "know" a theory about his behavior, he is no longer bound by it but becomes free to disobey it. This is the "existentialist axiom." In the same section is the "free will argument," which points out that it is harder to believe that one's free choice will be as predicted by another than to believe that one can predict his (the other's) free choice; yet such a bias is illogical. Finally, in Section 4.3 we propose an interpretation of the axiom of choice based on the consideration of an imaginary experiment (which could actually be performed) in which to *reject* the axiom of choice is to assert that a certain player cannot choose certain strategies.

THE BREAKDOWN OF RATIONALITY. We not only discover facts about the metagame tree, we try to interpret these facts. And in so doing we accept the discipline of experiment: if it predicts wrongly, it is wrong. This enables us to proceed without appealing, as other game theorists have done, to arguments based on elaborate concepts of "rationality." Instead we find a use for the simplest and most straightforward definition of "rational behavior" (namely, *optimizing behavior*) as a building block in our theory. We say that *rational* behavior consists in choosing the alternative one prefers.

Adherence to this simple definition leads us, however, to point out that people are not rational. First, sometimes two people cannot both be rational (our first breakdown). Second, sometimes both are better off if they are both irrational (our second breakdown). These facts are well known to game theorists—who, however, have generally preferred to change the definition of rationality, often making it abstruse and hard to accept, rather than admit that the concept has "broken down."

Our third breakdown, however, appears not to have been noticed before. It is described in Section 6.4, where a theorem is proved (Theorem 9) to the effect that to be rational in two-person games is usually to be a sucker. It is suggested that this is the reason why, even when they are rational, people such as political leaders, businessmen, and those involved in the battle of the sexes seldom talk as if they are.

Why has Theorem 9 been overlooked? It is a very simple theorem.

The reason may lie in the attitude adopted by most game theorists toward so-called sure-thing strategies—this being that a player lucky enough to have such a strategy should "obviously" pursue it. Theorem 9 flatly contradicts this, showing that sure-thing strategies are, in many realistic situations involving two players, incredibly silly.

THE ANALYSIS OF OPTIONS. This book is primarily about theory. But Section 5.3 sketches a method, called the "analysis of options," whereby our theory can be applied usefully to real-life political conflicts. The method, now under intensive further development at the University of Waterloo, is described more completely in Management Science Center (1969b). It takes full advantage of the non-quantitative approach to game theory.

As development has increased the scope and power of this method, it has been applied with increasing success to larger and more complex models. The first applications, starting in 1967 on a tentative experimental basis, were carried out in Washington with the cooperation of A.C.D.A. personnel and concerned such problems as the strategic arms race (the A.B.M. problem, etc.) and nuclear proliferation. In 1968 the Vietnam conflict was studied. As we have said, an excerpt from this analysis is given in Section 5.3. Later, T. L. Saaty and the author, at the Urban Institute in Washington, D.C., helped to analyze two urban conflict problems: one, a problem involving urban transportation systems, was analyzed in cooperation with Henry Bain; the other, a historical analysis of the New York school strike problem, was conducted with the help of Betsy Levin. But the most complex and thoroughgoing analysis so far (also the most sensitive) was an analysis of many aspects of the Arab–Israeli conflict, conducted in the spring of 1970 by T. L. Saaty and the author with a group of interested individuals in Beirut, Lebanon, under the auspices of the Royal Jordanian Institute. In all this, the extent of the author's indebtedness to Thomas L. Saaty cannot be overemphasized. Indeed, without his efforts no applications of the theory would have been made.

I believe that further development of this technique will significantly improve the methods presently used in international politics. This

could be important for the future of mankind. For this reason, we should indicate the limitations of the game-theoretic approach.

There are two main causes of all the evil that exists in the world. One is that humans are very wicked; the other is that they are very stupid. Technocrats tend to underestimate the first factor and revolutionaries the second. But both are important.

Now there is no reason to hope that applied game theory will diminish human wickedness. It will not affect our preferences for killing, persecuting, and displacing one another. These can be affected only by changes in political consciousness and by the creation of *moral* theories, which to that extent are *not* scientific—though they may be based on scientific findings. To the extent that a theory is scientific, it is value-free and has no tendency to make people morally better or worse than they were before.

Nor will applied game theory diminish most kinds of human stupidity. Only certain kinds will be affected. Our approach is to take as given whatever misconceptions and delusions (or possibly sound information) decision-makers may have about the preferences and alternative courses of action open to the participants in a game situation. Having accepted these as premises, we can correct the stupid and illogical assertions (or possibly agree with the sound assessments) that the same decision-makers have derived from these premises. In other words, all we can say is "*If* your view of the world is correct, still this does not follow" (or possibly "then you are quite right for these reasons").

Even so, the area in which we *can* bring about improvements is quite significant. What we might call "game-theoretic" stupidity is both extremely pervasive and usually damaging to the interests of the "players." This, in any case, is the area which we have to explore.

THE NONMATHEMATICAL READER. This book is supposed to be mathematically self-contained, the required background being given in Appendix A. No mathematical knowledge is assumed beyond this, except in a few examples that may be skipped if necessary. If, however, I am lucky enough to attract a nonmathematical reader who really wishes to understand, he will have to work extremely hard. My advice to him would be to imagine continually that he is not

learning from the book but is *teaching* from it, and to construct (a) examples to illustrate the material being taught and (b) exercises such as he would have to give to a class of students to test their understanding. Such examples and exercises will be far more valuable than any that I could give.