Reprinted from the Philosophy of Science
Vol. 31 No. 2 April 1964

DISCUSSION: ON AN EXAMPLE OF UNPREDICTABILITY IN HUMAN BEHAVIOR

BY

PATRICK SUPPES

Reprinted from the Philosophy of Science
Vol. 31 No. 2 April 1964

DISCUSSION: ON AN EXAMPLE OF UNPREDICTABILITY IN HUMAN BEHAVIOR*

PATRICK SUPPES

Stanford University

Scriven's example of essential unpredictability in human behavior goes like this. Assume a rational predictor P who wants to infer from information available to him the choice behavior of an individual C, where (i) C is choosing rationally and intelligently, (ii) C is *a contrapredictive* relative to P, i.e., C wishes to falsify any prediction made by P about his, C's choice behavior, (iii) C knows the information available to P. It is easy to make these conditions more precise and then to prove that P cannot predict C's behavior. Scriven fills out the bare bones of this example with many illustrations and comments, which I shall not attempt to sketch. His example of essential unpredictability is one I accept as in the main correct. The critical tenor of my remarks is to place the example in perspective from the standpoint of the theory of games and to challenge the significance of the example for the development of a quantitative and predictive behavioral science.

Consider the following simple game played under two different conditions. Following Scriven, let us call the two players C and P. We shall require that player P (the predictor) move first. Player P chooses action b_1 or b_2. Player C then moves, choosing action a_1 or a_2. Related to Scriven's example, we may interpret b_1 as a prediction by P that C will choose a_1, and b_2 the prediction by P that C will choose a_2. The payoff matrix for the game we may take to have the following simple form

C \ P	b_1	b_2
a_1	-1	1
a_2	1	-1

We shall make it zero-sum by assuming that when C obtains 1 unit, P loses 1 unit, and vice versa. (The exact payoff outcomes are not essential for the discussion, and it is not necessary to assume the game is zero-sum or even constant sum, but we may as well for simplicity.)

From the standpoint of the theory of games, the fundamental question concerning the rules of this game is whether C does or does not know P's choice when he makes his own move. If C is informed of P's choice, then the game is one of perfect information and according to a theorem that originates with Zermelo, the game is strictly determined. (A game is strictly determined when optimal strategies for both players are non-random pure strategies.) Zermelo's theorem asserts that this is the case for

* These comments were read on May 2, 1963 at the annual meeting of the Western Division of the American Philosophical Association at Columbus, Ohio. They were directed at Scriven's paper, "An Essential Unpredictability in Human Behavior," which is to be published in *Scientific Psychology* (edited by E. Nagel and B. Wohlman). I have very briefly abstracted his main argument in order to make my remarks intelligible without familiarity with his paper.

games of perfect information. As is obvious for this simple game, when played under conditions of perfect information, C is in a position always to win. This is certainly a trivial and obvious conclusion.

If the rules of the game are now changed in a fundamental way so that P's move is not known to C, then the game is no longer one of perfect information, and pure, non-random strategies are no longer optimal. For the outcome payoffs indicated, the optimal, minimax strategies for both P and C are to choose one of their two alternatives with probability 1/2. If changes are made in the outcomes to upset the perfect symmetry of the present matrix, then the exact numerical randomization that should be followed for optimality will be something different from 1/2. But these minor quantitative variations do not seem important enough to pursue. Moreover, the restriction to two choices each for P and C is not essential; it is a straightforward matter to consider the case of n alternatives, and it is technically but not conceptually complicated to consider a continuum of possible choices for each of the two players.

Scriven's example of essential unpredictability of human behavior seems to me to be simply a special case of such a simple two-person competitive game. His second condition on C is the condition that assures us that C is playing the game, namely, the condition that C is a contrapredictive relative to P.

To show why I think that Scriven is making claims that are too large for his example, I want to distinguish first various cases of prediction for the behavior of players in a game. For those who prefer Scriven's language of contrapredictives, the translation is simple and direct. Before considering these cases, it is important to emphasize that prior to the analysis of the predictability of C's moves, we must fix upon the game C is playing. We can scarcely be expected to analyze or make assertions about the predictability of C's moves if after we have made the predictions, we are told that C is not playing this game, but some other game. So I take it in the following analysis that C and P are playing a game with moves and outcomes as specified above. If, in fact, they are not playing this game but some other game, then what is to be said about predictability must be changed. This talk about games should not be taken in too restrictive a sense, for it is a way of saying what C is motivated to do. Consider, for example, the first case I shall describe. The game is the game of perfect information already described. P's problem is to predict what move C will make. Now as a side bet, with another person P', P says that he will predict, separate from the game, what C's moves will be. In this game of perfect information, if C is playing it with serious intent, it is perfectly clear that on the basis of side bets P can indeed make such a prediction in a deterministic fashion. Namely, when P chooses move b_1, C will choose a_2 and when P chooses b_2, C will choose a_1. That is, when P predicts in the game that C will choose a_1, then in actual fact, C will choose a_2 in the game, etc.

There is nothing surprising or paradoxical in this. As most parents know, a situation very close to this arises in the age of most children. Namely, the parents will tell or ask the child to do something and in a very high probability of cases the initial response of the child will be negative. *Sotto voce* or in conversation with another person, the parent can reliably predict this response on the part of the child. We are able to make such predictions by first specifying the game that C, (here, the child,) is playing. If C or the child is concerned with the actions, side bets, motives, etc., of person P' he can begin to play another game and behave in a different way, but it is a matter of observation to decide what game he is playing. By these remarks I don't mean to suggest that if we begin to analyze complicated social behavior on the part of candidates for office,

lawyers in courtrooms, or lovers in parked cars, we can precisely describe the game the participants are playing and readily predict their moves. Our understanding of the details of behavior are not yet that good. My point is just that the conceptual schema introduced by Scriven does not raise any new or essentially surprising difficulties.

Holding these remarks in mind, let us look at some of the possible cases of prediction that may arise in the context of simple games.

Case 1A. C and P are playing a game of perfect information. P's moves in the game, intended to be predictions of C's moves, are bad indeed. The game is as defined above, and P is always wrong.

Case 1B. P is making side bets about Case 1A with person P'. The side bets are always correct. If C does want to change the game and now play a game of perfect information defined against both P and P', then it will in general be possible for P to make new side bets with person P'' predicting with probability 1, C's behavior in the enlarged game, provided of course that C is playing the new enlarged game with the same seriousness he played the original game.

Case 2A. C and P are now playing the game in the condition of imperfect information described above. In this case the optimum strategy for both players is to use a randomized strategy and no deterministic predictions within the game or as side bets are possible. If essential unpredictability is defined in terms of strictly deterministic predictions then the existence of this case certainly supports Scriven's thesis. Simple competitive games of imperfect information played well by both players provide immediate examples of essential unpredictability in the sense of strictly deterministic predictions.

Case 3A. We give up the hope for deterministic predictions, as has already been done in much of empirical science, and ask what are the possibilities of probabilistic predictions. In the game of imperfect information played by C and P it is now possible for P to make a side bet of a probabilistic sort concerning the behavior of C. For the payoffs as defined above, and in the environment in which C is playing the game seriously, P would do well to bet that C will use a minimax randomized strategy. P can profitably balance his bets to come out well in either case. If C does use this strategy, P will win the side bet. If C does not use this strategy, P can win at the game by taking advantage of C's departure from a minimax strategy.

Case 3B. C becomes aware that P is making side bets about his, C's probabilistic behavior in the original game and decides to take the falsification of P's side bets more seriously than the original game. In this case we enter a new game and the same analysis can be repeated, although the strategy space is now a much enlarged one. Having fixed this new game, P can in the same way enter upon side bets concerning C's behavior in the new game.

This enumeration of cases is not meant to be exhaustive. More importantly, the talk about games and side bets can be replaced by talk about motives and the study by P of C's motivations and goals. My point thus far has been twofold. In the first place, I agree with Scriven that the deterministic prediction of human behavior is impossible in many situations. Simple competitive games provide perhaps the best examples of such situations. But, secondly, I maintain that probabilistic prediction is not only possible but feasible for a very wide class of situations. If classical physics occupied the position it did at the end of the nineteenth century, modern behavior theory would undoubtedly be criticized and judged inferior for its probabilistic rather than its deterministic character. In actual fact, however, the kind of probabilistic

behavior theory which I would now like briefly to discuss, is in general methodology very similar to quantitative theories in physics. Modern physical theories are no longer deterministic but thoroughly probabilistic in character. It is not possible to argue effectively that probabilistic theories are needed for specifically human behavior, whereas fully deterministic theories are appropriate and adequate for non-human or inanimate behavior. To substantiate the claim that a predictive behavior theory and predictive behavior laws can be developed, what is wanted, it may be said, is the kind of predictive quantification found in physics: From a knowledge of the present state of the organism and its environment we should be able to predict its future state, at least in a probabilistic sense, for the not too distant future and for a moderately diverse even if restricted set of environmental conditions.

Mathematical behavior theory, as developed in recent years by psychologists, provides perhaps the most important example at present of the kind of theory just described. It would be out of place to describe in detail the nature of this theory, but its main features can be sketched as well as some of the problems it successfully handles. It should be emphasized from the start that modern behavior theory is thoroughly probabilistic rather than deterministic. Abandoning the development of a deterministic theory has undoubtedly been one of the main reasons for the considerable number of quantitative successes of the new theory. It is recognized that the very large number of underlying physiological mechanisms cannot yet be connected in an explicit and detailed way to overt behavior. On the other hand causal relationships of a probabilistic character, holding between behavioral variables, have been postulated and approximately verified.

A causal quantitative analysis of human behavior is the sort of thing often considered impossible in principle. It will be useful to consider one simple but fundamental example of a learning experiment that may be very satisfactorily analyzed within the framework of mathematical behavior theory.

The task for a subject in this experiment is to learn the arbitrary associations set by the experimenter between a given list of nonsense syllables and two telegraph keys. On each trial the subject is shown a nonsense syllable, for example, *XUH*, and then asked to punch that key with which it is associated. After one of the keys is punched, the trial is terminated by the experimenter flashing a light, let us say, or using some other similarly simple device, to show the subject which key was in fact the correct one. Of course, on the first trial it is necessary for the subject simply to guess which is the correct response to make, that is, which key to press, but as would be expected, the subject soon uses the information furnished by the lights above the keys. Even the most gross empirical analysis of the data from such experiments shows that the relation between receiving the information as to which key is correct and behaviorally learning to respond correctly is not deterministic. For example, neither the number of errors nor the trial number of last error is the same for all the nonsense syllables. On the other hand, a very simple probabilistic behavior theory gives an excellent account of the data. The intuitive idea is that each nonsense syllable may be represented as a single stimulus. At the beginning of the experiment no stimulus is conditioned or connected by an association bond to its correct response. It is postulated that on each trial there is a constant probability that the appropriate association bond will be formed. It is also postulated that until this association bond is formed the subject is simply guessing the correct answer. After the correct association is formed, the subject makes the correct response on all subsequent trials. From these postulates it is possible to derive a large number of probabilistic predictions about the behavior of subjects,

and these predictions have been confirmed to a remarkably accurate degree in a number of experiments.[1]

Now those who challenge the very idea of quantifying human behavior will have at least two objections to this simple learning experiment. First, they will wish to point out that in the usual experiment of this kind data are analyzed for an entire group of subjects together and not in terms of individual behavior. Their line of criticism is then, "Yes, we will admit it is possible to develop some sort of theory for the gross characteristics of group behavior, but it is not possible to do this for the fine nuances and differences of individual behavior. Those characteristics that are special for a given individual can in no sense be accounted for by the kind of theory you describe, which is wholly devoted to the average or norm of the group." However, this view represents both a misunderstanding of the experiment and the theory used to explain it. For several kinds of reasons, many of which are administrative rather than scientific, it is often convenient to run learning experiments for a short period but with a large number of subjects. The theory however applies just as well to the learning behavior of a single subject over a number of hours. In our laboratory at Stanford we have been performing such extensive experiments on individual subjects and, as a result, analyzing and testing the theory in terms of individual behavior. A typical example of such an experiment is one in which the subject is asked to learn a large number of vocabulary items from a foreign language. In one experiment, for instance, the subject hears a Russian word and then is shown three English words. His task is to select the English word that has approximately the same meaning as the Russian word he has heard. The experimental setup and the theory used to analyze the data are very simple and direct extensions of the experiment with nonsense syllables already described. The important point is that we can ask in a direct and meaningful way if the behavior of an individual subject over the course of fifteen or twenty days satisfies the theory.

The second and perhaps more important objection to the experiment with nonsense syllables, or the second experiment with Russian vocabulary, is that although principles of stimulus-response association may be used to explain the simplest sort of learning, it is not possible to use these principles to develop a quantitative theory for more complicated human behavior. It will perhaps be claimed that no simple principles of association between stimuli and responses can possibly account for the learning of even the most elementary concepts, in particular, mathematical concepts, the learning of which must go beyond any completely simple principles of stimulus association. For the past several years we have conducted extensive experimental investigations on the learning of mathematical concepts by young children. A typical instance of one of these experiments is the following. Children of kindergarten age are asked on each trial to push one of three keys. Above the first key is shown a small triangle, above the second a small quadrilateral, and above the third a small pentagon. On each trial the child is shown a large figure that is either a triangle, quadrilateral, or pentagon and is asked to push the appropriate key. The child cannot learn the concept of triangle, quadrilateral or pentagon in this experiment by any simple principles of stimulus association, because on each trial he is shown a different figure, that is, the same triangle, the same quadrilateral or the same pentagon is never shown on more than one trial. On each occasion when a quadrilateral, for example, is shown,

[1] See, e. q. G. H. Bowen, *Psychometrica,* 1961, **26,** 255-280 and P. Suppez and R. Ginsberg, *Psychological Review,* 1963, **70,** 139-161.

the child is presented with a new quadrilateral having a different shape and orientation from any of the previous ones he has seen. On the other hand, the kind of theory described above gives a very good quantitative account in probabilistic terms of the data of this experiment, provided that the role played earlier by the nonsense syllables or Russian words is now played by the three simple geometrical concepts. In other words, the association is established between the concept and the appropriate response rather than between a single stimulus and the appropriate response.

In claiming that we are adequately predicting quantitative aspects of the behavior of children in learning mathematical concepts, I would not want to be interpreted as saying that we have anything like the final story on such behavior, for there are many deep questions concerned with how the concepts are formed, or with how the concept-response association is established, which we cannot yet answer. For the present purpose, however, the important point is that we do have sufficient theory to provide a very satisfactory quantitative analysis of the behavior.

Contrary to Scriven's casual claim that geography and geology rather than physics are the appropriate models for psychology, I wish to assert, and have tried to show, that we are already well on the way to a mathematical behavior theory adequate for the analysis of many simple experiments and of such a character that it has the kind of mathematical viability and predictive "feel" about it that we expect of physical theories. Let me conclude by saying that the kind of behavior theory I have described has very good philosophical roots. The associationist psychology of Hume is an obvious precursor. Hume, I feel, would be very happy with the all-or-none learning laws now well confirmed in a variety of experiments concerned with stimulus-response associations or the formation of simple concepts.

PRINTED IN BRUGES, BELGIUM, BY THE ST CATHERINE PRESS LTD.