1 Principles of Rational Decision

So also, the games in themselves merit to be studied and if some penetrating mathematician meditated upon them he would find many important results, for man has never shown more ingenuity than in his plays.

—G. W. Leibniz (quoted in Ore, 1960)

Two Paradigms of Rational Decision

There are two well-developed theories of rational decisionmaking: the theory of coherent individual decision of Ramsey, de Finetti, and Savage, which is based on the principle of *maximum expected utility*, and the theory of games of von Neumann and Morgenstern, which is based on the concept of *equilibrium*.[1] In the most satisfactory part of the theory of games, the theory of zero-sum two-person games, von Neumann and Morgenstern showed that a game-theoretic equilibrium corresponds to each player playing his or her *security* strategy—the strategy which maximizes the minimum possible gain (the *maximin* strategy).

Example: Justice as Fairness 1. A group of ten people, of which you are one, is to divide one million dollars. All agree that a just distribution scheme is one that would be chosen by a rational, self-interested agent if he or she were to be one of the recipients but had no information as to which one. Accordingly, each person is assigned a numeral from 1 to 10 by lot and you are chosen to divide the money by numeral among the members. We assume, for present purposes, that utility is proportional to money. If you choose by *maximin*, then you will evaluate each distribution scheme according to the least amount of money you could receive under that scheme. Your unique *security* strategy consists in choosing to divide the money equally. If you choose to *maximize your expected utility*, you evaluate each distribution scheme according to the average of the payoffs to each of the ten num-

bers, weighted by the probability that you have that number. As it is part of the conception of fairness that you have equal probability of holding any number, all distribution schemes look equally good. If, for example, the entire million is to be given to number one and you have a one in ten chance of being number one, your expected payoff is one hundred thousand dollars. This is just the payoff that you would get if the million were shared equally.

What is the proper relationship between the rationality concepts of expected utility theory and game theory? The two theories have developed in the absence of a univocal answer to this question. Luce and Raiffa (1957) suggest in one place that for a decisionmaker facing a situation of risk (where the chances of factors other than the decisionmaker's own choice are known), the proper rule is to maximize expected utility, but in situations of uncertainty (where the chances are not known) one should choose one's security strategy. But this suggestion makes nonsense of the theory of subjective probability which they develop later and which is meant to apply under uncertainty. Shubik (1982, p. 2) makes the cut in a different place:

> The general *n*-person game postulates a separate "free will" for each of the contending parties and is therefore fundamentally indeterminate. To be sure, there are limiting cases, which game-theorists call "inessential games," in which the indeterminacy can be resolved satisfactorily by applying the familiar principle of self-seeking utility maximization or individual rationality. But there is no principle of societal rationality, of comparable resolving power, that can cope with the "essential" game, and none is in sight. Instead, deep-seated paradoxes, challenging our intuitive ideas of what kind of behavior should be called "rational," crop up on all sides.

Can the elusive concept of free will bear the weight put on it here? Must we decide such issues before we know how to interact rationally with a person—or an automaton? Would it not be preferable to have a unified theory of rational action, such that in game situations each player can treat the others as part of nature?

To complicate matters further, both classical game theory and decision theory have been vigorously criticized, notably by Herbert Simon (1957, 1972, 1986), for ignoring computational, procedural, and other bounding aspects of the process of reasoning. I believe that such considerations hold part of the key to the correct view of the relation of game theory to individual decision theory.

Deliberation can be modeled as a dynamic process with informational

feedback, a process that is carried out by deliberators motivated by considerations of expected utility and having finite computational resources. Consideration of games played by such bounded Bayesian deliberators grounds and illuminates equilibrium concepts of classical game theory under certain special assumptions, and suggests how that theory must be modified in situations where these assumptions fail.

This chapter will provide an introduction to some essential concepts of game theory and expected utility theory. Chapters 2 and 3 will show how models of dynamic deliberation can provide a bridge between them.

Expected Utility

The origin of the concept of an *expected value* is contemporaneous with the origin of mathematical probability theory itself. The *utility* concept was introduced later, and went through a considerable evolution before taking its present form.

The mathematical theory of probability was conceived as an instrument for evaluating gambles in games of chance. It was assumed that the natural measure of value of a gamble was the *expectation* of the payoff—with the payoff of each outcome being measured in terms of liquid assets: gold or coin of the realm. The expected value is just the sum over outcomes of the probability of the outcome times the payoff associated with that outcome. For example, consider two gambles on independent flips of a fair coin with the following payoffs:

Gamble 1: 4 ducats if heads; 0 if tails
Gamble 2: 7 ducats if two heads (HH) on two flips; 0 otherwise

Evaluating by expected payoff in ducats, gamble 1 has an expected payoff of $(\frac{1}{2})(4) + (\frac{1}{2})(0) = 2$; gamble 2 has an expected payoff of $(\frac{1}{4})(7) + (\frac{3}{4})(0) = \frac{7}{4}$; and gamble 1 is to be preferred to gamble 2.

This expected value or "moral hope" was from the beginning taken as the correct quantity for assessing gambles. Intellectual effort was focused more on the question of computing the probabilities than on the philosophical justification of the expectation principle. Later on, with the law of large numbers, a frequentist gloss became available: the expected payoff is almost surely the average payoff one would achieve in a very long series of independent trials of the gamble.

The move from expected ducats to expected *utility* was precipitated by a puzzle, known as the *St. Petersburg* paradox after the journal in which Daniel Bernoulli published a landmark discussion of the problem

in 1783. In the St. Petersburg game, a fair coin is to be flipped until it comes up heads. If it comes up heads on the nth toss, you will be paid 2^n ducats. There is no limit to the number of tosses. What is the value of this game?

The St. Petersburg Game

Possible outcome	Probability	Payoff
H	½	2 ducats
TH	¼	4 ducats
TTH	⅛	8 ducats
•	•	•
•	•	•
•	•	•

The expected payoff is the infinite sum: (½)(2) + (¼)(4) + (⅛)(8) . . . = 1 + 1 + 1 . . ., which exceeds any finite value. Should everyone be willing to pay any amount whatsoever to get into this game? Would you? Bernoulli's rationale for a negative answer (following Gabriel Cramer)[2] is that *value* is not properly measured in monetary units, but rather in terms of a theoretical quantity, *utility*. Money, and most other real goods, have declining marginal utility: an extra ducat on top of a 10-ducat gain adds more utility than an extra ducat on top of a 1,000-ducat gain. Bernoulli even suggests a typical utility function: utility = log(money). The St. Petersburg game has finite expected utility given the logarithmic utility function.[3]

> *Example: Justice as Fairness 2.* Again, a group of ten people of which you are one is to divide a million dollars. You are to select a distribution scheme by number, as before, which will maximize your own expected utility under the fairness assumption that all numbers are equally likely to be your number and under the Bernoullian assumption that utility = log(money). The unique maximal distribution scheme here is the egalitarian one with everyone getting $100,000 or 5 utiles (log of 100,000 is 5). The scheme which gives all the money ($1,000,000 or 6 utiles) to number one has an expected utility for you of only 6/10 of a utile. Any utility function with strictly declining marginal utility will give the result that equal shares maximize your expected utility. The concave shape of the utility function here gives the egalitarian conclusions

which depended on preference for security (on the *maximin* decision rule) in the previous example, *Justice as Fairness 1*.[4]

The utility hypothesis did not come to Bernoulli—or Cramer—from out of a vacuum. Utilitarian ideas were in the air, floated by (among others) Francis Hutcheson, professor of moral philosophy at Glasgow and teacher of both Adam Smith and David Hume.[5] Questions about the nature of utility did not much occupy probabilists following Bernoulli but they remained within the province of economists and moral philosophers, who discussed them from an ethical and psychological point of view.

It was generally assumed that utility was a psychological quantity, and that there was no difficulty in principle in comparing the utilities of different individuals. Thus, we have Alfred Marshall's remark (quoted in Savage, 1954) that the declining marginal utility of money is well illustrated by the fact that the rich man will take a taxi while the poor man will walk. It is here assumed as a matter of course that the disutility of

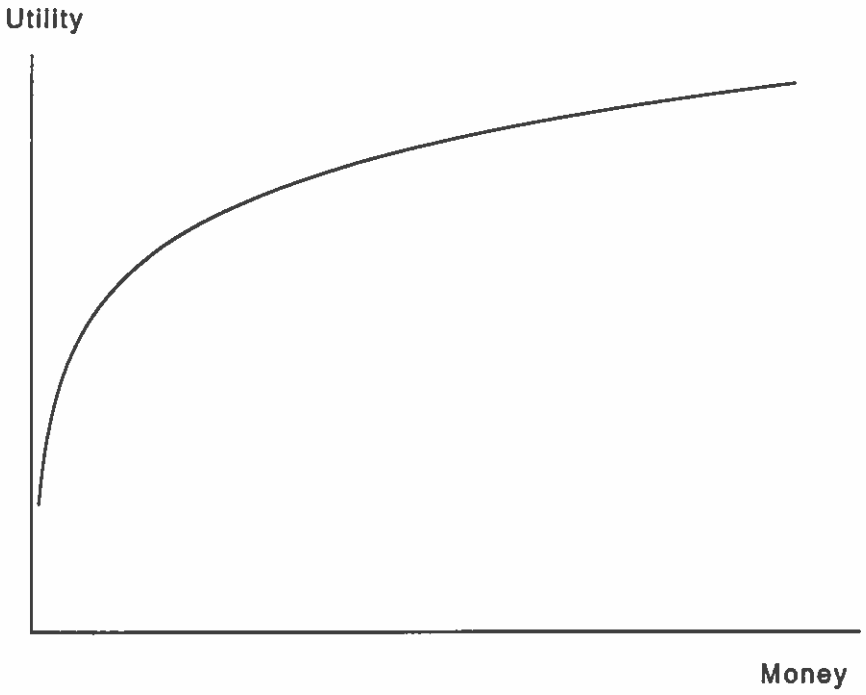

Figure 1.1. Bernoulli's utility function

walking is approximately the same for the rich and the poor, and so it must exceed the marginal utility of cab fare for the rich but not the poor. It is easy to appreciate how, in the hands of James Mill, Jeremy Bentham, and John Stuart Mill, utilitarianism could have generated a powerful movement for radical social reform.

> *Example: Justice as Fairness Meets Utilitarianism.* You believe in the version of justice as fairness that takes maximization of expected utility as the principle of individual rational decision. You have a friend, a utilitarian, who believes in maximizing total utility of the group. Each of you is to come up with a just plan to distribute one million dollars among 10 people of three different types: 3 of type 1, 4 of type 2, and 3 of type 3. You both believe that utility is an empirical psychological property and that the types in question are so specified as to determine the individual's utility function for money. You agree what the utility functions are for each type.[6] Then you and your friend will agree on the correct distribution scheme. You maximize your expected utility under the fictitious supposition that you have an equal chance of being any one of the ten, that is, that you have 0.3 probability of being of type 1, and so on. Then the quantity that your distribution scheme must maximize is just one tenth of the quantity that your friend's scheme is maximizing.

By the end of the nineteenth century, the sort of conception of utility that we find in the English utilitarians came under positivistic attack, notably by Vilfredo Pareto. If utility was to be interpreted in terms of consumers' behavior, rather than in the manner of introspective psychology, then interpersonal comparisons of utility seemed to make no sense. Indeed, even with respect to one consumer the measurement of utility on a numerical scale appeared to be unjustified. Consumers simply made their choices, and the operational part of utility talk seemed to consist solely of preference ordering.

In his manual of political economy (1927) Pareto used the ordinal indifference curve approach developed by Francis Ysidro Edgeworth. But while Edgeworth thought that utility was measurable and that the indifference curves derived from utility functions, Pareto took them as primitive. The entire theory, he wrote, "rests on no more than a fact of experience, that is on the determination of the quantities of goods which constitute combinations between which the individual is indifferent." This acerbic footnote follows: "This cannot be understood by literary economists and metaphysicians. Nevertheless, they will want to inter-

fere by giving their opinions; and the reader with some knowledge of mathematics can amuse himself by perusing the foolish trash they will put out on the subject."

To some extent, Pareto's position was anticipated by William Stanley Jevons. In his *Theory of Political Economy* (1871) Jevons addressed these skeptical remarks to the question of interpersonal comparisons of utility: "The susceptibility (to pleasure) of one mind may, for all we know, be a thousand times greater than that of another. But, provided that the susceptibility was different in a like ratio in all directions, we should never be able to discover the difference. Every mind is inscrutable to every other mind."

Lionel Robbins' influential discussions in the thirties (1932, 1938) echoed Jevons'. In 1938, he noted how his faith in utilitarian welfare economics had been shaken:

> I am not clear how these doubts first suggested themselves; but I well remember how they were brought to a head by my reading somewhere—I think in the works of Sir Henry Maine—the story of how an Indian official had attempted to explain to a high cast Brahmin the sanctions of the Benthamite system. "But that," said the Brahmin, "cannot possibly be right. I am ten times as capable of happiness as that untouchable over there." I had no sympathy with the Brahmin. But I could not escape the conviction that, if I chose to regard men as equally capable of satisfaction and he to regard them as differing according to a hierarchical schedule, the difference between us was not one which could be resolved by the same methods of demonstration as were available in other fields of social judgement.

Lord Robbins needn't have gone to India for his illustration. Consider the following specimen of nineteenth-century chauvinism from Edgeworth's *Mathematical Psychics* (1881, pp. 77–78):

> But equality is not the whole of distributive justice . . . in the minds of many good men among the moderns and the wisest of the ancients, there appears a deeper sentiment in favor of aristocratic privilege—the privilege of man above brute, of civilized above savage, of birth, of talent, and of the male sex. This sentiment of right has a ground of utilitarianism in supposed differences of *capacity* . . .
>
> If we suppose that capacity for pleasure is an attribute of skill and talent . . . we may see a reason deeper than Economics may afford for the larger pay, though often more agreeable work, of the aristocracy of skill and talent. The aristocracy of sex is similarly grounded upon the supposed superior capacity of the man for happiness . . . upon the sentiment—
>
> Woman is the lesser man, and her passions unto mine
> Are as moonlight unto sunlight and as water unto wine.

Illustration: Pareto Optimality. If we agree with Pareto that utility has only a personal and ordinal significance, what is left of utilitarianism as a social philosophy? One can still express certain qualitative features that every utilitarian measure of social welfare would have. We say that a social state S2 *weakly Pareto dominates* a state S1 when some member of society prefers S2 to S1 and no member prefers S1 to S2. S2 *strongly Pareto dominates* S1 if and only if all members of society prefer S2 to S1. The judgment that a social state is preferable to one which it strongly Pareto dominates is an ordinally expressible remainder of the utilitarian doctrine that social utility is the sum of individual utilities. The stronger principle that a social state is preferable to one which it weakly Pareto dominates is motivated by the additional utilitarian principle that everyone counts equally. The ordinal remainder of this principle is, roughly, that everyone counts for something. We will say that a Pareto optimal social state is one which is not weakly Pareto dominated by any other.

The concept of utility was partly rescued from Paretian skepticism by Frank Ramsey in 1926[7] and by John von Neumann and Oskar Morgenstern in 1944. Ramsey's analysis went deeper than that of von Neumann and Morgenstern, but was for a long time little known among economists.[8] When von Neumann and Morgenstern independently rediscovered one of Ramsey's key ideas and published it in *Theory of Games and Economic Behavior*, they carried cardinal utility back into respectability.

The idea in question was to consider preferences not only for goods or prospects, but also for *gambles* over goods. If you know only that my preference order for desserts is

Raspberries and cream
Chocolate mousse
Blueberry pie
Cheesecake

you cannot sensibly answer the question as to whether the *difference* in utility between the first two items is equal to the difference in utility between the second two. But if you are, in addition, told that I am indifferent between a gamble which gives me raspberries and cream (RC) if heads; cheesecake (Ch) if tails, and one which gives me chocolate mousse (CM) if heads; blueberry pie (BP) if tails, then I can conclude that these differences are equal:

$$\text{Utility Gamble 1} = \text{Utility Gamble 2}$$

$$\tfrac{1}{2}U(\text{RC}) + \tfrac{1}{2}U(\text{Ch}) = \tfrac{1}{2}U(\text{CM}) + \tfrac{1}{2}U(\text{BP})$$

$$U(\text{RC}) - U(\text{CM}) = U(\text{BP}) - U(\text{Ch})$$

Pareto had already remarked that if a natural criterion for equality of differences of utility could be found, then utility could be measured on a numerical scale. Given a conventional choice of the zero point and the size of a unit "utile," the utility values would be unique.[9]

The extension of the preference ordering to gambles provided the requisite criterion for quantifying utility. Cardinal utility was, after all, justified in a way strictly in accord with the Paretian methodology. Given von Neumann–Morgenstern utility, Bernoulli's idea about the declining marginal utility of money again makes perfect sense. It is an empirical claim about utility differences for a given individual, and it reduces to a claim about that individual's preferences over gambles. The interpersonal comparisons of utility assumed by the classical English utilitarians, however, are not rescued by the von Neumann–Morgenstern construction. Since the data do not determine the choice of zero point and unit, quantities such as (1) the sum of utilities over different persons and (2) the utility of that member of the group who is worst off in terms of utility are not given any determinate sense.

> *Example: Personal Justice.* Ten followers of von Neumann and Morgenstern are deciding how to distribute one million dollars among them. Each is an expected utility maximizer. Each attempts to find a just distribution scheme among numbers 1 to 10, where players will be assigned numbers later by fair lottery. Each seeks a distribution scheme that maximizes expected utility according to his or her *own* utility function. So each player has a personal conception of justice, and no interpersonal comparison of utilities is involved. Nevertheless, if each player's utility function exhibits a declining marginal utility for money then all players will agree that the egalitarian distribution scheme is the fair one.[10]

Von Neumann and Morgenstern quantified utility by bringing in known chances. But at least some thinkers with empiricist worries about utility also have empiricist worries about chance. This wider skepticism raises questions about the availability of probability to quantify utility. Ramsey had already answered this question in 1926 by constructing both personal utility and personal probability out of preferences over gambles.

The key move is to find subjective surrogates for the *chances* that von Neumann and Morgenstern used to scale the decisionmaker's utilities. Ramsey started by identifying propositions that have no values in and of themselves to the decisionmaker, and whose truth or falsity does not modify the value of payoffs. He calls these propositions "ethically neutral." A proposition, p, is ethically neutral with respect to a collection of payoffs, B, for an agent if she is indifferent between B with p true and B with p false. A proposition, p, is ethically neutral for an agent if it is ethically neutral for her with respect to maximal collections of payoffs.[11] The nice thing about ethically neutral propositions is that the expected utility of gambles on them depends only on their probability and on the utility of the contemplated payoffs. The utility of the ethically neutral propositions themselves is not a complicating factor. In our illustration of the von Neumann–Morgenstern utility scale, I tacitly assumed that the propositions describing the outcome of the coin flip (H or T) were ethically neutral.

We can identify an ethically neutral proposition, H, for which the decisionmaker has personal probability (degree of belief) ½ when there are two payoffs, A and B, such that she prefers A to B but is indifferent between the two gambles: (1) Get A if H is true, B if H is false; (2) get B if H is true, A if H is false. For the purpose of scaling the decisionmaker's utilities, such a proposition is just as good as the proposition that a fair coin comes up heads.

The procedure just described extends in a straightforward way to identify more surrogates for chance. Consider a wheel of fortune with 100 possible outcomes which are ethically neutral. The decisionmaker prefers A to B but is indifferent between (1) A if outcome i, B otherwise, and (2) A if outcome j, B otherwise, for all possible outcomes i and j. Then the decisionmaker regards the 100 possible outcomes as equiprobable, and the disjunction of N outcomes as having probability of $N/100$. For a richer assortment of surrogate chances consider a wheel of fortune with more sides, or consider sequences of outcomes for repeated flips of a coin. A rich enough preference ordering has enough ethically neutral propositions to approximate the external scaling probabilities used by von Neumann and Morgenstern to any desired degree of precision.

These are the key ideas of the procedure by which Ramsey extracted from a rich and coherent preference ordering over gambles both a subjective utility and a subjective degree of belief such that the preference ordering agrees with the ordering by magnitude of expected utility. Utility now has shed the psychological and moral associations with which it was associated in the eighteenth and nineteenth centuries. The theory of expected utility is now a part of *logic:* the logic of coherent preference.

Ramsey proved a *representation theorem* to the effect that any coherent preference ordering over a rich enough set of gambles has associated with it a unique probability and a utility unique up to choice of zero point and magnitude of the unit utile, such that the ordering by expected utility agrees with the preference ordering. If we hold that rational preferences over a meager set of gambles (a) should be coherent and (b) should be embeddable in a coherent set of preferences over an enriched set of gambles, we keep the existence of a probability-utility representation whereby rational preferences over the meager set of gambles goes by expected utility, although we lose uniqueness. In this sense, we can say that the only normative content implied by the use of an expected utility model is that preferences should be coherent.[12]

If one has some sort of coherent moral preferences for society, these preferences must as well admit of an expected utility representation. So coherent social preferences over a rich set of prospects give rise to a corresponding social utility function unique up to choice of a conventional zero point and unit of measurement. When does the social utility have a utilitarian representation? That is, when is it the case that there are choices of zero points and units of measurement for individual utility scales and for the social utility scale, such that social utility is the sum of individual utilities? John Harsanyi answered this question in 1955 by proving the appropriate representation theorem. Neglecting technical details, the theorem says that coherent social preferences which satisfy the Pareto condition[13] with respect to individual preferences have a utilitarian representation. So, it seems that the moral content of modern utilitarianism consists of nothing more and nothing less than the Pareto condition.[14]

Rawls (1971) identified two great traditions in Western ethics: the social contract tradition, whose essential leading idea is justice as fairness, and the utilitarian tradition. When seen in the light of modern utility theory, there is considerable convergence between these traditions. As Harsanyi (1955) pointed out, under certain conditions justice as fairness *entails* utilitarianism. Suppose (1) you derive your social utility ordering by applying the expected utility version of justice as fairness to a society with categories and (2) your preferences conditional on being in a category coincide with the preferences of anyone in that category.[15] Then you are a utilitarian.

It is evident that the whole question of utilitarianism has been profoundly transformed by the evolution of the utility concept. Contemporary philosophy has been a little slow in catching up; a considerable amount of contemporary discussion of the ethics of utilitarianism remains stuck in the nineteenth century. We cannot pursue these impli-

cations of the modern utility concept for social philosophy here. The interested reader must be content with a reference to the literature.[16] Returning to our central topic of maximization of expected utility in individual decision, I would like to end this section with a question. *If expected utility theory presupposes only coherent preferences, how could it fail to be applicable in game-theoretic situations?*

The Theory of Games

The theory of games is almost entirely a creation of the twentieth century. Game-theoretic problems were considered in the 1920s by Borel (1924) and von Neumann (1928), and the subject emerged full-blown in von Neumann and Morgenstern's *Theory of Games and Economic Behavior* in 1944.

Von Neumann and Morgenstern argue that there is a difference in principle between rational decisionmaking for a single individual, such as Robinson Crusoe, acting against nature, and for a member of a group of interacting rational individuals (1947, pp. 11–12):

> The difference between Crusoe's perspective and that of a participant in a social economy can also be illustrated in this way: Apart from those variables which his will controls, Crusoe is given a number of data which are "dead"; they are the unalterable physical background of the situation . . . Not a single datum with which he has to deal reflects another person's will or intention of an economic kind—based on motives of the same nature as his own. A participant in a social exchange economy, on the other hand, faces data of this last type as well: they are the product of other participants' actions and volitions . . . His actions will be influenced by his expectation of these, and they in turn reflect the other participants' expectation of his actions.

The difficulty that is here supposed to arise in the case of the social exchange economy is not so much the "free will" of the other participants as the role of mutual expectations, expectations of expectations, and so forth that can exist in a community of utility maximizers. These may seem to threaten a kind of self-reference in which the mutually interacting optimization problems of the various actors are not capable of joint solution. This is really the problem to which the theory of games is addressed. It will become clearer when we discuss von Neumann and Morgenstern's justification of their concept of a solution for a game.

First, however, let me introduce a few of the concepts of the theory. In the simplest sort of situation, a *normal-form* game, the players all choose simultaneously and independently among their respective possible acts, and the payoffs are determined by the combination of acts

chosen. A game is called *zero-sum* if (for some scaling of the players' utilities) the total payoff for every possible combination of acts is zero.

The simplest zero-sum games are those involving only two players, and it is here that the von Neumann–Morgenstern theory has its greatest success. Such a game can be specified by a payoff matrix for one player, since the second player's preferences can be represented by payoffs which are just the negative of those of the first player. Here is an example:

Row's payoff matrix

	C1	C2	C3
R3	1	−1	−1
R2	1	0	1
R1	−1	−1	1

If Row does his act 3 and Column does her act 1, then Row gets payoff 1 and Column therefore gets payoff −1; whereas if they both do their respective act 2, they both get payoff 0. A simultaneous choice of acts by all players is called a *Nash equilibrium* if no player can improve his or her payoff by a unilateral defection to a different act. In other words, at a Nash equilibrium, *each player maximizes his or her utility conditional on the other player's act*. For example, [R1, C1] is not a Nash equilibrium, because if Column plays 1 then Row is better off playing either 2 or 3. Likewise, [R3, C1] is not a Nash equilibrium because if Row plays 3, Column would do better playing 2 or 3. (Remember that Row's losses are Column's gains.) You can verify that [R2, C2] is the unique Nash equilibrium of the game. If Column plays 2 Row can do no better than to play 2 himself (indeed in this case the alternatives are strictly worse) and Column is in a similar situation when Row plays 2. In a famous passage, von Neumann and Morgenstern (1947, p. 148) argue for the centrality of the Nash equilibrium concept:

> Let us now imagine that there exists a complete theory of the zero-sum two-person game which tells a player what to do, and which is absolutely convincing. If the players knew such a theory then each player would have to assume that his strategy has been "found out" by his opponent. The opponent knows the theory, and he knows that the player would be unwise not to follow it . . . a satisfactory theory can exist only if we are able to harmonize the two extremes . . . strategies of player 1 "found out" or of player 2 "found out."

The harmony in question is Nash equilibrium, and the argument is supposed to show that an adequate theory of rational behavior in game-theoretic situations will have the consequences that if each player makes a rational choice, the result will be a Nash equilibrium. Then, if there is such an adequate theory of rational behavior, in any game like our example in which a unique Nash equilibrium exists the rational choice for each player must be to choose the act that is a constituent of the unique Nash equilibrium. In this way, if we have existence and uniqueness of Nash equilibria in general, we can turn the Nash equilibrium concept itself into a theory of rational choice.

Let us pause to consider this argument in the light of subjective expected utility theory. It is far from airtight. Here I want to emphasize one rather large and important gap. Suppose that there is a theory of rationality which is absolutely convincing, and suppose not only that both players know it but that it is common knowledge. (Each knows that the other knows it and that the other knows that he knows it, and so forth.)[17] Suppose that the calculation needed to extract the relevant information is finite,and that it is moreover small enough to take virtually no effort on the part of the players. Does it follow that each player can find out the other player's strategy? Only if the *inputs* to the theory are themselves common knowledge among the players. Von Neumann and Morgenstern are assuming that the *payoff matrix* is common knowledge to the players, but presumably the players' subjective probabilities might be private. Then each player might quite reasonably act to maximize subjective expected utility, believing that he will *not* be found out, with the result *not* being a Nash equilibrium.

For a rather crude illustration of this possibility, suppose in the foregoing example that Row plays 3 because he is almost sure that Column will play 1, thinking that Column is almost sure that he will play 1 in response to the mistaken belief that Column will play 3, etc. And suppose that Column plays 3 because she is almost sure that Row will play 3, because she thinks that Row thinks that she will play 1, etc. Each maximizes expected utility with the result being [R3, C3], which is not an equilibrium. For simplicity, this example uses degrees of belief that are nearly zero or one, but more complex and interesting examples of the phenomenon are possible. It has been recently studied by Bernheim (1984) and Pearce (1984), who call it *rationalizable strategic behavior*.

In two-person games, there is a relatively simple way to identify the rationalizable acts. An act is *strongly dominated* for a player if she has some other act that gives a better payoff no matter what the other player does. For example, in the zero-sum game for which Row's payoff matrix

is as follows, Row's act 2 *strongly dominates* Row's act 1, but Column has no strongly dominated acts.[18]

	C1	C2
R2	1	2
R1	0	0

It is clear that no strongly dominated act is rationalizable, since no beliefs about the other player's acts can make it look as attractive as the act which dominates it. Conversely, Pearce (1984) and Bernheim (1984) showed that in two-person games those acts which remain after iterated deletion of strongly dominated strategies are all rationalizable. This principle extends to n-person games, if we do not require that one player's beliefs make other players' actions probabilistically independent.[19] Thus, in a game like our first example, where no strategies are strongly dominated, all strategies are rationalizable.

Von Neumann and Morgenstern neglected the possibility of rationalizable nonequilibrium strategies, even though they developed a theory of personal utility, because they never took the extra step with Ramsey to subjective probability. They assume that the information in the payoff matrix is the complete input for a theory of rational decision—an attitude that is perhaps still the prevalent one in game theory. I can find no principled argument for this assumption, and the theory of personal probability as developed by Ramsey, de Finetti, and Savage appears to contradict it.

In order for the von Neumann–Morgenstern argument for Nash equilibrium to work, all the inputs for rational decision must be common knowledge; otherwise the hypothesis that the other player can find out your strategy and make a best reply lacks a foundation. Therefore let us suppose that each player's prior personal probabilities are common knowledge, as is both the payoff matrix and the fact that the players are expected utility maximizers. Does the argument for Nash equilibrium now succeed? To see that it does not, let us modify our example as follows:

	C1	C2	C3
R3	1	0	−1
R2	0	0	0
R1	−1	0	1

[R2, C2] is still the *unique* Nash equilibrium. Now suppose that each player's prior probabilities over the other player's actions are [⅓, ⅓, ⅓] and that this is common knowledge, as is the payoff matrix and the fact that the players are subjective expected utility maximizers. These assumptions are compatible with any play by any player (with no prospect of being found out) and thus with any outcome of the game. Here uniqueness of the Nash equilibrium does not entail uniqueness in the recommendations of the underlying theory of individual rational behavior, so even common knowledge of the prior probabilities does not guarantee that a player's choice of action will be found out. It is apparent that we must build in even stronger assumptions about common knowledge and tell a more complicated story in order to make the argument for Nash equilibria valid. I will return to this question in the next chapter. For the moment, let us take the Nash equilibrium concept with a grain of salt, and proceed.

In the case of two-person zero-sum games, von Neumann and Morgenstern were able to establish a deep connection between security (maximin gain) strategies on the parts of the individual players and Nash equilibria of the game. Suppose that you and I are playing such a game, and our strategies form a Nash equilibrium. Then by the definition of *equilibrium*, neither of us would profit by a unilateral change of strategy. Since the game is zero-sum, my profit is your loss and conversely. So neither of us can lose by the other's unilateral change of strategy. In other words, we are both playing security strategies. So, in this special case, it is a necessary condition for a Nash equilibrium that each player play his or her security strategy.

Is it a sufficient condition? The *prima facie* answer is *no*. Consider the game of Matching Pennies. You either hide a penny in your hand, or not. I guess whether you did. If I guess correctly, I give you a penny, otherwise you give me a penny. Inspection of the payoff matrix will disclose that there *is* no Nash equilibrium:

Matching Pennies

	C1	C2
R2	1	−1
R1	−1	1

At each combination of acts, unilateral deviation will pay one of the players. Each act is a security strategy for each player.

The picture changes radically, however, if one follows Borel (1924) in allowing *mixed strategies*. Each player turns his or her choice of an act over to a chance device, with the operative choice being the choice of the chances. (The chance devices of different players operate independently.) So, in Matching Pennies, Row can choose any number between zero and one as the chance of R2; likewise for Column. Any point in the unit square thus represents mixed strategies for Row and Column. The payoffs for a combination of mixed strategies are the expected utilities, using the chance probabilities to compute the expectation. Figure 1.2 shows Row's payoffs for mixed strategies in Matching Pennies, plotted as a surface above the unit square. Enlarging the set of objects of choice to include mixed strategies has created an equilibrium: the combination of mixed strategies in which each player gives equal chances to each of her alternatives. This consists of the *security* mixed strategy for each player, where each player has an expected payoff of zero. Von Neumann proved in 1928 that it is true in general for finite two-person zero-sum games that if mixed strategies are included, there is always a Nash equilibrium and it will be attained if both players play security strategies.

So we always have *existence* of an equilibrium. But it may fail to be *unique*. There may be many Nash equilibria in a finite two-person zero-sum game with mixed strategies—but the failure of uniqueness is relatively painless. This is because the equilibria are *interchangeable*, in the

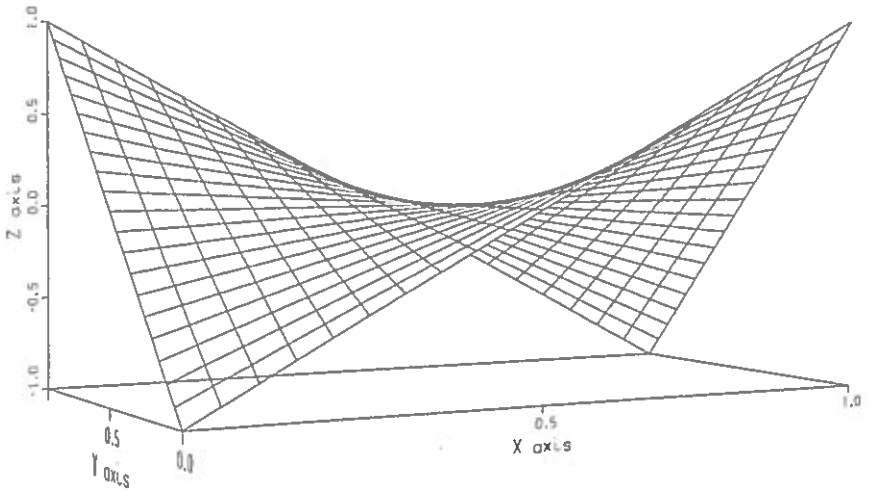

Figure 1.2. Payoff surface for mixed strategies in Matching Pennies

following sense: (i) if I play my end of one equilibrium, and you play your end of another, then the combination of our plays is a third equilibrium, because any combination of security strategies is an equilibrium and (ii) all equilibria have the same value for each player, namely that player's security level.

If one is willing to buy the rather shaky argument advanced by von Neumann and Morgenstern for the equilibrium concept, the problem of rationality has been solved for this special class of games. Rational action consists here in choosing a security strategy. That all players do so is a necessary and sufficient condition for their combined play to be at a Nash equilibrium (that is just as good as any other equilibrium). It is on this basis that the idea of a security strategy was rather widely applied in the heady decade following the publication of *Theory of Games and Economic Behavior*. Everyone knew, of course, that the tight connection between security and equilibrium did not hold for games in general, but the significance of that fact was not always fully appreciated.

When we pass from finite two-person, zero-sum games to finite *n*-person non-zero-sum games, the connection between security and equilibrium is broken and the picture becomes immensely more complicated. One remarkable fact is that here mixed strategies still guarantee the existence of equilibria. This fact was demonstrated by John Nash (1951), after whom they are named.

We now assume that there are a finite number of players, each with a finite number of strategies. Each player's payoff is a function of the choices of strategy of all players. There need be no special relationship between payoffs of different players. Two-person zero-sum games model situations of strict competition, but this larger class of games can model a whole spectrum of payoff profiles from strict competition to pure cooperation.

As an illustration of the difficulties introduced by even the simplest non-zero-sum games, consider the game of Chicken:

Chicken

		Column	
		Don't swerve	Swerve
Row	Don't swerve	−10, −10	5, −5
	Swerve	−5,5	0,0

The name comes from the image of two teenagers driving down the center of the road on a straight course toward a head-on collision; the first one who swerves is "chicken." The ordered pairs in the payoff matrix list the payoff for Row first and the payoff for Column second. If one swerves and the other doesn't the first loses face and the second gains it. If both swerve their relative status is unchanged. If neither swerves the result is worse than losing face. The story is slightly misleading here since, as before, we want to think of the players as making independent decisions at a given time. So let's suppose that the drivers must make an irreversible decision at the start by pushing a button in their computerized hot rods and then ride it out.

There are two pure Nash equilibria in this game: [Row swerves, Column doesn't] and [Column swerves, Row doesn't], as well as an equilibrium in mixed strategies where each flips a fair coin to decide if she will swerve.[20] We do not have interchangeability. If each picks her end of that Nash equilibrium that she prefers, neither will swerve, since each prefers the pure equilibrium in which the other swerves. This would lead to a definitely nonequilibrium outcome. And the connection with security strategy is broken because each player's security strategy is not to swerve—so the result of both players going for security is again nonequilibrium.

Although uniqueness fails rather dramatically, we do still have existence of equilibria. Nash proved this by exhibiting a continuous function[21] which maps the space of mixed strategies (of all players) into itself for a non-zero-sum game. This function leads each player to put more weight on strategies which look better than the *status quo*. By a well-known theorem of Brower, this function has a fixed point: a mixed strategy that gets mapped onto itself. At such a fixed point, no pure strategy looks better to any player than the mixed strategy associated with that point. This point is therefore a Nash equilibrium.

Nash also investigated special classes of non-zero-sum games where interchangeability of equilibrium points does hold. It is evident, however, that in the general case the status of the equilibrium concept as a touchstone of rationality is here even shakier than in the case of zero-sum game theory. In games with multiple equilibria, even if your opponent knows that you will pick your end of an equilibrium, he cannot figure out your strategy. And if he cannot figure out your strategy and the equilibria are not interchangeable, what is your rationale for even picking an equilibrium in the first place?

Since the pressing problem is too many rather than too few equilibria,

there has been reason to look for stronger solution concepts that are met by only some subset of the Nash equilibria. And, indeed, some Nash equilibria seem to look better than others. Consider the following game,[22] supposing you are Row:

	C1	C2
R2	0,0	0,0
R1	1,1	0,0

In this game there are two Nash equilibria, one at [R1, C1] and one at [R2, C2]. Each is a genuine equilibrium; if Column plays her end of the equilibrium you can do no better than to play yours. But if you have even the slightest doubt that she will play C2, R1 will have greater expected utility.[23] And if you reason this way, isn't it likely that Column—being in a symmetric situation—will do so as well?

The two equilibria in this game are distinguished by Selten's (1975) notion of a *perfect* equilibrium. A *completely mixed* strategy is a mixed strategy in which every pure strategy gets some positive probability. If your opponents play mixed strategies, one of your strategies is a *best reply* if it maximizes expected utility where the expected utility is calculated using your opponents' mixing probabilities. An *ε-perfect equilibrium* is a completely mixed strategy (for all players) in which any pure strategy which is not a best reply has weight less than ε. A *perfect equilibrium* is a limit as ε approaches 0 of ε-perfect equilibria. Every perfect equilibrium is a Nash equilibrium, but the converse is not true. In the example, [R1, C1] is a perfect equilibrium, but [R2, C2] is not since for any completely mixed strategy, R1 is Row's unique best reply and C1 is Column's unique best reply.

Selten (1975, p. 35) views the model as embodying "A Model of Slight Mistakes": "There cannot be any mistakes if the players are absolutely rational. Nevertheless, a satisfactory interpretation of equilibrium points in extensive games seems to require that the possibility of mistakes is not completely excluded. This can be achieved by a point of view which looks at complete rationality as a limiting case of incomplete rationality." Myerson (1978, p. 74) commented: "The essential idea behind Selten's perfect equilibria is that no strategy should ever be given zero probability, since there is always a small chance that any strategy might be chosen, if only by mistake." For this reason, Selten's concept is often called "trembling hand" perfection.

Myerson pushed the idea one step further. The previous example can be converted to the next by giving each player a third unattractive alternative:

	C1	C2	C3
R3	−9, −9	−7, −7	−7, −7
R2	0,0	0,0	−7, −7
R1	1,1	0,0	−9, −9

Now there are three Nash equilibria: [R1, C1], [R2, C2], [R3, C3]. There are two perfect equilibria for now not only [R1, C1] but also [R2, C2] is a perfect equilibrium. (To see this consider a path of convergence along which R1, R3, C1, C3 are equiprobable and grow smaller while R2 and C2 are equiprobable and converge to 1.) Yet [R1, C1] still may seem the "best" equilibrium for much the same reason as before. Is one inclined to think that a "tremble" to R3 or to C3 is not really as likely as one to R2 or C2? To capture this intuition, Myerson introduced the notion of a *proper* equilibrium.

An ε-*proper equilibrium* is a completely mixed strategy such that if a pure strategy, A1, is a better response than a pure strategy, A2, then the probability ratio $p(A2)/p(A1)$ is less than ε. A *proper equilibrium* is a limit of a sequence of ε-proper equilibria as ε approaches zero. Every proper equilibrium is a perfect equilibrium but not conversely. In the example, [R1, C1] is proper; [R2, C2] is perfect but not proper; [R3, C3] is a Nash equilibrium but not a perfect one.

Selten and Myerson proved that perfect and indeed proper equilibria *exist* in every game of the kind under consideration. But the problem of nonuniqueness persists. For example, in the game of Chicken, all equilibria are perfect and proper. Considering the symmetry of games like Chicken, it seems unrealistic to expect a stronger equilibrium concept to deliver both existence and uniqueness.[24] Therefore it seems that the situation for non-zero-sum game theory can never be as simple and elegant as in the von Neumann–Morgenstern theory for zero-sum games.

There is a further complication in game theory which I have postponed discussing until now. That is the status of games in *extensive form*. Here we generalize from the simple model where all players simultaneously and independently make their choices to a model which allows a sequence of moves by different players in varying states of information

about the moves already made by other players. Von Neumann and Morgenstern have an argument to the effect that all the complexity of the extensive-form model can, in the end, be reduced to the simple normal-form model of one round of simultaneous independent choices. But, as we shall see, this argument has been seriously challenged.

Following Kuhn (1953), we will model the temporal, causal, and informational structure of a game in extensive form as a tree. A simple example is shown in Figure 1.3. Player A moves first, choosing either act A1 or A2. Then player B either finds herself at the top node, in which case she knows that A has played A1, or at the bottom node, in which case she knows that A has played A2. In either case, she must choose between B1 and B2. Then the game is over; the course of play has traversed a path through the tree; and the payoffs at the end of that path are received.

The foregoing is a game of *perfect information*. Each player at each choice point knows whatever preceding choices have been made. The Kuhn model also allows for games where this may not be the case. Consider the tree in Figure 1.4. Here player A begins by choosing A1, A2, or A3. Then B must move. If A chose A1, B knows it. But otherwise, B knows only that A chose either A2 or A3. This is indicated by the dotted line between A2 and A3. The set containing the nodes which these two

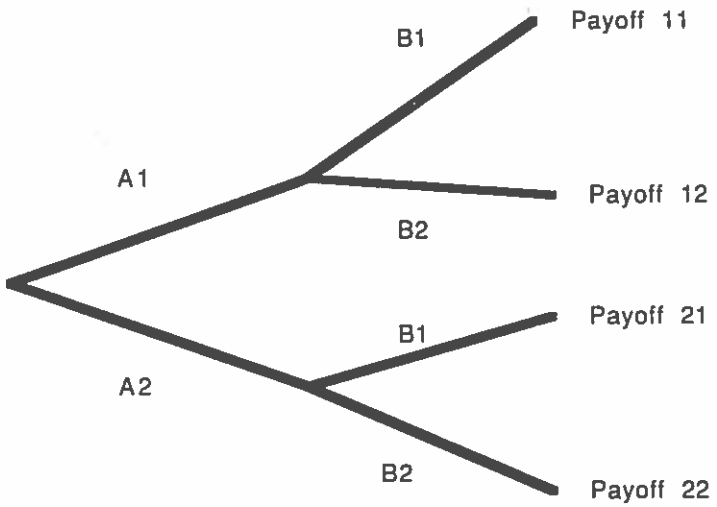

Figure 1.3. Kuhn tree for a game of perfect information

acts lead to is known as B's *information set* at each of these nodes. A player's knowledge of preceding play by other players may thus be limited by his information set. We do, however, assume perfect recall: a player remembers whatever he knew earlier in the course of play.

A player's *strategy* for an extensive-form game is a comprehensive contingency plan: a function that maps each information set at which he could find himself into a choice of action. If each player thought about how to play the game and independently chose such a strategy, these strategies would jointly determine the course of play. For this reason, von Neumann and Morgenstern argued that a game in extensive form was equivalent to the normal-form game where players choose between its strategies: the *strategic normal-form game* for the original extensive-form game.

The adequacy of the strategic normal-form representation of extensive-form games was generally taken for granted until questioned by

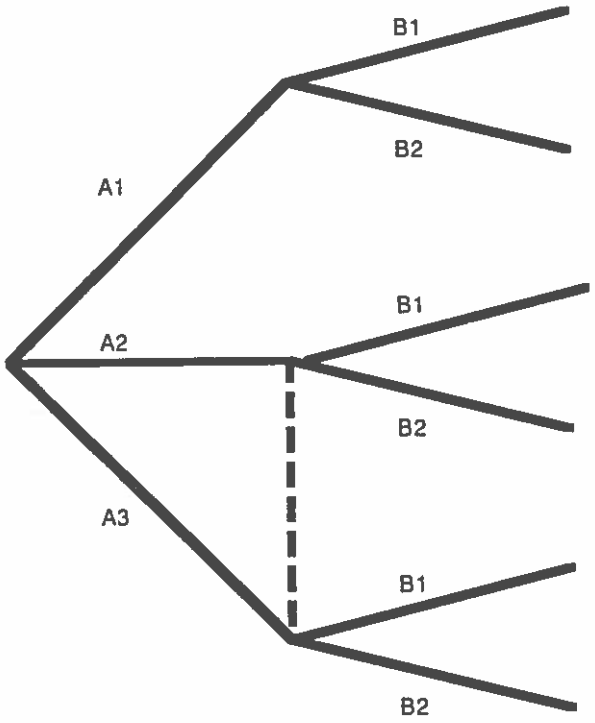

Figure 1.4. Extensive-form game with information set

Selten (1965).[25] Consider the game tree in Figure 1.5, which is the strategic normal form of the game:

	B1 if A2	B2 if A2
A1	0,0	0,0
A2	1,1	−1,−1

Since A moves first, he has only two strategies. B has two strategies, depending on what she plans to do if A does A2. If we look at the matrix of the normal form, we find two Nash equilibria: [A2, B1 if A2] and [A1, B2 if A2]. But if we look back at the game tree, we see that the second equilibrium is quite wacky. B would be foolish to choose B2 if A had chosen A2, since choosing B1 would surely give her a greater payoff. (Here B's maximizing expected utility does not depend on B's particular subjective probabilities, because B knows with certainty at the choice point at issue that B1 will give a greater payoff.) Thus the strategy B2 if A2 is not a credible option for B. Seeing this, A will choose A2 and B will choose B1. The point becomes more vivid in the setting of questions of nuclear deterrence. Hermann Kahn (1984, p. 59) reports a typical beginning to a discussion of the policy of mutually assured destruction (MAD):

> One Gedanken experiment that I have used many times and in many variations over the last twenty-five or thirty years begins with the statement: "Let us assume that the president of the United States has just been informed that a

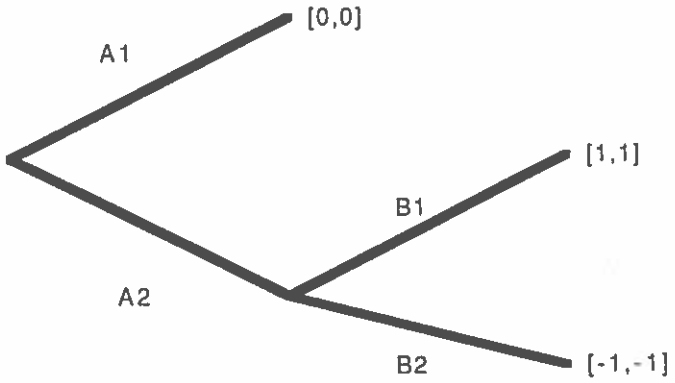

Figure 1.5. A challenge to strategic normal form

multimegaton bomb has been dropped on New York City. What do you think that he would do?" When this was first asked in the mid-1950s, the usual answer was, "Press every button for launching nuclear forces and go home." . . . the dialogue between the audience and myself continued more or less as follows . . .

Kahn: "What happens next?"
Audience: "The Soviets do the same!"
Kahn: "And then what happens?"
Audience: "Nothing. Both sides have been destroyed."
Kahn: "Why then did the American President do this?"

A general rethinking of the issue would follow, and the audience would conclude that perhaps the president should not launch an immediate all-out retaliatory attack.

What the audience is beginning to see, in our terms, is that there is an essential difference between the strategic normal-form representation of MAD, which treats it as equivalent to Dr. Strangelove's doomsday machine, and the extensive-form representation, which pays attention to the causal and informational context of the decisions involved in implementing the policy. Even in a situation in which a doomsday machine would be an effective deterrent a policy of mutually assured destruction would not be, because it rests on a noncredible threat. The threat is not credible because in the relevant situation it would not be in the best interests of the nation to carry it out.[26]

It is evident that the strategic normal form of an extensive-form game may fail to capture important causal and informational structure, and consequently that the Nash equilibrium concept applied to strategic normal form may be inadequate. In an important paper, Kreps and Wilson (1982b) proposed to remedy the situation by an application of expected utility theory. A strategy (for all players) is *sequentially rational* if the strategy of each player, starting at each information set, maximizes expected utility according to her beliefs and the strategies of all the other players.

But what of information sets to which players initially assign probability zero? What should a player's beliefs be conditional on reaching such an information set? Kreps and Wilson put a "consistency" condition on these conditional probabilities. They must be obtainable as a limit of well-defined conditional probabilities in a sequence of assessments (degree of belief, strategy pairs) which give each information set nonzero probability. A *sequential* equilibrium is then defined as an assessment which is both consistent and sequentially rational. The fishy equilibrium in the example of Figure 1.5 cannot be sequential.

The attentive reader has perhaps noticed that the fishy equilibrium is

also not *perfect* in the sense of Selten, and indeed his concept of perfect equilibrium was introduced with these problems of extensive-form games in mind. Selten's notion is somewhat stronger than that of Kreps and Wilson. It can be thought of as adding to the requirement of sequential rationality an additional requirement of a certain kind of robustness under "trembles." Stronger kinds of robustness conditions have been suggested—for example, the "persistent" equilibria of Kalai and Samet (1984) and the various types of structural stability investigated by Kohlberg and Mertens (1980). For most of these refinements of the Nash equilibrium concept, it can be shown that at least one such refined equilibrium exists in every finite non-zero-sum game. None of them, however, is strong enough to guarantee uniqueness, and thus none solves the problem of multiple equilibria in non-zero-sum games.

Integrating the Two Paradigms

The theory of subjective probability and utility provides a foundation for a univocal rationality principle: *maximize expected utility*. The analysis initiated by Ramsey shows that the normative content of this theory is just that preferences should be *coherent*. There is nothing in the foundations of expected utility theory to limit its applicability in the sort of situations studied by the theory of games.

In the theory of noncooperative games, rational action is discussed in terms of a cluster of equilibrium concepts. The central notion is that of a *Nash equilibrium*, and the concept of a *security strategy* derives its license from its connection with Nash equilibria in the special case of two-person zero-sum games. The concept of a Nash equilibrium rests on that of expected utility. A Nash equilibrium is just a combination of strategies for each player such that if each player has found out the other players' strategies each is maximizing expected utility.

But the rationale for assuming that each player will have found out the other players' strategies is murky. Von Neumann and Morgenstern's argument for this assumption appears to fail even in the case of two-person zero-sum games. And it is in worse trouble in the non-zero-sum case as a result of multiple, noninterchangeable equilibria. Doubts about the von Neumann–Morgenstern argument are, in a way, behind both proposals to weaken (to rationalizability) and to strengthen (to perfect or proper equilibrium) the concept of Nash equilibrium. In games in extensive form the expected utility principle applied at the players' choice points comes in conflict with the Nash equilibrium principle applied to overall strategies. The disparity between these two principles

here provides a further motivation for refinements of the Nash equilibrium concept.

All this suggests that the picture of expected utility theory and game theory as separate theories dealing with separate domains is wildly inaccurate. Instead, game theory is and should be founded on expected utility theory, but the details of the foundation are open to serious question. If anything like classical game theory is to emerge, it must be under stronger assumptions than just common knowledge of Bayesian rationality (that is, expected utility maximization). The strength of assumptions needed to derive classical game theory, and the effect on game theory of weakening those assumptions, are subjects which merit investigation.