The Stanford Institute for Mathematical Stu
in the Social Sciences

APPLIED MATHEMATICS AND STATISTICS LABORATORIES
STANFORD UNIVERSITY

Reprint No. 42

The Philosophical Relevance of Decision Theory

PATRICK SUPPES

Reprinted from
The Journal of Philosophy
Vol. LVIII, No. 21, Oct. 12, 1961

Reprinted from THE JOURNAL OF PHILOSOPHY, Vol. LVIII,
No. 21, October 12, 1961
Printed in U. S. A.

THE PHILOSOPHICAL RELEVANCE OF DECISION THEORY *

1. *Introduction.* There is, I am sure, a sense in which any developed scientific theory has philosophical significance. It is equally clear that some scientific theories are of considerably more philosophical importance than others. For philosophy, quantum mechanics is more important than hydrodynamics, learning theory than social psychology, the theory of sets than topology, and so on. It is the primary point of the present paper to discuss the philosophical relevance or importance of decision theory, a theory I classify as a new branch of mathematical statistics and economics, with certain ramifications in psychology. I hope to be able to show that in its own way decision theory has the kind of primary relevance for philosophy that we associate with quantum mechanics or the theory of sets.

To begin with, we may characterize the fundamental problem of decision theory in the following way. A person, or group of persons, is faced with several alternative courses of action. In most cases the decision maker will have only incomplete information about the true state of affairs and the consequences of each possible act. The problem is to choose an act that is optimal relative to the information available and according to some definite criteria of optimality.

In a very natural way, the most important branches of decision theory may be characterized by a 2×2 table as illustrated below. The left column is for the category of individual decisions and the right column for the category of group decisions. The first row is for the category of normative theory and the second for the category of descriptive theory.

* This paper has been written in connection with research supported by the Group Psychology Branch of the Office of Naval Research under Contract NR 171-034. To be presented in a symposium on "Decision Theory and Its Philosophical Significance" at the fifty-eighth annual meeting of the American Philosophical Association, Eastern Division, December 28, 1961.

	Individual decisions	*Group decisions*
Normative theory	Classical economics Statistical decision theory Moral philosophy	Game theory Welfare economics Political theory
Descriptive theory	Experimental decision studies Learning theory Survey studies of voting behavior	Social psychology Political science

Rather than comment on the philosophical relevance of decision theory in general, I shall attempt to indicate what I think are the most interesting ramifications for philosophy of each of the quadrants shown. The emphasis will be on normative theory.

2. *Individual Normative Theory.* A common problem besetting both the theory of induction and moral philosophy is that of giving an adequate account of the concept of rationality. The normative theory of individual decision making has been concerned to explicate the notion of rationality in what is, in some respects, a very thorough fashion. It would be rather absurd in this general paper to attempt to survey even a small fraction of the many substantial results, primarily of a technical nature, that have been achieved in the last two decades. There is, however, a central kind of difficulty that has arisen and that I think is of great philosophical importance.

Let me begin in an indirect fashion by a comparison with the situation in the foundations of mathematics. As work on the arithmetization of analysis and the development of the theory of sets progressed in the nineteenth century, it seemed possible, at least for a short period, that the whole of mathematics could be derived from three simple postulates: the principle of abstraction, that is, that for any property there exists a set of elements having this property; the principle of extensionality, that is, that two sets are identical just when they have the same members; and the axiom of choice. Mathematical work in the nineteenth century indicated that the theory of sets was a natural framework within which to construct the rest of mathematics.[1] The intuitive and simple reasonableness of these three principles (with the possible exception of the axiom of choice) seemed overwhelming.

The discovery of paradoxes derivable from the first two assumptions by Burali-Forti, Russell, and others shook the foundations of mathematics to such an extent that a full recovery has not yet been made. But the discovery of the paradoxes has had an affect that

[1] This sketch is not meant to be historically exact.

has not been entirely negative. The many attempts, ranging from intuitionism to formalism, to put the foundations of mathematics on a secure basis have brought our understanding of the nature of mathematics to a new level of sophistication. Related negative results like the incompleteness theorems of Gödel have shown that what seemed to be immediate and obvious intuitions about even so restricted a branch of mathematics as elementary number theory are unreliable.

Recent work in decision theory has shown in similar fashion that there is no simple coherent set of principles capable of precise statement that corresponds to naive ideas of rationality. Just as research in this century in the foundations of mathematics has shown that we do not yet know exactly what mathematics is, so the work in decision theory shows that we do not yet understand what we mean by rationality. I mean by this not merely that we have no adequate general definition of rationality, but that, even for highly restricted circumstances, it turns out to be extremely difficult to characterize what we intuitively would want to mean by a rational choice among alternative courses of action. To focus on a collection of special situations and to attempt to characterize a rational strategy of choice for them is very similar and, in fact, closely related to an attempt to solve particular problems of induction without necessarily resolving "the" problem of induction.

The formidable problems besetting the rational decision maker may be illustrated by considering the difficulty of formulating an adequate principle of choice for finite games against "nature." Such games are special cases of the description given above of the general decision situation, although a wide variety of decision situations may be mathematically represented as such finite games. The game may be represented by a matrix in which a player must choose a row and a column is chosen by nature. The entries in the matrix represent the payoff to the decision maker, when he chooses a row and nature chooses a column.—The metaphorical talk about nature should not mislead anyone. This is merely a way of referring to a situation in which a course of action must be taken against an opponent or in an objective situation about which there is no information. The restriction to no information is not so severe as it may seem, for a situation in which partial information is incorporated may be redefined to yield a game with no information—a familiar practice in the mathematical theory of decisions. The happy situation when the entries in some one row are for every column better than those for any other row presents no difficulties. We simply apply the sure-thing principle; i.e., we choose that clearly preferred row. Unfortunately this is the unusual situa-

tion, and this principle in itself is seldom selective of a unique course of action.

The classical principle of indifference of Laplace states that when we have no information about the different possible states of nature (the columns of the game) we should assume that the probabilities of each are equal and then choose that row whose expectation is maximal. The much more conservative minimax principle states that we should choose according to the hypothesis that we may expect the worst and thereby minimize our maximum loss. The minimax principle was first proposed by von Neumann as a sound principle of strategy in games against an intelligent opponent. Its extension to games against nature as a fundamental principle of statistical decision theory was first made by Wald. A variety of other principles have been proposed.

In view of the numerous suggestions that have been made, particular interest may be attached to John Milnor's analysis [2] of what would seem to be the desirable characteristics of any rational principle of selection. His results, like those of Russell's paradox for the foundations of set theory, yield an impossibility theorem. Briefly, Milnor proposes the following nine axioms for any fully acceptable principle of choice. First, the principle must order the alternative courses of action. Second, this ordering must be independent of the arbitrary ordering of the rows and columns of the matrix. Third, the principle must be compatible with the sure-thing principle mentioned earlier; that is, if one row dominates another row in every column, that first row must be preferred. Fourth, the principle must satisfy an obvious condition of continuity. Fifth, the preference ordering of the courses of action must be unaffected by a linear change in all entries in the matrix. (This principle reflects the general result that the utility of consequences is measured only on an interval scale.) Sixth, the principle of choice must satisfy the condition of independence from irrelevant alternatives, namely that the ordering between old rows must not be changed by the addition to the game of a new row. The seventh axiom asserts that the principle must be invariant under a linear change of any column; in other words, the ordering among courses of action must not change if a constant is added to some column. Eighth, the ordering generated by the principle of choice must be indifferent to a column duplication; that is, the ordering must not change if a new column identical with some old column is added to the game. The ninth axiom asserts the require-

[2] "Games against Nature," in *Decision Processes*, edited by R. M. Thrall, C. H. Coombs, and R. L. Davis (New York, Wiley, 1954), pp. 49–59.

ment of convexity, which means simply that, if a row is equal to the average of two other rows judged equivalent in the preference ordering, then the first row must not be judged worse than the two equivalent rows of which it is an average. This axiom is sometimes described as asserting that the decision maker should not be prejudiced against randomization.

None of the familiar criteria of rationality for the decision maker is compatible with all nine of these axioms. These nine axioms are obtained by extracting desirable properties of various criteria of rationality that have been proposed. Milnor shows, for example, that the Laplacean criterion is characterized by axioms 1, 2, 3, 6, and 7 and the Wald minimax criterion by axioms 1, 2, 3, 4, 6, 8, and 9. The undesirable negative result is that no criterion satisfies all nine together.

Closely related paradoxical results in confirmation theory suggest that the naive theory of rationality, like the naive theory of sets, cannot easily be systematically reconstructed in any simple and consistent fashion.

Certain moral philosophers will undoubtedly be inclined to dismiss the kind of results I have been discussing as applicable only to the technical problems that arise in the theory of induction. In their minds, the concept of rationality I have been discussing has little if any relevance to the concept of rationality that arises in moral theory. This I think is clearly a mistake, as the discussion in the next section is meant to show.

Elsewhere [3] I have discussed the difficulties of characterizing the pure theory of rationality even when we accept a principle of choice, in this case the Bayesian principle that enjoins the decision maker to maximize his expected utility (the expectation being relative to a subjective probability distribution on the possible states or strategies of nature). Roughly speaking, the pure theory permits no structural assumptions about the environment, but is intended rather to hold always and everywhere. An example of an axiom of the pure theory is the postulate that the preference relation on the set of acts is transitive. A typical structural axiom is the assumption that the decision maker can partition the set of states of nature as fine as he pleases in terms of probability, an assumption which seems to me irrelevant to the pure concept of making a rational decision. I shall not try to summarize the technical results obtained, but only remark that they are all essentially nega-

[3] "Some Open Problems in the Foundations of Subjective Probability," *Information and Decision Processes*, edited by R. E. Machol (New York, Wiley, 1960), pp. 162–169.

tive and indicate the difficulties of axiomatizing even the simplest part of the pure theory of rationality.

By concentrating on the difficulties facing the development of any adequate concept of rationality, I do not mean to imply that the philosophical significance of decision theory under the heading of individual normative theory is restricted entirely to negative results. The revival of subjective probability, for instance, within the framework of decision theory by L. J. Savage and others has already had and will undoubtedly have further repercussions in the foundations of the theory of probability. The simple relative-frequency theories of von Mises and Reichenbach are already beginning to seem old-fashioned.

The development of utility theory within the general framework of decision theory has brought the kind of calculus envisioned long ago by Bentham to a high degree of technical perfection. It is unfortunate that the word "utility" is connected in most philosophers' minds with hedonism. The formal calculus of utility developed in recent decades is no more committed to a calculus of pleasure than to one of duty. The important intellectual contribution of the hedonistic tradition has been the recognition that some principle of calculation is required for rational action in the face of partial or incomplete information. Some years ago Davidson, McKinsey, and I [4] attempted to show the close relations that exist between the measurement of utility and a formal theory of value. I shall not attempt here to restate the arguments we gave, but only reiterate my conviction that the recent formal theories of utility are as important for moral philosophy as are recent theories of subjective probability for the philosophical problems of induction.

3. *Group Normative Theory.* In the table shown above three disciplines are listed in this quadrant: game theory, welfare economics, and political theory. The relations of game theory to the concept of rationality as discussed in the preceding section are apparent and will not be considered in any detail here. Suffice it to say that one of the most satisfactory analyses of the concept of rationality exists for competitive games. This is the classical minimax theory of von Neumann already mentioned.

In this section I should like to turn from the concept of rationality to the more specifically moral concept of justice.

I have the impression that for many years the consideration of political theory by either philosophers or political scientists has been nearly identical with consideration of the history of political

[4] "Outlines of a Formal Theory of Value, I," *Philosophy of Science*, 22 (1955): 140–160.

theory. It is my feeling that the newer welfare economics has broken out of its specific economic context and has already laid the foundations of a new approach to political theory. The central problem of welfare economics has been the Benthamite one of devising and analyzing a variety of schemes for the distribution of economic goods in the society. It has gradually come to be realized that the restriction to economic goods can be dropped and that the problem may be regarded as the more general one of setting social and political policy. It is assumed in this theoretical work that in some sense decisions reached reflect in an equitable and just manner the values and tastes of the members of the society. In fact, the preferences of the individuals making up the society are usually and somewhat unrealistically taken as given data.

To indicate the kind of results that may be obtained by the methods of welfare economics, I should like to sketch another impossibility theorem. This is due to Kenneth J. Arrow [5] and is concerned with the existence of a just or equitable method of social decision. We suppose that there are a number of possible social states and that each member of the society has a preference ordering for these states. The problem is to construct an intuitively reasonable social preference ordering from the given individual orderings. One simple proposal is, of course, the method of majority decision. Social state A is preferred to social state B by the group as a whole if a majority of the members of the group prefer A to B, otherwise not. Just as in the case of the Laplacean principle of indifference or the minimax principle for choosing a strategy in games against nature, there are intuitively desirable axioms that are violated by the method of majority decision. Perhaps the easiest way to illustrate the difficulties is to describe the so-called paradox of voting. Suppose there are three issues A, B, C and three people voting on these issues. Let us assume that the first person prefers A to B to C, the second person prefers B to C to A, and the third person prefers C to A to B. The issues are voted on in pairs. It is easily checked that if the first choice is between A and B the selected issue will be C. If it is between A and C the outcome will be B, and if it is between B and C the outcome will be A. In other words, the outcome chosen is completely dependent in this symmetrical situation on the arbitrary choice of which issues are to be voted first.

What Arrow has done is to proceed as Milnor did, namely, to write down reasonable axioms that any social decision method should satisfy and then to ask if indeed there exist any methods

[5] *Social Choice and Individual Values* (New York, Wiley, 1951).

satisfying the axioms. In essence his four axioms are the following. The first axiom postulates a positive association of social and individual values. In particular, if one alternative social state rises in the ordering of every individual without any other change in the orderings, it is natural to postulate that it rises or at least does not fall in the social ordering. The second axiom states the independence of irrelevant alternatives. The meaning of this postulate is that if, for instance, a set of candidates is being considered for an office and the preferences of voters for these candidates are known, then the deletion of one candidate from the list will not affect the relative preferences for the other candidates. In thinking about this postulate it is important to emphasize that strategic considerations are not being considered. We are concerned with the actual preferences of the group members and not with their behavioral use of a strategy in those situations where they feel their first choice could not possibly be elected. For instance, in the 1948 presidential election a strong states' rights advocate might have preferred J. Strom Thurmond to Thomas E. Dewey, but, because he felt that Thurmond did not have a chance, he may have voted for Dewey. It is this kind of strategic consideration that is being ignored in this postulate. The meaning of the postulate is that, if the states' rights conservative preferred Thurmond to Dewey to Truman to Wallace, then if Dewey were no longer a candidate he would retain the same ordering of preferring Thurmond to Truman to Wallace, and similarly the deletion of any one of the four candidates would not disturb the preference ordering for the remaining three, even though the deletion of one of the four might affect his voting behavior. The third axiom asserts that the social decision method is not to be imposed. A decision method is said to be imposed when there is some pair of alternative social states X and Y such that the community can never express its preference for Y over X no matter what the preferences of all the individuals concerned may be. The existence of outmoded taboos furnishes examples violating this condition. The fourth axiom asserts that the social decision method shall not be dictatorial; that is, the preferences shall not simply correspond to those of some one individual in the social group. Unfortunately, Arrow is able to prove that, if there is any degree of variety in the individual preference orderings, then there exists no social decision method satisfying the four axioms stated.

The point I have been concerned to make here is that the Milnor and Arrow impossibility theorems are as philosophically relevant to the foundations of the concepts of rationality and justice as are

the paradoxes of set theory for the foundations of mathematics. These impossibility theorems demonstrate that our naive intuitions about rationality and justice cannot be counted upon to yield a coherent and consistent theory.

4. *Descriptive Theory.* Perforce the descriptive or behavioristic theory has less direct relevance to philosophical problems than the normative theory, but there is one particularly important point I should like to describe in the brief space remaining. There are a rapidly increasing number of experimental studies of the actual decision behavior of human beings. Studies have been made of value choices, of actual inductive behavior, and of behavior in competitive or cooperative game contexts. The literature is too large to review here, but an important general conclusion seems to be that actual behavior deviates rather sharply from the normative models, even in the case of strictly competitive games. On the other hand, the behavior of experimental subjects in many cases corresponds well with quantitative predictions derived from learning theory formulated in terms of stimulus sampling and conditioning.[6]

Stimulus-sampling learning theory was first given a quantitative formulation in 1950 by the psychologist W. K. Estes. It has since been developed by a number of investigators. In a highly simplified form, the basic ideas run as follows. The organism is presented with a sequence of trials, on each of which he makes a response that is one of several possible choices. In any particular setup it is assumed that there is a set of stimuli from which the organism draws a sample at the beginning of each trial. It is assumed that on each trial each stimulus is conditioned to at most one response. The probability of making a given response on any trial is postulated to be simply the proportion of sampled stimuli conditioned to that response, unless there are no conditioned stimuli in the sample, in which case there is a "guessing" probability for each response. Learning takes place by the following mechanism. At the end of a trial a reinforcing event occurs which identifies that one of the possible responses which was correct. With some fixed probability the sampled stimuli become conditioned to this response if they are not already, and the organism begins another trial in a new state of conditioning.

Independent of the question of empirical adequacy in predicting actual choice behavior, behavioral scientists have more general

[6] A number of experiments of this sort are reported in the book by Richard C. Atkinson and me, *Markov Learning Models for Multiperson Interactions* (Stanford: Univ. Press, 1960).

reasons for preferring a learning theory (like the stimulus-sampling variety just sketched) to decision theory. To the experimental psychologist, the static character of the concepts of subjective probability and utility is suspect, and these two concepts are the central concepts of decision theory. The psychologist resists accepting them as basic or primitive concepts of behavior. Ideally, what he desires is a dynamic theory of the inherent or environmental factors determining the acquisition of a particular set of beliefs or values. If these factors can be identified and their theory developed, the concepts of probability and utility become otiose in one sense. I have recently tried to show how stimulus-sampling learning theory provides the beginnings of such a development.[7] The philosophical interest of the behavioristic approach lies in the possibility of constructing a more realistic framework than the static one of decision theory for discussing the normative theory of choice. This is not to say that the distinction between normative and descriptive questions is to be abolished. It is rather that the more detailed and fundamental behavioral theory opens up the possibility of analyzing normative questions at a deeper conceptual level, but this is a matter that cannot be explored here.

PATRICK SUPPES

STANFORD UNIVERSITY

[7] "Behavioristic Foundations of Utility," *Econometrica*, **29** (1961): 1–17.