mayer PPS 55

From
Choice and Consequence,
Thomas C. Schelling, Harvard
University Press, 1984

[redacted]

10 What Is Game Theory?

YOU ARE on the station platform, ready to board the train, and meet an old friend who has reserved a seat in a different car from yours. You agree to meet in the diner. After you board the train a steward comes through making reservations, and you discover that there is a first-class diner and a second-class buffet car. You'd somewhat rather eat in first class, you suspect that your friend would prefer the buffet car, but mainly you want to make a reservation that coincides with his. Do you elect the diner or the buffet car?

Again you are on the platform and meet a friend you are trying to avoid; he is going to coax you onto some committee. Your reservations are in different cars, but he suggests meeting in the diner. When the steward comes through you discover to your relief that there are two diners, first-class and buffet, and if you choose correctly you may "innocently" miss your friend. You have to be careful; he can guess that you'll evade him if you can. Normally you'd dine first class and he knows it. For which car do you make your lunch reservation?

Once again, you are on the train without a reserved seat. You find a seat but a few passengers are left standing. When the steward announces lunch, the standing passengers watch eagerly to see who will vacate a seat in favor of lunch. If you go to the diner you will have no claim to your seat when you return. If you do not vacate your seat you cannot eat; if you do not eat nobody gets

PPS0839

10 What Is Game Theory?

YOU ARE on the station platform, ready to board the train, and meet an old friend who has reserved a seat in a different car from yours. You agree to meet in the diner. After you board the train a steward comes through making reservations, and you discover that there is a first-class diner and a second-class buffet car. You'd somewhat rather eat in first class, you suspect that your friend would prefer the buffet car, but mainly you want to make a reservation that coincides with his. Do you elect the diner or the buffet car?

Again you are on the platform and meet a friend you are trying to avoid; he is going to coax you onto some committee. Your reservations are in different cars, but he suggests meeting in the diner. When the steward comes through you discover to your relief that there are two diners, first-class and buffet, and if you choose correctly you may "innocently" miss your friend. You have to be careful; he can guess that you'll evade him if you can. Normally you'd dine first class and he knows it. For which car do you make your lunch reservation?

Once again, you are on the train without a reserved seat. You find a seat but a few passengers are left standing. When the steward announces lunch, the standing passengers watch eagerly to see who will vacate a seat in favor of lunch. If you go to the diner you will have no claim to your seat when you return. If you do not vacate your seat you cannot eat; if you do not eat nobody gets

Used by Permission:Thomas Schelling, (1967). "What is Game Theory?", Contemporary Political Analysis, 212-238. Free Press

your seat, not even for the time you would like to be in the diner. What arrangement can you work out?

Finally, you are in your air-conditioned parlor car when the steward gives you your ballot. Passengers are asked whether they wish smoking to be permitted in these cars. You suspect it will be a close vote. A second item on the ballot asks whether, if smoking is permitted in response to the passengers' wishes, it should be confined to cigarettes. You'd love a cigar, which is all you ever smoke, would evidently answer no to the second question, but suspect that all nonsmokers and some cigarette smokers would vote to exclude cigars. How do you vote? Can you make a deal with someone?

As you leave your parlor car at the station, the steward stands expecting his tip. Fifty cents would be a reasonable tip, but the steward disposes of enough favors to make it worth some small expense to be among his favorites. You suspect that some of the other regular commuters try to tip a little above average. You'd like to tip a little above average. How much do you give the steward?

Interdependent Decisions

Game Theory is the formal study of rational decision in situations like these. Two or more individuals have choices to make, preferences regarding the outcomes, and some knowledge of the choices available to each other and of each other's preferences. The outcome depends on the choices that both of them make, or all of them if there are more than two. There is no independently "best" choice that one can make; it depends on what the others do.

For some problems, like choosing the route that minimizes distance from home to office, you can reach a solution without solving anybody else's problem at the same time. To drive through an intersection, though, you want to know what the other driver is going to do—to stop, slow down, speed up, or just keep going—and you know that a main element in his decision is what he thinks you are going to do. Any "solution" of a problem like this is necessarily a solution for *both* participants. Each must try to see the problem from the other's point of view, but when he does he sees himself trying to reach a decision.

What game theory did was to identify this class of situations as

one of practical importance and intellectual challenge, and to propose that any satisfactory solution for rational participants ought to be a solution for them jointly. Each must base a decision on his expectations. Unless we are willing to suppose one or more among them merely to expect wrong—and then we have to decide *ad hominem* who is going to be wrong—there must be some consistency, not only of their choices with their expectations but among their expectations of each other. Game theory is the formal study of the rational, consistent expectations that participants can have about each other's choices.

It is, though, abstract and deductive, not the empirical study of how people make decisions but a deductive theory about the conditions that their decisions would have to meet in order to be considered "rational," "consistent," or "noncontradictory." Of course defining "rational," "consistent," or "noncontradictory" for interdependent decisions is itself part of the business of game theory. Take the case of the man whom we do not want to meet in the diner: could there be a theory that tells us unequivocally which diner to choose in order not to meet him? Only if we deny our opponent access to the theory. If logic could tell us which diner to choose, the same logic could tell him which diner we would choose, and as von Neumann and Morgenstern said in the monumental work that launched game theory nearly four decades ago, we can hardly be satisfied with the generality of any theory whose success depends on its not becoming known!

Strictly speaking, this kind of theory is not predictive. It is what is sometimes called "normative" theory in contrast to predictive or explanatory theory. Still, it is doubtful whether theorists would put forth so much energy and receive so much attention if their deductions were not felt to provide some bench mark for the analysis of actual behavior. This method, which might be called "vicarious problem-solving," has been traditional in economics; for the study of how business firms maximize profits, even for the study of whether they try to, it is helpful to know how they would behave if they actually tried and succeeded.[1]

Solving the Problem

Let us look at how the problems that began this chapter are approached through game theory. First, with an exception to be noted later, from the point of view of game theory none of these

problems involves dining cars. The dining cars are merely an interpretation; the man I did not want to meet in the diner could as well be a disarmament inspector along whose route I do not want to leave evidence of violation, or a submarine commander about to fire a torpedo in the direction he thinks my ship will go. Second, the problem does not involve particular individuals; game theory eschews solutions based on personal idiosyncrasy or the ability of one individual to outguess another. Third, in game theory one does not care why the one individual wants to meet the other and the second wants to avoid the first; they are treated as "rational" in the way they try to achieve their goals, but their goals are *their* business, and game theory takes them as data.

In the case of *opposed interests*, if either of us has to make his choice first, in a way that the other can see, the solution is easy: the first loses, however he chooses, and the second wins. This result is trivial but its implications are not. It points to the value of postponing decision, of gaining intelligence about the choice another has already made and denying intelligence in case one has to move first.

This dining-car case is simplified by the occurrence of only two possible outcomes, *meet* and *don't meet*, and using *S* and *F* for success and failure, the problem can be depicted in a 2 × 2 matrix (Figure 1). In the lower left corner of each cell is the outcome from my point of view, I being the one whose decision corresponds to

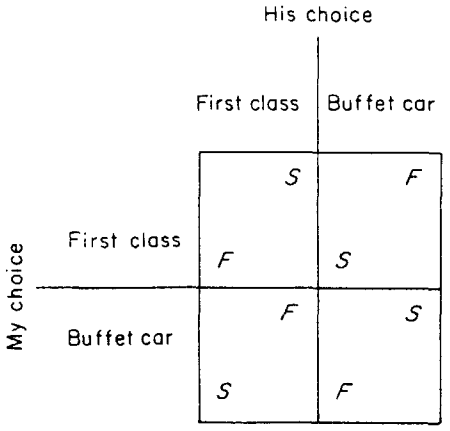

Figure 1

choosing the upper or lower row; and in the upper right corner of each cell is the outcome from his point of view, his decision corresponding to the choice of the left or right column. We can make the problem look somewhat quantitative by using numerical scores in place of *S* and *F*—a 1 for success and a 0 for failure, or perhaps a −1 for failure, choosing numbers for sheer mnemonic and typographic convenience, just remembering that the larger of the two numbers means success. We may as well use the same pair of numbers for both players, although this again is just for convenience. We can now say that each player tries "to maximize his score," but this merely means that he tries to achieve success or to maximize his chances of success.

This is one of the situations that in game theory is known as "zero-sum." It is often described as a situation in which he loses what I gain and vice versa, but actually in game theory the scoring systems of the two individuals are invariably treated as incommensurate. If two feudal noblemen play a game of cards, one to lose his thumb if he loses and the other to lose his eyesight, the game is "zero-sum" (as long as neither cares about the other's loss) though nobody's loss is the other's gain and there may be no way of comparing what they risk losing. It is precisely *because* their value systems are incommensurable that, if their interests are strictly opposed, we can arbitrarily represent them by scales of value that make the scores or payoffs add up in every cell to zero. Visually it is often more convenient to use positive numbers and zeros; the sum then will be some positive number.

What, now, does game theory say about this dining-car or disarmament-inspection or torpedo-target problem that is abstractly represented in our matrix? The reader can probably guess: it says that each participant should have a fifty-fifty chance of succeeding. Why? Because the positions are symmetrical, and in game theory we agree not to pick favorites. Is it quite true that their positions are symmetrical, when one wants to meet and the other wants not to meet? Yes, the same situation arises in matching pennies; one wants both coins heads or tails and the other wants a head and a tail, but if we match a nickel against a penny it is arbitrary whether we call the Indian or the buffalo "head." So we not only eliminate the dining car, we eliminate the concept of "meeting." We can interchange columns in the matrix and get another that is superficially changed but essentially the same;

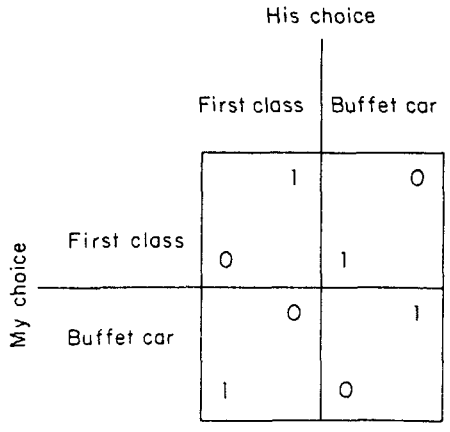

Figure 2

"meet" and "not meet" are only labels, and in game theory we ignore the labels unless there are special reasons for using labels as part of the communication process.

We might say that it is a "toss-up" who wins this game and indeed one may as well flip a coin. But I can flip the coin for either of two reasons: because I just don't care and, like a person who doesn't know which shoe to put on first, want some arbitrary way to decide; or alternatively because if I deliberately flip a coin you cannot guess what I will do, any better than you can guess the toss of a coin. In game theory it is discovered that some games of wits (usually, "zero-sum" games of pure conflict) can be converted into games of chance by appropriate randomization of one's decision.

There is a consistency here: if I flip a coin you can have no better than a fifty-fifty chance at meeting me, and if you flip a coin I can have no better than a fifty-fifty chance of avoiding you. In game theory this fifty-fifty chance of success or failure for each participant is considered the "value of the game," and the "solution." This does not quite say that a person should flip a coin. What it says is that two rational participants, in this situation with alternative outcomes, cannot rationally expect more than a fifty-fifty chance of success unless there are special reasons for supposing that one of the opponents just does not understand the game. If you can think of any line of reasoning by which to choose

one car or the other with a better than fifty-fifty chance of meeting me, I can spoil your strategy by flipping a coin. No mediator could talk the two of us into any scheme that gave odds of less than, or more than, a fifty-fifty chance of meeting, because one of us could always do better by flipping a coin.

Where is all the mathematics? The mathematics is of two sorts. One relates to logical generalization: it is interesting to know whether every problem of this kind has this kind of solution, and what kinds may not. Second, if we complicate the problem it may take some practical mathematics to figure out what kind of coin to flip. Suppose for example that there is one dining car in which your acquaintance is bound to find you but another in which he has only a fifty-fifty chance even if you both go to that car. The latter is like two dining cars coupled together, and to decide where to go you must choose among the equivalent of three dining cars, rolling dice to determine which of the three to go to. He then has one chance in three of finding you, and could himself guarantee one chance in three by choosing one or the other dining car with odds of two to one. Complicate the problem all you please, the principle remains the same; complicate it all you please, and the services of a mathematician or a computer become necessary. The intellectual achievement is in recognizing which complicated problems of disarmament inspection, torpedo fire control, and dining-car selection can be reduced to the general principle of flipping a coin or using random numbers. For generations people presumably chose safe combinations at random in order not to be outguessed by burglars, but it was game theory that saw the same principle (with the odds suitably chosen) in the allocation of a quota of on-site disarmament inspections among the months of a year or the sections of a territory.

Notice that communication is of no significance in this strictly adversary relation. The submarine commander and the captain of the target ship can have no rational interest in sending each other messages; any message worth sending is not worth reading, unless somebody thinks that he is a little smarter than his adversary and can think one step further in a game of mutual deceit.

Alternative Solutions

Now turn to the two friends who want to meet in the same dining car. They succeed or fail together. (If we want symmetrical termi-

nology we can call the situation a "zero-difference game" in exactly the same sense as the pure-conflict situation is called a "zero-sum game.") Their choices are represented in the matrix depicted in Figure 3. Their problem is an "embarrassment of solutions." There are two, and they do not know which to choose. If

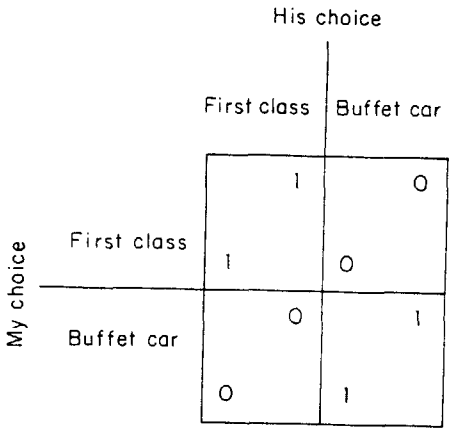

Figure 3

either can move first, letting the other follow, the situation is trivially easy. This is a "team" situation and it takes only one-way communication, or a leader-follower relation, or a "rule" known to both participants, to solve their problem. If they flip coins, they guarantee the same fifty-fifty chance that the adversaries did. What they might do is search for clues; a clue is a kind of signal that each can recognize as an arbitrary instruction worth following in the interest of getting together. Here is the place where "labels" can make a difference, but only as a kind of surrogate for an instruction or a communication. If one dining car is named "The Rendezvous" and the other "Solitaire," they may agree tacitly that they have the signal they need. Members of a squad separated in combat, two people with a lunch date who failed to mention where to meet, or two cars keeping to opposite sides of the road need such clues and signals. Communication makes the problem trivial, but communication is not always available. What is interesting conceptually about this problem is that there are too many "solutions," posing a problem.

Consider now the man who will lose his seat if he goes to the diner. His interest, and that of the man who wants his seat, are neither strictly opposed nor wholly coincident. Both will be better off if the man can reclaim his seat when he returns, because if he can he will eat, and the other will get to sit down for a while. The "solution," if the man would rather sit than eat and has no way of reclaiming his seat, is an *inefficient* one: he goes without lunch, the other stands up all the way. What is needed is a one-way promise that the man who sits down will get up, or an enforceable contract, or a scheme to rearrange the incentives of the man who takes the seat (such as his going to the dining car second, not first, and being hungry enough to vacate his seat when the first returns). Game theory helps to discover some of these "inefficient" situations; it can also try to discover some rules or procedures, legal arrangements, or enlargement of the range of strategies available, to achieve better outcomes for both participants. Game theory also provides a framework for studying the bargaining that then occurs if there are two or more such outcomes and they discriminate differently among the participants.

A Framework for Analysis

So far I have mentioned only some rudiments of game theory, and none of the subtle or elaborate analysis that has attracted the attention of mathematics. But what may be of most interest to a social scientist is these rudiments. The rudiments can help him to make his own theory, and make it in relation to the particular problems that interest him. One of the first things that strike a social scientist when he begins to experiment with illustrative matrices is how rich in variety the relationships can be even between two individuals, and how many different meanings there are for such simple notions as "threat," "agreement," and "conflict." He is struck by how many configurations of information and misinformation there are, how many different communication systems, and what a variety of alternative "legal" constraints on bargaining and tactics. Even the simplest of situations, involving two individuals with two alternatives apiece to choose from, cannot be exhaustively analyzed and catalogued. Their possibilities are almost limitless. For this reason, game theory is more than a "theory," more than a set of theorems and solutions; it is a framework for analysis. And for a social scientist the framework can be useful

in the development of his own theory. Whether the theory that he builds with it is then called game theory, sociology, economics, conflict theory, strategy, or anything else is a jurisdictional question of minor importance.

Consider two individuals with two choices each, four possible outcomes. For each participant, rank the four outcomes from first choice to fourth, without yet using numbers to represent the intensity of preferences; eliminate ties, that is, assume that no two outcomes are equally attractive or unattractive for either of the participants. How many different 2×2 situations can we get? The answer is 78. Furthermore, in 66 of these situations the positions of the two participants are different; and there are a total of 144 different positions a man can be in vis-à-vis his partner.

This number is large enough to surprise most people; but if it seems manageably small, we need only to make allowance for some tied preferences and the number of distinguishable 2×2 matrices exceeds a thousand. Just give each participant three alternatives to choose among, rather than two, with nine outcomes that can result from the joint decision, and the number of distinguishable positions a man could be in vis-à-vis his partner is more than a billion. That is to say, if we prepare a table with three rows and three columns and put, in each of the nine cells, one of the numbers from one to nine for the player who chooses column, and similarly for the one who chooses row, there are more than a billion different ways of inserting those numbers, even after we eliminate all the duplications that result from arbitrarily rearranging rows and columns. (To be more exact: the number is $[9!]^2 \div [3!]^5 = 3{,}657{,}830{,}400$.)

No wonder there is no exhaustive catalogue of even the simplest kinds of interdependence that can exist between the decisions of two people. Add a third person, or add for each person his estimate of the other person's preferences, or add an opportunity for one person to make his choice conditional on the other's choice, and the number of different possibilities quickly becomes astronomical. Let the population explosion go to any imaginable extreme and form all the possible pairs of human beings on this planet; there will not be enough pairs to illustrate the full variety of the situations that can occur when two people contemplate between them a dozen possible outcomes they jointly determine by choosing, in a brief sequence of moves, among three or four alternatives each.

These numbers are not meant to daunt the theorist but to encourage him. Since a definitive catalogue of even the simplest situations and their analyses could not be physically provided nor humanly read if it could be, and since evidently not all differences are important differences, one needs a system, or some criteria, for handling whole classes of situations that, though different, need not be distinguished. One needs to identify the models that have the greatest generality or some unique interest. And one needs a few theorems that permit him to make general statements based on a few salient characteristics of a model, without having to examine all the possibilities.

Some Illustrative "Moves"

The use of matrices and explicit preferences can be helpful both in discovering and in communicating distinctions that need to be made (and in recognizing false distinctions or inessential ones). How does one distinguish a threat from a warning? How does one distinguish the potency of a threat from its credibility? How does one distinguish a bluff from an insufficiently credible threat? When does a threat need to be coupled with a reassurance to be effective? In what situations can both parties be interested in threats, in what situations can only one party have an interest? When is misinformation of value to both parties, when is it of value to one party, and when harmful to both? What is the minimum communication system required for the effectiveness of a threat, of a promise, of a threat coupled with a promise; and what kinds of insurance against failure will enhance the credibility of a threat, what kinds will degrade it? What definitions break down, or have to be replaced by more complicated notions, if the number of relevant alternatives increases from two to three, or from three to some larger number?

It turns out that many of these concepts and distinctions can be operationally defined by reference to an explicit "payoff matrix" that shows the preferences of two parties among the several outcomes. It also turns out that some cannot, and it is useful to see explicitly why they cannot. Some concepts can be operationally defined, and quite simply represented, as a change in a single number or preference ranking in a single cell of a matrix; some can be defined as simultaneous changes in two or more of the payoffs—two payoffs of the same person in different cells, or one payoff for each of the players.

This is hardly high-powered theory, and surely does not yet involve mathematics, but it can lead to discoveries and it can reduce ambiguity in communication.

We can make threats that are bluffs or bets that are bluffs: does "bluff" have the same meaning in both cases? My dictionary says that to bluff is to frighten someone by threats that cannot be made good. What about "will not" be made good? Is there a difference? What is it if I make a threat that I want you not to believe will be made good? Am I bluffing if I try to make you underrate either my capability or my willingness to do what I said? As von Neumann and Morgenstern pointed out, in situations like poker one may not only bluff to win an occasional hand on poor cards but also, quite rationally, bluff to be occasionally caught bluffing, so that a partner may think one is bluffing when one is not and put more money in the pot. It is extraordinary how rich in alternative meanings some of these apparently simple concepts are; the surest way I know to identify the necessary distinctions, to get away from verbal ambiguities, even to discover significant motives and actions that one had not thought of, is to use some of the rudimentary paraphernalia of game theory in making a model that one can manipulate.

Another superficially simple concept is *immunity*. An important problem in a rebellious area is to get people to give information that they want to give but are afraid to. The same problem arises in getting blacks to testify when their rights have been violated, or hotels integrated whose owners are afraid of reprisal. Medical authorities have the same problem in getting drug abusers to seek medical advice, since disclosure of the addiction makes the patient subject to prosecution. Grand juries often have to grant a witness immunity from self-incrimination. (A committee can even give an immunity that a witness does not want, to deny him the excuse that otherwise resides in the danger of self-incrimination.) In elections the secret ballot is mandatory, not an optional privilege, so that no one can give evidence of how he voted and thus cannot be made to comply with a bribe or a threat. This concept of immunity is susceptible to formal analysis, and the analysis could lean on some of the concepts and techniques of game theory. The situation is a "game" of n persons, where n is typically three or more but can be as small as two; there are payoffs to be identified, channels of communication and a structure of information, a distinction between verbal communication and evidence, and a set of choices that go in a certain sequence. There are alternative ways of providing immunity, such as privacy, protection, and coercion. Privacy can be personal or statistical; the protection can be based on defense against third parties or deterrence of them; the coercion can be secret or it can be made visible to third parties to discourage countercoercion. These situations do not especially belong to economics, law, political science, criminology, strategic intelligence, or any of the traditional disciplines; it cuts across them.

Still another example is the interesting subject of locks, alarms, warnings, and safety catches. We usually do not need much theory to help us buy a lock for the garage door, but a lock on nuclear weapons is rich in its theoretical possibilities. There are many kinds of locks and many motives, and even a classification of them requires something that looks a little like game theory. A lock on radium in a doctor's office has, among its purposes, the anomalous one of protecting the thief himself. A lock on the bathroom door is intended to keep people out who prefer to stay out and is equivalent to a sign saying "occupied"; and in bathrooms in some new buildings, to keep children from locking themselves in, there is an anomalous lock that can be unlocked from either side of the door. A lock on an ammunition chest may be designed to keep the contents from being used by somebody, and a mechanism that destroys the contents when the box is violated is almost as good as one that keeps the thief out; if the lock is to keep someone from destroying our ammunition, though, a destruct mechanism merely eases his task. A lock that makes the ammunition explode when the mechanism is joggled will not protect the ammunition if it works secretly, but if the burglar knows that it will explode in his face it can deter him. Some locks are designed only to measure the urgency of entry and are designed to give way under stress; fire alarms and emergency brakes are protected by a piece of glass to which a small metal hammer is conveniently attached. Some locks are meant to catch the intruder by blocking escape, some to catch his identity by photograph, some merely to report his intrusion by giving an "alarm," and they are hidden or made conspicuous according to whether one wants to trap the burglar or to deter him. And some, like the time lock on a bank vault, are designed to keep the owners themselves from being able

to open them, so that they are immune to coercion during times of day that the place is unprotected.

And so on. Similar problems arise in handling confidential information, the reaction to a radar alarm and the authority to launch warfare, systems to guard the legal rights of apprehended suspects, and disciplinary systems. What we are discussing is devices or institutions that can be construed as a "move" in an *n*-person game, where interesting values of *n* may be anywhere from one up to half a dozen and the game can profitably be described by reference to the payoffs, the information structure, and the strategies available to the participants. The garage door, as I said, may be an easy one, but designing an appropriate device for a nuclear weapon, a fallout shelter, or an ammunition convoy in Vietnam requires explicit attention to the rich array of alternatives, the tradeoffs and compromises, the probabilities of contingent events, the relative magnitudes of payoffs, and what needs to be communicated, what guarded against revelation. The richness of the problem, and the value of explicit analysis, is occasionally brought home to us on those occasions when we lose a credit card, lock ourselves out of a house, or can't find something we hid to keep it away from the children.

I am not trying to advertise something called "game theory" that will provide instant insight into these interesting problems, but rather to illustrate the kind of problem that stimulated the development of game theory and to show how ubiquitous these problems are.

Voting Strategy as an Example

Voting schemes provide nice illustrations of the domain of game theory. Voting is notorious for inviting strategy—the calculation of how one ought to cast his ballot in view of how others may cast their ballots. Someone who dislikes public housing may vote in favor of a civil rights amendment he despises, knowing that only with the amendment can the bill itself be killed on a subsequent ballot. Voting also invites coalitions; and implicit coalitions can be exploited by designing "package" proposals to be voted all at once as a means of enforcing the coalition. "Packages" eliminate alternatives; a rule obliging the President to enact or to veto an appropriations bill in its entirety permits the Congress to exploit the President's preferences.

I am going to work through an example, and to keep it simple I shall restrict the number of voters to two. I can do that if I use a rule of unanimity. You and I are members of a two-man committee to determine the career of an employee who would normally be considered for promotion but has been charged with a blunder that he might be fired for. Our committee has to decide two things. First, is the man's overall record so excellent that, leaving the blunder aside, we ought to promote him? Second, is he guilty of this blunder? If his record is excellent and he is innocent he will be promoted; if his record is only ordinary and he is guilty he will be fired. If we find him guilty, but find his record excellent, he will be demoted but not fired; if we find him innocent but of ordinary record he will be kept but neither promoted nor demoted.

I have been through the evidence and reached the conclusion that, all things considered, the man ought to be demoted, but I'd rather keep him than fire him, even promote him than fire him. You are convinced the man ought to be fired; if you can't fire him you'd like him demoted, least of all promoted. Under the rules we must vote on both issues, his record and his innocence. Under the rules, it takes two to find him guilty of the blunder, and two to award him an excellent. Under the rules, we do not vote whether to promote, keep, demote, or fire the man; we vote these two issues.

The normal procedure is to vote first on guilt or innocence and, having that out of the way, to proceed with whether his record is excellent or ordinary. If both of us prefer, however, to take his record up first, we may. So together we first vote on which question to take up first, unanimity being required to take up his record first.

Both of us are interested in the *outcome*, not in the abstract notions of innocence and excellence. And both of us have made no secret of our preferences.

We can sketch this problem in the form of a branching tree (Figure 4).

There are eight ways that the balloting can go in arriving at one of these four results. The first branching point at the top determines which issue is taken up first; the second determines the answer on that first issue and the third the issue voted last. The numbers in the sketch refer to how many votes it takes to determine the choice; at the top of the diagram the 2 means that it

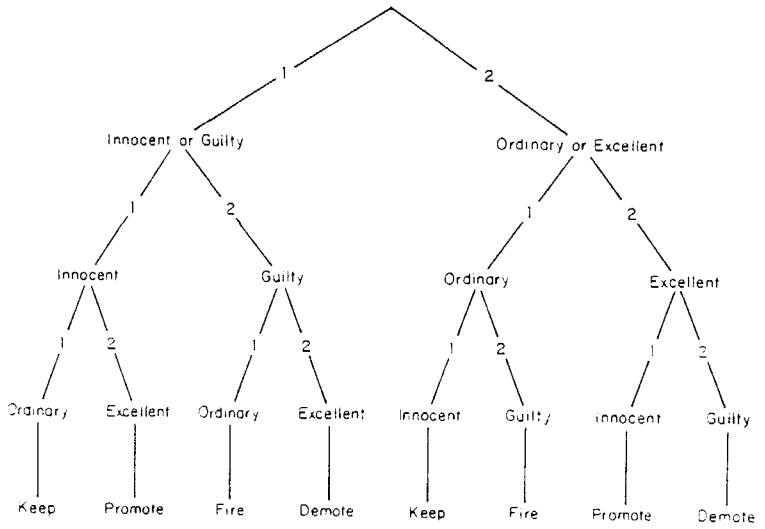

Figure 4

takes two favorable votes to elect the right-hand procedure (under which his record is voted before his innocence or guilt), the 1 means that it takes only one vote to have his guilt taken up first. It would take three unanimous ballots to reach the right-hand outcome where the man is promoted; either of us can bring about the outcome on the left because it takes only one vote for guilt to be considered first, one vote to find him innocent, and one vote to deny him an excellent rating. Under one procedure if I vote him guilty you can alone decide to have him fired; under the alternative procedure, if you vote him a good record I can get him promoted by finding him innocent. How should we expect each other to vote?

One way to work this problem is to start from the final votes and work up. At the far left (at the point reached if either of us votes for the normal procedure and if either of us votes him innocent) the final ballot (nominally on his excellence) is a choice between *promote* and *keep*, and it takes two to promote him. Evidently we'll both vote to keep him. At the final vote second from the left the choice is between firing and demoting, and it takes two to find his record excellent and thus to demote him; you prefer to fire him, and your vote would do it. Foreseeing this, at the preceding stage when we vote innocent or guilty, we know that

the choice is between keeping him and firing him, so I shall vote him innocent, after which we shall both find his record ordinary. This means that if either of us supports the normal procedure on the first ballot the result will be that the man is kept.

Similarly, on the far right if we have found him excellent, we shall both vote to demote him; if instead we have found him ordinary I shall vote to keep him. So when we vote on his rating we know we are voting to demote or to keep him, and we both vote "excellent" in order to demote him at the next stage.

So, if we both vote to reverse the normal procedure and take up his record first, we can expect the man to be demoted. Since we both prefer his demotion to his merely being kept, we should both vote to reverse the normal procedure, then to find his record good, then to find him guilty.

There are several points to note. First, *the procedure makes a difference;* the man is demoted or merely kept according to which of the two questions we vote first. Second, one of the two procedures is *less satisfactory for both of us* than the other procedure, even though our interests do not coincide. Third, the reason why voting first on guilt or innocence leads to this less satisfactory outcome is that I must expect you to find his record poor after we both find him guilty. Because I do, I have to find him innocent. It is your power of decision on the final ballot that diverts me down another branch, to an outcome that we both like less than demotion. If you could promise in advance to vote his record good, I could go ahead and vote him guilty and we'd both be better off. The alternative procedure, down the right-hand branch, can be thought of as a way for you to give me that promise; by voting a good record in advance, you deny yourself the possibility to get the man fired after I vote him guilty, leaving me free to vote him guilty.

Each of us would have to reexamine his strategy if the other's preferences were switched. If you knew that I really wanted him promoted, for example, you would not dare to vote as you just did; nor would I if I knew that you wanted him promoted.

A Matrix of Strategies

One technique of game theory is to identify all of these "strategies"—all of the different contingent plans that the voter may have for deciding along the way how to vote next. If I had to be absent, and sent a deputy to represent me, I could not simply tell him how to vote on each ballot. Each vote should depend on how

the preceding ballot went. I can, however, if I'm willing to be sufficiently explicit, anticipate all possibilities and tell my deputy what to do in each case that could arise. I could say, for example, "Vote yes on the first ballot. If that loses, vote no on the next two ballots; but if it wins, vote yes on the next ballot and yes again if it wins or no if it fails."

This is a *sufficient* instruction; it tells him how to do everything I would have done as the situation unfolds. In the language of game theory, this is a "strategy." Every such contingent instruction, if it covers all possible contingencies, is a "strategy." In this voting problem, the number of different strategies is limited, and any advance plan or instruction that covers every contingency can be thought of as a selection of one strategy from among all the possible strategies. If we identify all of the alternative strategies, we can construct a matrix consisting of all possible plans that the two of us might have, and thus convert our dynamic sequential problem to a static simultaneous-choice equivalent, in which I mer)ely choose a strategy in advance, considering all the strategies open to you, and you do the same, and the outcome is the joint result of these two strategies.

To see how this is done, without cluttering the page with too large a matrix, suppose that we have already voted to reverse the usual procedure and to take up excellence first, and are about to decide on our remaining strategies. Since a no vote is decisive while a yes vote can carry or lose, I have one completely definite strategy: voting no on both ballots with the result that the man is kept, independently of how you vote. I also can vote no on the first ballot and yes on the second; if I want him fired, this may be a way of achieving my aim. If I vote yes on the first ballot there are four possible plans I could have for continuing: (1) to vote him guilty whether or not the first ballot finds his record excellent; (2) to vote innocent however the first ballot comes out; (3) to vote guilty if his record is found excellent, otherwise innocent; and (4) to find him innocent if his record is found excellent, otherwise guilty. I have, then, a total of six possible ways of playing the game when two ballots remain. You have the same alternatives, so there are thirty-six different ways our contingent plans can combine in reaching one of four possible outcomes. These are shown in Figure 5.

The numbers have to be explained. To represent my prefer-

Your strategies

My strategies	No No	No Yes	Yes No	Yes No/ Yes *	Yes Yes	Yes Yes/ No **
No, No	1 / 2	1 / 2	1 / 2	1 / 2	1 / 2	1 / 2
No, Yes	1 / 2	3 / 0	1 / 2	3 / 0	3 / 0	1 / 2
Yes, No	1 / 2	1 / 2	0 / 1	0 / 1	0 / 1	0 / 1
*Yes, No/Yes	1 / 2	3 / 0	0 / 1	0 / 1	0 / 1	0 / 1
Yes, Yes	1 / 2	3 / 0	0 / 1	0 / 1	2 / 3	2 / 3
**Yes, Yes/No	1 / 2	1 / 2	0 / 1	0 / 1	2 / 3	2 / 3

*Vote *yes*, followed by *no* if it carries, *yes* if it fails.

**Vote *yes*, followed by *yes* if it carries, *no* if it fails.

Outcomes:

Promote, [0 / 1], Keep, [1 / 2], Demote, [2 / 3], Fire, [3 / 0]

Figure 5

ences I have arbitrarily given a score of 3 to *demote*, 2 to *keep*, 1 to *promote*, 0 to *fire*. Since your preference order is *fire, demote, keep, promote*, I've scored you with 3 if he is fired, 0 if he is promoted, and so on. These numbers just remind us of our preference order, the magnitudes do not matter. (A little later we shall see where numerical values would make a difference.)

Neither of us has a "dominant" strategy, that is, a strategy that he would be satisfied to have chosen no matter what the other chose. Row 6 looks good to me unless you choose Column 3 or 4, in which case I'd rather have chosen Row 1. Column 5 looks good to you if I choose Row 2, bad if I choose Row 3 or 4, pretty good if I choose Row 5. There are some columns you might choose that leave me indifferent—Column 1, for example. There are some columns in which my score can be anything from 0 up to 3 according to what row I pick.

Though no row or column is an obvious "best" choice, we can still ask whether there is a pair of expectations we can have about each other that will lead us to choices that confirm those expectations. Is there a column such that if I expect you to choose it I will choose precisely the row that, if you expected it, would lead you to choose that column? Yes, Row 6 and Column 5 have that "equilibrium" property. If I expect you to choose Column 5, I am content with Row 6, and if you expect me to choose Row 6 you are content with Column 5. We cannot quite say that I "prefer" Row 6 when you choose Column 5, because I would do just as well in Row 5, but if you expected me to choose 5 you would choose Column 2. The intersection of Row 6 and Column 5 is an "equilibrium point," or an "equilibrium pair" of strategies. It has the property that if we both make the corresponding choices, each expecting the other to do so, each has behaved correctly in accordance with his expectations and each has confirmed the other's expectations.

Furthermore, the intersection of Row 6 and Column 5 is an "efficient" outcome, as economists use the term. There is no other cell in the matrix that can improve the outcome for one player without worsening it for the other. The same cannot be said for the cell in the upper left corner, which is also an equilibrium point but a weak one. (It is a "weak" one, a kind of "neutral equilibrium," because neither of us has an actual preference for that cell above any others in the corresponding row or column.)

If we draw up the corresponding matrix for the two-stage ballot under the normal procedure, with guilt or innocence being decided first, we get the matrix in Figure 6.

This matrix differs from Figure 5 in several ways. One is that you now have a dominant strategy: Column 3 in every row is as good as any other column and sometimes better. You can eliminate the other 5 from consideration. Since you can, I can assume you will, and I choose Row 1 or 2.

But though 3 dominates, your outcome is not especially favorable. Knowing your choice, I pick a row that gives me a score of 2 and you but 1. You cannot wish that you had chosen differently, all you can wish is that I could have expected you to. Then I might have chosen differently.

If Columns 3 and 4 could be suppressed, I would have a dominant strategy, Row 6, and you could choose Column 5 or 6, and both of us would be ahead. But in the matrix as it stands, the two of us cannot hold a consistent pair of expectations that would lead us to Row 6, Column 5. This pair of strategies has not the equilibrium quality; there is no line of reasoning by which we can reasonably expect each other to expect it.

The Complete Matrix

The very first ballot, then, deciding the order in which to take up the two questions, can be construed as a ballot for deciding which of these two matrices to confront. We could of course construct a matrix corresponding to the whole three-ballot game. It would be hard to get on a single page, but we can at least ask what it should look like.

How many rows and columns would it have? A complete strategy has to indicate how to vote on the first ballot and how to vote thereafter in either of two cases. Since it takes two of us to reverse the normal procedure, one to keep it, a vote of no on the first ballot need only be coupled with a choice of a row (or column) in the matrix (Figure 6) corresponding to the left-hand branch. So there are six complete strategies corresponding to a vote of no on the first ballot. If I vote yes on the first ballot, my strategy must specify a row in each matrix, since I shall have to choose a row in whichever matrix the first ballot selects. There are, thus, 36 possible strategies containing a yes vote on the first ballot. Altogether,

Your strategies

My strategies

	No No	No Yes	Yes No	Yes No/ Yes *	Yes Yes	Yes Yes/ No **
No, No	1 2	1 2	1 2	1 2	1 2	1 2
No, Yes	1 2	0 1	1 2	0 1	0 1	1 2
Yes, No	1 2	1 2	3 0	3 0	3 0	3 0
*Yes, No/Yes	1 2	0 1	3 0	3 0	3 0	3 0
Yes, Yes	1 2	0 1	3 0	3 0	2 3	2 3
**Yes, Yes/No	1 2	1 2	3 0	3 0	2 3	2 3

*Vote *yes*, followed by *yes* if it carries, *no* if it fails.

**Vote *yes*, followed by *yes* if it carries, *no* if it fails.

Outcomes:

Promote, [0 / 1], Keep, [1 / 2], Demote, [2 / 3], Fire, [3 / 0]

Figure 6

then, there are 42 strategies for me and the same number for you. This 42 × 42 matrix has 1,764 cells, each containing one of the four outcomes. What else do we know about it without taking the trouble to draw the matrix?

We know, without any more theory, that the outcome is bound to be asymmetrical; no outcome has the same rank in our two preference scales. We might guess, and with a little more theory we would know, that this large matrix shows an equilibrium pair of strategies corresponding to a yes on the first ballot for both of us and Row 6, Column 5, of the matrix in Figure 5. That is, the solution we arrived at by working backward from final outcomes corresponds to an equilibrium pair in our larger matrix.

Actually there is a further characteristic that game theory tells us to expect. There is at least one "dominated" row or column in the matrix—a row or column that is inferior to some other row or column in at least one cell and nowhere superior. If we strike out dominated rows and columns, compressing the matrix, we shall still find dominated rows and columns (because some that were originally not dominated are dominated after the eliminations). We can go on doing this until the residual matrix contains only cells with the *demote* outcome. Game theory is interested in which kinds of problems generate matrices that have various properties, like this one.

A few more things can be observed in this example. One is that a "dominant" strategy is not necessarily a good strategy *to have*. It is necessarily a good strategy *to play*, because no matter what the partner does the dominant strategy proves never to have been inferior to any other choice. But its mere availability can induce the other player to make a choice that condemns one to a poor outcome.

Another point, not illustrated in our matrix, is that in general a matrix need not show an equilibrium pair of strategies. It may show more than one; if it shows more than one they may differ, and they may differ by both payoffs' being lower in one cell than another or by one payoff's being lower, the other higher. (Game theory also tells us that if a matrix shows no equilibrium pair of strategies, one can be generated by a randomized choice, with suitable odds, among some or all of the strategies in the matrix; this procedure, though, requires that there be a suitable interpretation of the numerical values of the payoffs.)

Collective Decisions

This voting example illustrates that a "gamelike" situation can be viewed as a *collective-decision* process—a process by which two or more individuals jointly decide on an outcome. The analysis also has ethical implications: we assumed the voters to be concerned with outcomes, not with strategies for their own sake; with consequences, not actions; with ends, not means; with justice, not truth. A voting scheme also illustrates how the organization of authority, leadership, and bargaining arrangements can affect the outcomes—can affect whether an outcome is efficient, can affect in whose favor the outcome discriminates. And evidently if we had been willing to enlarge our committee and have a majority-vote procedure, coalitions would have been important; communication might have been important to coalitions, and so might discipline. And it is evidently important what people know, or think they know, about each others' preferences.

The "legal" arrangements are important. If binding promises can be enforced, the alternative voting procedure is unnecessary: you just promise to vote the man excellent if I will join in finding him guilty. In fact, the first ballot can be thought of as a "bargain" that you have an incentive to keep because I have a credible incentive to vote for promotion if you back out of the bargain.

Probabilistic Uncertainty and Numerical Preferences

The numbers in our matrices had only ordinal significance. To illustrate how the numerical values could assume importance, and how numerical values are assigned in game theory, suppose that any award of excellence or verdict of guilty is subject to a review procedure that we believe to have only a fifty-fifty chance of confirming our unanimous vote. If a man is found innocent and excellent there is a fifty-fifty chance that he will be kept or promoted; if a man is found guilty and excellent there are equal probabilities of 0.25 that he will be promoted, kept, demoted or fired.

To handle the problem we now need a more complicated set of preferences. It is not enough to know that I prefer demoting to keeping the man, keeping to promoting, and promoting to firing. We now have to know whether I'd rather keep the man or take a fifty-fifty chance between demoting and firing. And we may have to know whether I'd prefer a fifty-fifty chance between demoting and firing or a four-way split over the four outcomes. We can assume a few things, such as that if I prefer demoting the man to keeping him I prefer a fifty-fifty chance between demoting and keeping to the certainty of keeping him, and prefer any odds between demoting and keeping to any odds between promoting and firing.

Two points are worth mentioning. First, not only can these "critical odds" or "critical risks" be subjected to certain consistency postulates in a way that may permit us to go ahead and solve our problem, but it even turns out to be possible and convenient (though not necessary) to derive numerical values for the different outcomes from a limited number of expressed critical-risk preferences. These numbers can be operated on *as though* one were trying to maximize the mathematical expectation, that is, the expected value in a probabilistic sense. One can alternatively just postulate that a decisionmaker associates numerical values with all the outcomes and tries to maximize expected value; but the postulate need not be that heroic. It needs only to be that he can answer a few simple questions like those we asked earlier about the critical odds between a pair of outcomes that would make him just willing to settle for the certainty of a third that lies between the other two. If our man then obeys a few other "consistency" rules to avoid some kind of contradiction, we can often handle the problem. For convenience we can attach numerical values to outcomes, based on these critical odds, even calling these numbers "utilities" or something of the sort, but this is only a convenience for combining and compounding a limited set of expressed preferences in the form of critical odds.

The second point is that the need for numerical values arises only in the presence of uncertainties of this sort (and only when the number of alternative outcomes is at least three), when one has to place his bets in a probabilistic environment. (The uncertainty may be about another's choice or, in case of deliberate randomization or faulty control, about one's own.) If there is no such uncertainty, numerical values prove unnecessary (as they were in our original voting situation). And in the face of uncertainty one *has* to make choices of this kind, so it is not an outlandish assumption that one actually can. "Numerical utilities," though often thought unique to game theory, are by no means peculiar to game

theory; they arise in the same fashion in any theory of decision under uncertainty.

These numerical values are arrived at separately for all the participants, and there is no intended interpersonal comparability among value scales. In some calculations it may appear that arithmetic is done on the numerical values of two or more players together, but it invariably turns out in game theory that an expression involving the "utilities" of two participants contains only *ratios of increments,* from which any units of measure would cancel out.

It is of some philosophical interest whether the value scales of two individuals are *assumed* inherently incommensurable, or instead we just mean that we don't yet *know how* to compare them. Game theory typically assumes the first position. Some writers treat this as a limitation of the theory and look forward to some way to compare the scales of value between people. I know of none, though, that has indicated how he would use such knowledge if it were available. Just as absolute-cost comparisons in international trade are unnecessary and usually meaningless—the notion of "comparative advantage" or "comparative cost" being sufficient to solve every problem of interest in international-trade economics—the notion of *comparative* ratios of *utility increments* (in which any absolute scales would cancel out) is sufficient in game theory. In fact, so far as game theory is concerned, there really are no "utility scales" to compare. There are only preference rankings among outcomes that have to incorporate numerical probabilities when some of the outcomes themselves are probabilistic. To say that a rational individual "maximizes utility" is a little like saying that nature "conserves" momentum or that water "seeks" its own level. These figures of speech save a lot of circumlocution; but when we forget that they are figures of speech and try to compare actual measures of utility, or to measure the "frustration" of water when a valve opposes it, it is time to abandon the metaphor and get back to operational statements.

An Apotheosis of "Rationality"?

The question is often raised whether game theory restricts its empirical applicability by postulating mental giants with nerves of steel—perfectly rational amoral deciders who have access ex officio to the theoretical results of game theory.

The answer is: not quite. In principle there is no difficulty in imputing misinformation rather than true information, in supposing that calculation is costly or that people make mistakes or suffer from bad memories or display idiosyncrasies in their choices. In our voting scheme, for example, we can easily suppose that when a man votes on excellence he cannot remember whether or not a vote has already been taken on guilt or innocence; and in fact our review-board procedure can easily be interpreted as the likelihood that a vote will be recorded wrong or that one of the voters will shy away from the word "guilty" for unconscious reasons.[2]

But to handle these departures from perfection one has to specify them explicitly. And it greatly complicates the problem to depart from perfection, whether it be perfect memory or perfect absence of memory, perfect knowledge or perfect absence of knowledge, perfect calculation or perfectly random choice. The man with the perfect memory and the man without a memory are the easiest to handle in abstract analysis. To allow for an imperfect memory requires that we specify precisely how his memory misbehaves (and whether he knows how it misbehaves, whether his partner knows how it misbehaves and whether he knows whether his partner knows how it misbehaves, and so forth). Pretty soon we are tempted to give him either a perfect memory or no memory at all, or perhaps to provide him a simplified and idealized "imperfect" memory such that exactly half the time he forgets everything, knows that he does, and his partner knows it too.

But this is not a limitation of game theory; it is a limitation of any theory that tries to deal with the full multidimensional complexity of imperfect decisionmakers. Game theory usually does assume perfect knowledge or perfect absence of knowledge, because these are simple and unambiguous assumptions to make. Anything between the two extremes requires detailed specification, and game theorists can at least be forgiven for solving the simpler problems first and saving the more complicated ones for later.

Game theory usually supposes a few other things, such as that a man's ethics are what have recently been called "situation ethics"; he is concerned with *outcomes*, not intermediate processes. (In our voting example he is not seeking "truth" as to guilt or excel-

lence, but defines justice in terms of what is done with the man.)[3] The decisionmaker is assumed not trying to be bold or novel for the sake of boldness or novelty, not trying to surprise us for the sake of surprise itself; he is not concerned with *why* his partner may choose a particular strategy, but what strategy his partner will choose. Nothing but the *outcomes* enter his value system. If a man has good will or malice toward his partner, a conscience or a bent for mischief, it is all assumed to be reflected in his valuation of the final outcomes. It is assumed that all the elements of his value system are displayed—everything that matters to him is allowed for—in the ranking or valuation of cells in the matrix.

How much a limitation this is depends, as in any theory, on whether an abstract, somewhat perfectionist bench mark can be helpful, and whether we can keep in mind that the result is only an abstract perfectionist bench mark. Newton's laws don't work if atmospheric resistance is present; purely inertial motion is hard to observe in the earth's gravitational field; some voters are shrewd parliamentarians, some are naive or inept. Game theory runs the same danger as any theory in being too abstract, even in the propensity of theorists to forget, when they try to predict or to prescribe, that all their theory was based on some abstract premises whose relevance needs to be confirmed. Still, game theory does often have the advantage of being naked so that, unlike those of some less explicit theories, its limitations are likely to be noticeable.

Games, Theories, and Social Science

A word needs to be said about the name of this discipline, "game theory." The name has frivolous connotations. It is also easily confused with "gaming," as in war gaming, business gaming, crisis gaming—confused, that is, with simulations of decision or conflict.

The name arises from the observation that many parlor games have the key quality of interdependence among players' decisions. The best move in a chess game, the best way to bid or the best card to lay down in a bridge game, depends on what one's opponents are likely to do, even on what one's partners are likely to do. Furthermore, these games are usually well defined; there is an explicit and efficient set of rules; the information available to the players is specified at every point (even if in a probabilistic sense);

and the scoring system is complete. If we had a more general name for the subject now known as "game theory," it would be found that a great many parlor games fit the definition. It was this that led the authors of the first great work in the field to call their book, *Theory of Games and Economic Behavior*, and "game theory" stuck like a nickname.

Decades of usage have got professionals so used to the name that they occasionally forget that "game" is not only a technical term but a word in the English language. If they say that war is a *game,* elections are a *game*, industrial disputes or divorce negotiations are *games*, they usually have nothing playful in mind but are merely using a term that grew out of the recognition that some games, too, are *games*.

There is another problem of nomenclature: *game theory* already has the word "theory" in its name. We find it useful to draw distinctions between economics and economic theory, statistics and statistical theory, decisions and decision theory; but there is no accepted name for whatever the field is of which "game theory" refers to the theoretical frontier. Most game theory in fact has been substantially mathematical; some people prefer even to define it as the application of mathematics to this subject, and any bibliography of the discipline is dominated by accomplished mathematicians. Often the mathematicians have been more interested, for natural professional reasons, in mathematics than in law, social structure, diplomacy, economics, or sociology. Game theorists, and social scientists who deal with the subject of which game theory is the mathematical frontier, are out of touch with each other in a way that, say, economists and economic theorists are not, for a number of reasons including, often, the absence of a sufficient common interest to keep them in touch. The mathematical barrier is not the only one. There is an unusual dichotomy between the subtle, elegant, mathematical accomplishments of game theorists and the interests of social scientists.

Nothing in this essay begins to describe what mathematical game theorists actually do or even to give the flavor of it. For the social scientist, what is rudimentary and conceptual about game theory will be, for a long time, the most valuable. And it will be valuable not as "instant theory" just waiting to be applied but as a framework—one with a great deal of thought now behind it—on which to build his own theory in his own field.

Take the payoff matrix. This is hardly "theory," although a

good deal of theory underlies the definition of strategies and the interpretation of payoffs. Yet by itself, as a way of identifying alternatives and ordering choices, of laying out the structure of a situation to facilitate analysis, comparison, and communication, the payoff matrix may be, for the analysis of interdependent decision, what double-entry bookkeeping was for accounting, national-income accounts for economics, the truth table for logic, or even the equation for mathematics.[4]