***Topoi* 5** (1986) 3-19.

A Foundation for Causal Decision Theory*

Brad Armendt
Ohio State University

The primary aim of this paper is the presentation of a foundation for causal decision theory. This is worth doing because causal decision theory (CDT) is philosophically the most adequate rational decision theory now available. I will not defend that claim here by elaborate comparison of the theory with all its competitors, but by providing the foundation. This puts the theory on an equal footing with competitors for which foundations have already been given. It turns out that it will also produce a reply to the most serious objections made so far against CDT and against the particular version of CDT I will defend.[1]

Causal decision theory comes in four versions: Gibbard & Harper's (1976), Skyrms' (1979), Lewis' (1981), and Sobel's (1978). The theories are in spirit very much alike, but the differences between them are philosophically significant. A good description and comparison of the four versions is given in Lewis (1981). While Lewis prefers his formulation of CDT, I prefer Skyrms', and it is for Skyrms' version that the foundation is supplied.

In accordance with standard practice, this foundation consists of 1) a set of axiomatic conditions on rational preference systems, and 2) the derivation of a representation theorem which shows that for any preference system satisfying the axioms there exist a probability measure P and order-preserving utility functions U which represent the preferences, and which are related by the theory's general utility rule. The representation theorem has a uniqueness part which shows that P and U are not arbitrary: for each preference system P is uniquely determined and U is unique up to positive linear transformation. (Here I am describing the theorem presented in this paper; some representation theorems for other theories contain weaker uniqueness conditions.) I should make it clear from the beginning that the foundation I provide for the theory relies on formal results of Fishburn's in his (1973). My alterations of his formal theory are slight. I do reinterpret his theory somewhat. The application of that interpretation to causal decision theory is new.

* The ideas contained in this paper were presented and discussed in talks at the University of Minnesota at Duluth and Brown University; an early version of the paper was read at a conference on decision theory run by the Philosophy Department at Ohio State University in February, 1984. I benefited from each of those discussions. This paper is based on parts of my 1983 doctoral dissertation at the University of Illinois, Chicago. I especially thank Brian Skyrms and Paul Teller for their excellent criticisms and advice.

[1] There are, of course, additional defenses of the claim that CDT is a superior theory that can be given. I will mention those defenses as I describe CDT and the motivations behind the theory, but I will not here rehearse or comment on all the recent debate. Interested readers should see Gibbard & Harper (1976), Skyrms (1979), Lewis (1981), Jeffrey (1981, 1983), and Eells (1982).

Topoi **5** (1986), 3-19.

1. Causal Decision Theory.

It is an old and familiar idea that a rational agent with only human ability to know the future should choose that action which among all the alternatives has the maximum expected value (or, in case of ties, one of the actions with maximum expected value).[2] In Jeffrey's (1965) theory of rational decision this idea is incorporated as the principle that the rational choice is the action A with maximum utility $V(A)$, which can be computed from utilities and probabilities of A's possible consequences according to the formula

$$(1) \qquad V(A) = \textstyle\sum_i P(C_i /A)\, V(A \,\&\, C_i).$$

The quantities $P(C_i /A)$ are the agent's subjective conditional probabilities (degrees of belief) that consequence C_i is true (will occur), conditional on A (*i.e.* the performance of the act described by the proposition A). It is understood that—and formula (1) is only valid if—the set of propositions $\{C_i\}$ is a partition. The quantities $V(A \,\&\, C_i)$ are the values the agent's utility function gives to each of the possible outcomes, "I do A and C_i occurs." We should note that in fact (1) holds for *any* proposition A and any partition $\{C_i\}$ such that the conditional probabilities are well defined, but in decision-making contexts (where the values of *actions* concern us) the formula is most useful when applied to an act proposition A and consequence propositions C_i. The basic intuition behind the recommendation that acts be evaluated in terms of their expected values is the idea that the utility of an act is the weighted average of the values of its possible consequences, the weights being measures of the chances the consequences have of following the act. Jeffrey's decision theory is sometimes known as conditional expected utility (CEU) decision theory since the weights are taken to be the agent's conditional (on the act) degrees of belief that each consequence will result.[3] (It is also, since Gibbard & Harper (1976), known as *V-maximization* decision theory, in contrast to CDT's *U-maximization*.)[4]

It may be tempting to regard these conditional probabilities as the agent's estimates of the objective chances each C_i has for happening, which are (or would be) fixed by the state of the world and the agent's performance of A. Although this might be sometimes so, it important to remember that it is not so in general. It may be that the agent's partial ignorance of the world

[2] Jeffrey (1981) points out that the idea is at least as old as the *Port Royal Logic* (1662):

> ...to judge what one ought to do to obtain a good or avoid an evil, one must not only consider the good and the evil in itself, but also the probability that it will or will not happen; and view geometrically the proportion that all these things have together... (IV, 16)

[3] Another well-known decision theory, that of Luce & Krantz (1971), is also known as a conditional expected utility theory. In the expected utility rule of their theory, the weights are indeed conditional probabilities, but the theory is quite different from Jeffrey's: the objects of preference for Luce & Krantz are conditional decisions, which are functions from states to consequences; since the states and decisions are independent in their treatment, from our point of view their theory has more in common with Savage's and Fishburn's than with Jeffrey's, and it might even be developed into a causal decision theory. The feature of the Luce & Krantz theory which I find objectionable is their use of disjunctive decisions, each disjunct being conditional on different events. Even if isolated examples of such decisions make some sense, they are in general highly unintuitive.

[4] V-maximization has recently become more sophisticated than the description I provide indicates. Defenders of V-maximization have elaborated on the theory in response to the criticisms of causal decision theorists. Their sophisticated sorts of V-maximization are very interesting, but not because they are theories as good as causal decision theory. (Only unreasonably strong assumptions will save these V-maximizing theories from *approximating* the correct answers that CDT gives, or from getting the correct answers *almost* all the time.) I will confine my account to V-maximization in its unsophisticated form. See Jeffrey (1981, 1983) and Eells (1982) for the variations.

leaves him in doubt about, for example, which of two propositions B and $\neg B$ holds, and he may recognize that it makes a difference in his decision problem in the following way: Perhaps he believes that if B is true, the objective probability of C_1's following A is high, and he also believes that if $\neg B$ is true, the objective probability of C_1's following A is low. (We may suppose he believes this because B describes a part of the world which he takes to be a significant link in the (potential) causal connections between A and the C_i's.) If the agent has (and actually entertains, let us suppose) the sorts of beliefs described, then his degrees of belief $P(C_i /A \,\&\, B)$ and $P(C_i /A \,\&\, \neg B)$ are his estimates of the possible objective probabilities in the world; the $P(C_i /A)$'s will generally have values *different* from these estimates. Under the assumptions made above, the agent's degree of belief $P(C_1 /A)$ should be a weighted average of his beliefs $P(C_1 /A \,\&\, B)$ and $P(C_1 /A \,\&\, \neg B)$, the weights being the values $P(B)$ and $P(\neg B)$ respectively. (This follows from the probability axioms if $P(B) = P(B /A)$; if, in other words, there is no (epistemic) correlation between the agent's doing A and B's being true. A and B are then statistically independent. The importance of this condition will emerge below.) The degree of belief $P(C_1 /A)$ is a kind of *summary* of the agent's beliefs about the objective connections between the occurrence of A and the possible occurrence of C_1, but it may *not* coincide with any of the objective probabilities he thinks are serious possibilities. This is illustrated by the well-known example in which we imagine ourselves in the possession of a coin which we know to be biased either 2:1 in favor of heads, or 2:1 in favor of tails. If we think the probabilities that the bias is in either direction are equal, then the conditional (on the act of tossing the coin) degree of belief in the consequence heads is 1/2, which differs from our estimates of each of the objective probabilities we think possible.

The point of all this is to emphasize the epistemic nature of the probabilities appearing in (1). Not only are they subjective degrees of belief, but they generally are degrees of belief which are measures of epistemic connections among propositions, and not necessarily direct estimates of causal connections between the states the propositions describe. This can also be seen by simply observing that we need not have much evidence about the causal structures which connect some actions and their possible consequences in order to have useful conditional degrees of belief in the various C_i's given that A is done—observing repeated trials, or for that matter, having a hunch may produce them. And CEU theory is willing to use the conditional degrees of belief (if they are coherent), whatever their source.

Usually, reliance on the conditional probabilities (whether they are the kinds of summaries of beliefs about possible causal connections described above or not) and use of (1) leads to the intuitively correct answers in decision-making. CEU theory as it is briefly described above is usually a very good theory. In the account of the axioms and representation theorem for causal decision theory below, I will pay particular attention to showing that Skyrms' CDT does agree with CEU theory in all the cases where the latter gives the right answers.

What about the cases in which CEU theory goes wrong? Consider another sort of problem in which the agent has, in addition to the degrees of belief $P(C_i /A)$ useful in (1), some other beliefs about the causal structure of his problem.[5] Suppose the choice is between acts A_1 and A_2, and the agent believes (as a result of his knowledge of the causal structure, perhaps, though this is not the crucial point) that state S, to which he attaches a small positive utility, is likely to follow his action if he chooses A_2, but is not likely to follow A_1. So $P(S /A_2)$ is high and

[5] The following general description of a simple kind of causal decision problem is partly derived from Lewis (1981).

$P(S\,/A_1)$ is low. Suppose further that another state X is one which he greatly fears—he attaches large negative utility to it—and that the agent believes that the causal structure of the problem is this: his choice of A_1 or A_2 or anything else does not causally influence the occurrence of X, but X's occurrence or nonoccurrence *does* causally influence his choice. In fact, his choice of A_2 is much more likely than A_1 if X is true, so that choosing A_2 is a symptom of X's occurrence and his belief $P(X\,/A_2)$ is high.[6] Should the agent refrain from A_2 which he believes will probably lead to state S which he likes, in order to decrease the chances that X will (or did) occur? The answer should be *no*, since the agent believes his choice has no causal influence over X, but the example can be filled in so that CEU theory will say the opposite.

Fisher's smoking gene example.

I believe that my disposition to smoke cigarettes and my disposition to contract cancer are genetically influenced by the same factor, which accounts for the statistical correlation between smoking and cancer. I believe that I cannot influence my genetic makeup, and that smoking itself is not harmful. I enjoy smoking and attach small positive utility to the pleasure I would derive from smoking this cigarette. I attach a large negative utility to contracting cancer. If I believe that the causal connections and statistical correlations are strong enough, and if I disvalue contracting cancer enough, compared to how much I value the pleasure of smoking, then (1) and CEU theory will lead me to the decision to refrain from smoking, in order to minimize the chances I have the gene and contract cancer.[7]

Newcomb's Problem.

This best known of causal counterexamples involves a game played by the agent against a clever, accurate predictor which goes as follows: The agent is confronted with two boxes, one opaque and one transparent. In the transparent box is \$1000, and in the opaque box is either \$1,000,000 or \$0. The agent has two alternatives—he may either take only the contents of the opaque box (call this A_1) or he may take the contents of both boxes (A_2). *Before* he makes his choice the predictor, who has been very accurate in the past (say he has been correct 90% of the time,

[6] Alternatively, the choice of A_1 or A_2 might not be influenced by the bad state X but instead by a state Y which is a common causal influence of the choice and X. The choice is not believed to influence X's occurrence, but remains symptomatic of X's occurrence. This is a more accurate sketch of the smoking example and the Newcomb game discussed below.

[7] Suppose I value the pleasure I would derive from smoking this cigarette (act A_s) at 2 utiles, while I attach a large negative utility to contracting cancer, -1000 utiles. My utility function might say $V(S) = 2$, $V(\neg S) = 0$, $V(X) = -1000$, $V(\neg X) = 0$, where S is 'I enjoy this cigarette,' and X is 'I contract cancer.' Suppose I have these degrees of belief, where A_r is the act of refraining from smoking the cigarette:

$P(S \,\&\, X\,/\,A_s) = .7$	$P(S \,\&\, X\,/\,A_r) = .01$
$P(S \,\&\, \neg X\,/\,A_s) = .25$	$P(S \,\&\, \neg X\,/\,A_r) = .04$
$P(\neg S \,\&\, X\,/\,A_s) = .04$	$P(\neg S \,\&\, X\,/\,A_r) =.25$
$P(\neg S \,\&\, \neg X\,/\,A_s) = .01$	$P(\neg S \,\&\, \neg X\,/\,A_r) = .7$

(1) tells us that

$$V(A_s) = \Sigma_i\, P(C_i\,/\,A_s)\; V(C_i \,\&\, A_s)$$

$$= .7(-998) + .25(2) + .04(-1000) + .01(0) \approx -738$$

and $$V(A_r) = .01(-998) + .04(2) + .25(-1000) + .7(0) \approx -260.$$

whether one box or both boxes were taken by earlier players) forecasts the agent's choice and decides what the contents of the opaque box will be. He will leave it empty if he predicts the agent will take both boxes, and he will put $1,000,000 in it if he predicts the agent will take only that box. This is done before the agent chooses, and the agent strongly believes that the boxes are not tampered with. The agent is aware of the arrangement of the game and the predictor's past success. Given payoffs of sufficiently different utilities, (1) and CEU theory will lead to the recommendation that the agent take only one box, leaving the visible $1000, in order to increase the chances that in the past the opaque box was filled with the million dollars.[8]

Both of the answers given in the examples are wrong. If I believe (1) that nothing I do now can contribute to my having or not having the genetic factor (or if I just believe that the probability of such a contribution is very small), and (2) that the genetic factor is the main cause of my contracting cancer while smoking itself is not a cause of cancer, then the correlation between smoking and having the gene is *not* a good reason for avoiding the pleasure I would get from smoking. And if the agent believes that the money is arranged in the boxes before his decision and the arrangement is not thereafter altered by his choice, the predictor's reliability is not a good reason for leaving the $1000 that he might have. To act otherwise in either case is, as Gibbard and Harper (1976) remark, to knowingly act so as to produce *evidence* for a strongly desired state of affairs (absence of the gene, presence of the $1,000,000), without in any way *producing* the desired state, *even when such action has significant cost.*

An agent who follows CEU theory and the reasoning sketched above wrongly evaluates his alternatives because he fails to use his most specific and relevant information (or beliefs) in the given situations, namely his full information about the causal structure of his decision problem: As is the case in most decision problems, this agent has incomplete knowledge of the world and can identify possible states of affairs whose occurrence affects the chances that a given act will be followed by its various possible consequences. (I don't know whether I have the gene, but I do believe that having or not having it makes a big difference to my state of health which follows my smoking this cigarette). He can describe these possible states, and if he wishes, compute for each of them the value the act has if the state obtains. CEU theory tells him

(2) $V(A \,\&\, B_j) = \Sigma_i \, P(C_i / A \,\&\, B_j)\, V(C_i \,\&\, A \,\&\, B_j)$,

where B_j is the state. (This computation tells me what value I attach to smoking and not having the gene, in terms of my conditional degrees of belief in the possible consequences and the

[8] We can describe the possible consequences in terms of the money the agent receives:

T: $1000 MT: $1,001,000
M: $1,000,000 Z: $0

Suppose the agent's degree of belief that there is no cheating, etc. is very close to 1, to simplify the calculations. Then his degrees of belief might be:

$P(T / A_1) \approx 0$ $P(T / A_2) \approx .9$
$P(M / A_1) \approx .9$ $P(M / A_2) \approx 0$
$P(MT / A_1) \approx 0$ $P(MT / A_2) \approx .1$
$P(Z / A_1) \approx .1$ $P(Z / A_2) \approx 0$

CEU theory and (1) then yield (assuming the $V(C_i)$'s are measured by the money):

$V(A_1) = 0 + .9(1{,}000{,}000) + 0 + .1(0) = 900{,}000$
$V(A_2) = .9(1000) + 0 + .1(1{,}001{,}000) + 0 = 101{,}000.$

values I attach to each consequence conjoined with my smoking and not having the gene.) If the agent then wishes to use these values to compute the value of the act itself, CEU theory says

(3) $V(A) = \Sigma_j P(B_j /A) V(A \& B_j)$,

so by substitution, using (2),

(4) $V(A) = \Sigma_j P(B_j /A) \Sigma_i P(C_i /A \& B_j) V(C_i \& A \& B_j)$.

(4) might be useful to an agent who is more sure of his conditional degrees of belief which appear in that formula than in those which appear in (1), $P(C_i /A)$. But the formulas are equivalent: when the conditional probabilities in (4) are multiplied together, (1) follows from the probability axioms, the definition of conditional probability, and the assumption that the set of B_j's is a partition. So if the agent *can* determine the simpler conditional degrees of belief of (1), and if those beliefs are accurate summaries of his information about the effects of the different possible states B_j and their chances of actually obtaining, then he can rely on the simpler formula (1) to evaluate the acts. The unusual feature of the decision problems discussed above, in which our agent may follow the wrong recommendation, is that (1) and (4) are *not* the correct summaries of the agent's information or beliefs in those situations. This is so because (the agent believes that) his action is correlated with the states B_j, although the action in no way causes any of the states to obtain. In these situations, doing A (not smoking the cigarette) is not a cause but is a symptom of B_j's obtaining (not having the gene), so the degree of belief $P(B_j /A)$ is greater than the belief $P(B_j)$. CEU theory leads the agent to act so as to raise (at significant cost) the epistemic probability of the desirable state (not having the gene) even though the agent believes such action cannot cause the state to occur. CEU theory recommends this because it ignores the information about the causal independence of the action and the state (no dependence from action to state), while attending to the information about the epistemic dependence of one on the other.

The various causal decision theories are designed to correct CEU theory by incorporating the agent's beliefs about the absence of a causal connection from his action to the states in his evaluation of his action. The basic idea shared by all the CDTs is that if the agent believes that the world may be in one of several states (or have one of several structures) whose occurrence are not causally influenced by his action, and which each affect the chances the consequences have of being caused by the action (or which each affect the values of the consequences), then he should evaluate his action this way: For each of the possible states or structures, find the value the action has if that state holds; then find the value of the action by taking a weighted average of these values, using as weights the probabilities (degrees of belief) each state has of being the actual one. The theories differ in their description of the states and structures to be considered and in their analysis of the action's value for each state. The theories of Gibbard & Harper (1976) and Lewis (1981) both recommend that the agent consider various sets of counterfactual conditionals describing possible causal patterns the world might have that are relevant to the actions and consequences in question.[9] Which of these causal patterns obtains is taken to be outside the agent's control, and in both theories the agent is told to weight the values he gives to the possible consequences by his degrees of belief in the competing conjunctions of causal counterfactual conditionals. The appropriate conditionals are *causal* in the sense that in the smoking gene problem, for example, the agent would be expected to assign a substantial degree

[9] See Stalnaker (1972) for the suggestion these theories build on.

of belief to "I smoke □→ I enjoy my cigarette," but a very low degree of belief to "I smoke □→ I get cancer," given that he believes that his smoking does not cause him to have the gene, and that the only causal connection between the smoking and getting cancer is through this genetic factor. The appropriate causal counterfactuals are not "backtracking counterfactuals." Lewis' theory is more elaborate and general than the Gibbard-Harper theory in that he uses counterfactuals with chancy consequents. When his theory is filled in and the counterfactuals explained, it is formally equivalent to Skyrms' *K*-expectation CDT, which is not formulated in terms of counterfactual conditionals. See Lewis (1981) and Gibbard & Harper (1976) for more on their theories. I will concentrate on Skyrms' treatment of the decision problems presented above.

Let us return to the smoking gene example, recalling that the CEU formulas (1) and (4) do not incorporate correct summaries of my beliefs about the situation. The $P(B_j /A)$'s which appear in (4) will reflect my belief that smoking is correlated with having the gene, but not my belief that smoking does not cause me to have it. We might delete the misleading influence that this correlation contributes to the evaluation of A by replacing these conditional probabilities with the simple unconditional degrees of belief $P(B_j)$'s. If we take note of this alteration of our expected utility rule by writing the utility function U, (4) then becomes

$$(5) \qquad U(A) = \Sigma_j \, P(B_j) \, \Sigma_i \, P(C_i /A \,\&\, B_j) \, V(C_i \,\&\, A \,\&\, B_j).$$

(5) will agree with (4) and CEU theory whenever $P(B_j /A) = P(B_j)$ for all B_j; that is, whenever the states of the world which influence the chances or utilities of the consequences are statistically independent of the agent's choice of action. But when the states and the choice are not statistically independent, (4) and (5) will very likely give different values to A, and these differences may lead to different recommendations when the agent chooses the alternative with the highest utility. It is important to note that our justification for suggesting (5) is that the states B_j are believed *causally relevant* to the action A's production of its possible consequences C_i and they are believed *causally independent* of the action (no dependence from action to state). If the B_j's satisfy those conditions, but not otherwise, then we correctly summarize the agent's beliefs by using (5). Skyrms calls partitions which in a given decision situation describe the causally relevant (possible) states of the world which are outside the agent's influence *K-partitions*, and we may remind ourselves of the special features of these states by rewriting (5) as

$$(6) \qquad U(A) = \Sigma_j \, P(K_j) \, \Sigma_i \, P(C_i /A \,\&\, K_j) \, V(C_i \,\&\, A \,\&\, K_j).$$

K-expectation CDT recommends that an agent choose the action which has maximum *K*-expected value, given by (6).[10]

[10] It may be that the agent is unsure about the causal structure of his decision problem; he may have partial belief in a number of hypotheses about it. (Suppose in the smoking example that I am not sure whether the story we gave there is accurate, or whether smoking in fact causally influences my disposition to contract cancer, or whether Tralfamadorians are running cancer experiments from afar on humans selected from the smoking population.) It is important to note that *K*-expectation CDT adequately handles the more complicated and more realistic decision problems in which the agent is less than absolutely certain what the causal structure of the relevant part of the world is. The idea is simple: build his various hypotheses about the causal story into his appropriate *K*-partition. The hypotheses describe states of affairs beyond his influence which are relevant to the outcomes of his action. Each hypothesis will suggest a partition of factors which, according to the hypothesis, are appropriate *K*'s ; the expanded partition whose members are conjunctions of these factors with the hypothesis will be an appropriate *K*-partition. For the details on this, see Skyrms (1979, pp. 136-138). The idea goes back to Savage (1954).

The smoking gene and Newcomb problems again.

To apply (6) to the smoking gene problem, let K_1 be "I have the gene," and let K_2 be "I do not have the gene." Fill in (6) with conditional probabilities $P(C_i / A \& K_j)$ consistent with my beliefs about the correlations between having the gene and the various consequences C_i. Consider the sum of the products $P(C_i / A_s \& K_1)U(C_i \& A_s \& K_1)$, where A_s is the act of smoking, and also the sum of the products $P(C_i / A_r \& K_1)U(C_i \& A_r \& K_1)$, where A_r is the act of refraining. The first sum, the value of smoking while having the gene, is greater than the second, the value of refraining while having the gene. Similarly for the sums for K_2. Smoking and having the gene is better than refraining and having the gene, and smoking and not having the gene is better than refraining and not having it. So $U(A_s)$, the expectation of the values weighted by $P(K_1)$ and $P(K_2)$ rather than $P(K_1 / A_s)$ and $P(K_2 / A_s)$, is more than $U(A_r)$ no matter what my degree of belief in my having the gene happens to be (as long as it is not 0 or 1). So use of (6) leads to the correct recommendation that I smoke away.[11]

In the Newcomb problem, the agent should use K propositions such as

K_1 : The opaque box is empty.
K_2 : The opaque box contains $1,000,000.

and perhaps

K_3 : The opaque box contains something else.

But we will assume that as the problem was stated the degree of belief $P(K_3)$ is near zero, and ignore K_3. Reasoning similar to that for the smoking gene problem will yield the result that $A_2 \& K_1$, taking both boxes when the opaque box is empty, is better than $A_1 \& K_1$, taking only the opaque box when the opaque box is empty. And $A_2 \& K_2$ is better than $A_1 \& K_2$. The expectations of these values with $P(K_1)$ and $P(K_2)$ for the weights are, according to (6), the values $U(A_1)$ and $U(A_2)$. So $U(A_2)$ is greater than $U(A_1)$, no matter what the (nontrivial) degrees

[11] Suppose my degrees of belief are as follows:

	$P(S \& X / K_1 \& A_s) = .8$	$P(S \& X / K_2 \& A_s) = .15$
	$P(S \& \neg X / K_1 \& A_s) = .15$	$P(S \& \neg X / K_2 \& A_s) = .8$
	$P(\neg S \& X / K_1 \& A_s) = .045$	$P(\neg S \& X / K_2 \& A_s) = .005$
	$P(\neg S \& \neg X / K_1 \& A_s) = .005$	$P(\neg S \& \neg X / K_2 \& A_s) = .045$
and	$P(S \& X / K_1 \& A_r) = .045$	$P(S \& X / K_2 \& A_r) = .005$
	$P(S \& \neg X / K_1 \& A_r) = .005$	$P(S \& \neg X / K_2 \& A_r) = .045$
	$P(\neg S \& X / K_1 \& A_r) = .8$	$P(\neg S \& X / K_2 \& A_r) = .15$
	$P(\neg S \& \neg X / K_1 \& A_r) = .15$	$P(\neg S \& \neg X / K_2 \& A_r) = .8$,

which are consistent with those given in note 7. Then, by (6)

$$U(A_s) = P(K_1)[.8(-998) + .15(2) + .045(-1000) + 0] + P(K_2)[.15(-998) + .8(2) + .005(-1000) + 0]$$
$$\approx (-843)\, P(K_1) + (-153)\, P(K_2)$$

$$U(A_r) = P(K_1)[.045(-998) + .005(2) + .8(-1000) + 0] + P(K_2)[.005(-998) + .045(2) + .15(-1000) + 0]$$
$$\approx (-845)P(K_1) + (-155)P(K_2) .$$

So whatever $P(K_1)$ and $P(K_2)$ are—we might fill in any plausible values— $U(A_s)$ is greater than $U(A_r)$.

of belief $P(K_1)$ and $P(K_2)$ are, and the agent gets the correct recommendation to take two boxes.[12]

2. Objections to *K*-expectation CDT.

There has been extensive debate about whether or not there is any need to adopt one of the CDTs in order to make the best choices in problems like the two given above. Sophisticated *V*-maximization theories (see note 4) have been developed, and their defenders have argued that CDT is superfluous. This view is mistaken, but I will not discuss it here. Other criticisms have been made of *K*-expectation CDT in particular which I will consider in this section.

It is easy to see that the key to applying *K*-expectation CDT in the situations where it is most needed is the identification of an appropriate *K*-partition. An agent needs to find a set of propositions $\{K_j\}$ which are such that (a) their use in (6) is correct for the problem confronting him; that is, when they are used in (6) the evaluation of the action incorporates all of the agent's information/beliefs relevant to the problem; and (b) they are practical—the agent needs to have some idea what the values of the degrees of belief and utilities appearing in (6) actually are. Skyrms (1979) has described the *K*-partition appropriate to a given problem as a partition of maximally specific descriptions of the factors outside the agent's influence which are causally relevant to the outcomes of the alternatives available to the agent (or, better, relevant to the outcomes the agent cares about).

Skyrms' *K*-expectation theory has been criticized on the grounds that the appropriate *K*-partitions for decision problems must be carefully selected (which is true), that general use of the theory seems to depend upon a general way of finding an appropriate *K*-partition, and that this task requires understanding on the agent's part of the relations 'state *S* is outside my influence' and 'state *S* is a causally relevant factor to *A*'s having outcome *C*,' understanding which is more subtle than is reasonable to require for a useful decision theory, even a normative one. (Such remarks have been made by Eells (1982). David Lewis has also criticized Skyrms' characterization of appropriate *K*'s, except under a broad interpretation which renders them equivalent to Lewis' dependency hypotheses.) Further, Eells says that if Skyrms' theory lacks a representation theorem, it lacks a theoretical guarantee that it is broadly applicable—a guarantee possessed by other decision theories, including the theory he defends, Jeffrey's *V*-maximization. The axiomatic system and representation theorem I will describe provide a direct response to the

[12] We assume as we did in note 8 that the agent is so sure that there is no cheating, etc. that his conditional degrees of belief such as $P(Z/A_2)$ are nearly 0. (This is merely a simplifying assumption; other values might be supposed. For an agent who isn't sure which of several possible values might apply, use a finer *K*-partition, as described in note 10.) Then this *K*-partition might yield a simple set of conditional degrees of belief like this:

$P(Z/A_1 \& K_1) \approx 1$, and $P(-/A_1 \& K_1) \approx 0$ for other *C*'s
$P(M/A_1 \& K_2) \approx 1$, and $P(-/A_1 \& K_2) \approx 0$ for other *C*'s
$P(T/A_2 \& K_1) \approx 1$, and $P(-/A_2 \& K_1) \approx 0$ for other *C*'s
$P(MT/A_2 \& K_2) \approx 1$, and $P(-/A_2 \& K_2) \approx 0$ for other *C*'s

Now *K*-expectation theory and (6) tell us, assuming again that the $U(C_i)$'s are measured by the money:

$$U(A_1) \approx P(K_1)[0 + 0 + 0 + 0] + P(K_2)[0 + 1{,}000{,}000 + 0 + 0] \approx 1{,}000{,}000\, P(K_2)$$

and

$$U(A_2) \approx P(K_1)[1000 + 0 + 0 + 0] + P(K_2)[0 + 0 + 1{,}001{,}000 + 0] \approx 1000\, P(K_1) + 1{,}001{,}000\, P(K_2).$$

So $U(A_2)$ is greater than $U(A_1)$ no matter what the degrees of belief $P(K_1)$ and $P(K_2)$ are.

latter complaint about the theory. They also answer the former objections concerning the problem of the selection of *K*-partitions.

It is worth pointing out that for an agent with a given decision problem the task of selecting an appropriate *K*-partition is an empirical one. And a *K*-partition is correct for a given decision problem by being correct for the agent's *beliefs* and *preferences* about the problem. The causal decision theories are more successful than their predecessors by better incorporating the agent's beliefs, particularly his causal beliefs, about his problem into his act evaluations—the Newcomb game would be no problem at all for conditional expected utility theory if the agent did not have the beliefs the problem ascribes to him. This is not a point on which the critics of *K*-expectation theory go wrong, but it is worth mentioning because it leads to the idea that we should look to the agent's preference system, which underlies his beliefs and desires, if we want to describe the selection of appropriate *K*-partitions for his decision problems. In what follows, we will be able to state sufficient conditions for the propriety of *K*-partitions in terms of their behavior in the agent's preference system. And it will turn out that these conditions correspond quite well with Skyrms' description given above.

3. Preference and conditional preference.

The entities for which the agent is supposed to have preferences[13] are, first, propositions which may describe acts, states, or consequences, and second, mixtures of propositions. The propositions form a Boolean algebra. The preferences for propositions may be thought of as they are in Jeffrey's preference systems, as preference for news. And this metaphor may be extended to the preferences for mixtures of propositions: think of them as the agent's preferences for lotteries (run by a powerful genie, say), the outcomes of which are news that one of the mixed propositions is true. The mixing coefficients generate the odds on each outcome in the lottery. For example, an agent's preference system may include preferences for propositions R and S, where R is ‘It will rain tomorrow,’ and S is ‘It will be sunny tomorrow,’ which we think of as his preferences for the news that R and for the news that S. He may also have some preference for $.6(R) + .4(\neg R)$, a lottery on two possible bits of news, R, $\neg R$, with odds 3:2 in favor of the outcome, the news that R. The propositions which appear in a mixture need not form a partition; they need be neither incompatible nor exhaustive. An agent's system may include mixtures such as $.25(W) + .75(C)$, where W is ‘The White Sox win the pennant,’ and C is ‘The Cubs win the pennant.’ Again, this is a preference for a lottery whose outcome will be one of two news items, the lottery's odds being 3:1 in favor of the second outcome, the news that C. It should be clear in cases like this one, where the propositions do not form a partition, that the agent's preference for $.25(W) + .75(C)$ is not a preference for the Cubs' having a .75 probability of winning (the Sox and Cubs might both win). A preference for a mixture with a given set of mixing coefficients is not necessarily a preference for the news that the mixed propositions have probabilities corresponding to the mixture's odds, though in some cases, when the propositions do form a partition, it may be natural to think of the preferences this way.[14]

[13] In the following discussion I will sometimes use the phrase, ‘the agent has a preference for x’ to mean simply ‘x appears in the agent's preference ranking,’ rather than ‘x ranks higher than some y in the agent's preference system.’

[14] The use of preferences for mixtures over sets of propositions which do not form partitions is not essential to this theory. A simple modification of the axioms in Section **5** permits mixing to be restricted to partitions. This improves the theory, I believe; details appear in Armendt (forthcoming).

Mixing may be iterated, so we may have mixtures of mixtures, etc., which we may think of as lotteries whose outcomes are tickets to other lotteries (news that other lotteries will be held), and so on. The agent's preference system will be assumed to be very rich; for instance, it is closed under mixing. So for every *n* propositions $P_1,..., P_n$ (in fact, for every *n* propositions or mixtures) and every *n* mixing coefficients $\alpha_1,..., \alpha_n$ (nonnegative reals which sum to 1), the mixture $\Sigma\alpha_i P_i$ is in his preference ranking. The collection of all the propositions and mixtures is a *mixture set*. It is, of course, unrealistic to suppose that the preference system of a human agent is actually so rich. It is also unrealistic to expect actual human preference systems to be so finely graded that, for example, $\alpha P + (1-\alpha)Q$ and $\beta P + (1-\beta)Q$ are distinct elements of the ordering (possibly ranked together but generally not) whenever the real coefficients α and β are distinct, no matter how small $\alpha - \beta$ is. At this point I will only make two general remarks about the adequacy of these and other assumptions as conditions for rational human preference systems. First, the agent's preferences are thought of as dispositional, rather than occurrent. They are preferences the agent would display in a decision situation calling for the relevant choices (*e.g.*, when he is offered lottery tickets or wagers on relevant propositions). Such preferences are ascribed to agents who have never actually entertained them, much as are dispositional subjectivist degrees of belief. Second, even if we keep in mind the dispositional nature of the preferences, the above assumptions about the richness and fineness of the ordering and other assumptions made below are perhaps best interpreted as *embeddability* conditions: It is more plausible to view them as conditions on a structure into which rational human preferences should be embeddable, than as conditions which we should expect all rational preference orderings to satisfy.

The mixture set of preferences described so far is a von Neumann-Morgenstern type of preference system containing preferences for combinations of propositions with 'extraneous scaling probabilities' (the mixing coefficients). The precise description of the structure of the preference system is based on Herstein and Milnor's mixture set axioms for von Neumann-Morgenstern preference systems. Preference structures similar to this one appear in a number of decision theories, although the other theories which follow the von Neumann-Morgenstern approach do not take their preferences to be for *propositions* (some of which describe acts) and mixtures of propositions as this one does. This theory follows Jeffrey's theory in this respect, rather than the more common practice of interpreting preferences as preferences for nonpropositional *acts* (functions from states to consequences, *e.g.*).

In the present theory, the agent's preference system is enriched beyond the mixture set of preferences so far described by the addition of *conditional preferences*. His conditional preferences for *P*,*Q* and for *m*,*Q* (where *m* is a mixture) are understood as the preferences the agent has for *P* and for *m*, under his hypothesis that *Q* is true. These may or may not have the same ranks in his preference ordering as his unconditional preferences for *P* and for *m*. For example, under the hypothesis that it rains here this afternoon (*R*), my preference for swimming this afternoon at the local beach *S*,*R* is very plausibly ranked below my unconditional preference for swimming this afternoon, *S*. But under the hypothesis that it snows today in Tibet (*T*), my preferences *S*,*T* and *S* are very likely ranked together. A conditional preference, we must emphasize, is not a preference for a conditional proposition.[15] And *P*,*Q* is understood as the

[15] I think it is not clear that we very often *have* well-formed preferences for conditionals. But even if we do, it is unlikely that they coincide with corresponding conditional preferences, and I do not have a systematic account of

agent's *present* preference for *P*, under his hypothesis that *Q*, not (necessarily) as the preference he *would have* for *P* if *Q* were true, or if he were to *believe Q* true.

Sometimes a conditional preference may have the same ranking as an unconditional preference for a corresponding conjunction. The examples *S*,*R* and *S*,*T* given above would very likely have the same rankings in my preference ordering as *S* & *R* and *S* & *T* respectively. But this is not generally true; if it were then *P*,*Q* and *Q*,*P* should always be ranked together. Consider the conditional preferences *M*,*H* and *H*,*M* where *M* is 'I have medical insurance,' and *H* is 'I am hospitalized for a serious illness.' My preference for having medical insurance under the hypothesis that I am hospitalized is considerably greater than my preference for being hospitalized, under the hypothesis that I am insured. The present theory will of course permit conditional preferences to be ranked with unconditional preferences for conjunctions, but it does not assume that they are. (Treating conditional preferences as preferences for conjunctions may be adequate if one is only interested in making comparisons between preferences under the same condition. This is Lewis' treatment of conditional preference in (1969) and (1976). This treatment will not do, however, if all well-defined conditional preferences are comparable, as in the present theory.)

We may think of unconditional preferences for *P* or for *m* as preferences for the actual world's being a *P*-world or an *m*-world (a world in which lottery *m* is run). Conditional preferences *P*,*Q* and *m*,*Q* may then be described as the agent's preferences, under his hypothesis that the actual world is a *Q*-world, for its being a *P*-world and for its being an *m*-world. The metaphor used by Savage and Jeffrey, regarding preference as preference-for-news, may sometimes be useful in thinking about conditional preference: the agent's preference for *P*,*Q* is his present preference for the news that *P* is true in the actual world, assuming that the actual world is a *Q*-world. This makes sense, I think, for the conditional preferences used in the examples above. But for many conditional preferences this metaphor will not be useful. Suppose that I and many other contestants have entered two contests which award equal prizes P_1 and P_2, and I regard my chances of winning each contest as equal. It is plausible that I am indifferent between P_1 and P_2, and also between $\neg P_1$ and $\neg P_2$. Consider my conditional preferences P_1,F_1, P_2,F_1, $\neg P_1,F_1$, and $\neg P_2,F_1$, where F_1 is the hypothesis that I am one of two finalists for P_1. Now my conditional preferences for P_1, P_2, $\neg P_1$, and $\neg P_2$, under my hypothesis that I am a finalist for P_1 really are quite close to, if not the same as, my corresponding unconditional preferences. But if these conditional preferences are preferences under this hypothesis for the *news* that P_1, the *news* that P_2, etc. then, although I may be indifferent between

how our preferences for conditionals are structured. Here is an example taken from Lewis (1976) in which conditional preference and preference for truth-functional conditionals diverge:

> "A die will be thrown; I stand to win $2 if the 6 is up, $1 if the 5 is up, and nothing otherwise. I suppose the die to be fair and care only about the money I may win, so my preferences follow the computable expected payoffs. *X* holds iff a 2 or 6 is up, *Y* iff a 1 or 5 or 6 is up, and *Z* iff a 3, 4, 5, or 6 is up. I prefer *X* to *Y* conditionally on *Z*, but I prefer the conditional $Z \supset Y$ to the conditional $Z \supset X$."

(The truth of $Z \supset Y$ is consistent with 1, 2, 5, or 6. The truth of $Z \supset X$ is consistent with 1, 2, or 6. If each outcome is equiprobable, $EV(Z \supset Y) = 3/4$ and $EV(Z \supset X) = 2/3$.) As Lewis says, the question is more complicated if other kinds of conditionals are considered. I think that problems of interpretation of preference for such conditionals become more acute—the cases in which preference for conditionals have a clear interpretation seem fewer than those in which conditional preferences do. *E.g*,.I can understand a preference for going to the ball game, under the hypothesis that the home team wins, more easily than a preference for the conditional, if the home team wins, then I go.

P_1,F_1 and P_2,F_1, I will not be indifferent between $\neg P_1$,F_1 and $\neg P_2$,F_1: I will prefer the news of a loss of P_2 (which is a long shot anyway) to the news of a loss of P_1 (which, under my hypothesis, I had a good chance of winning). This all makes sense, I think, if we keep in mind that to make conditional hypotheses is not to acquire information about the world. But even if this is kept in mind, in cases where the conditional hypothesis provides full or partial information as to the occurrence of the proposition (or mixture) which the preference is for, treating the preference as preference-for-news is at least confusing, and perhaps misleading.[16]

This becomes especially apparent when we consider preferences of the form P,P and $\neg P$,P. My theory regards such conditional preferences as usually well-formed and nontrivial, and they play a significant role in the description of proper use of the K-expectation utility rule. Under his hypothesis that the actual world is a P-world, the agent has preferences for its being a P-world and for its being a ($\neg P$)-world. It is important to recall when considering $\neg P$,P that this is not a preference for $\neg P\&P$ (and certainly not a belief in $\neg P\&P$). One may have a well-defined unconditional preference for a proposition one strongly believes false, and one may have a conditional preference for a proposition assumed false—not just because the proposition may not be assumed *known* false, but also because preference and desire may be directed toward propositions in which one does not believe. I may desire that Pegasus be alive and willing to carry me wherever I please; and I may still desire this, under my hypothesis that Pegasus never lived in the actual world.

It will often turn out that making the hypothesis P (or the hypothesis $\neg P$) does not perturb an agent's preference for P. Consider these examples (read '$x \sim y$' as 'the agent is indifferent between x and y'; and '$x \succ y$' as 'x is preferable to y'):

(E_1) C: Candidate C wins the election.

My preferences C and C,C and C,$\neg C$ are quite possibly ranked equally. $C \sim C,C \sim C,\neg C$.

(E_2) S: I go swimming this afternoon.

Again, I think it plausible that, under the hypothesis that I do swim (or the hypothesis that I do not), my preference for swimming is simply equal to my unconditional preference for swimming, *i.e.*, that $S,S \sim S \sim S,\neg S$.

There may, though, be some room for disagreement here. See note 18 below. But also consider

(E_3) A_2: I take both boxes in the Newcomb game.

Here is a case in which it is highly plausible that A_2,A_2 is ranked below A_2, and A_2,$\neg A_2$ is ranked above A_2. Under the hypothesis that I take two boxes, my preference for taking two boxes is diminished, since worlds in which I take both boxes are worlds where an empty opaque box is likely. And since in worlds in which I do not take two boxes, filled opaque boxes are likely, the hypothesis that I do not take both boxes raises my preference for taking them. $A_2,\neg A_2 \succ A_2 \succ A_2, A_2$.

[16] Notice that the metaphor is in some ways problematic even for unconditional preferences (and so in Jeffrey's CEU theory). For example, if I believe that news of some proposition P could only reach *me* if some other state for which I have a strong preference (but only moderate belief) were true, then my preference for the *news* that P will not be a guide to my real preference for P.

(E_4) In the smoking gene example, S: I smoke the cigarette.

For reasons similar to those in example E_3, $S,\neg S \succ S \succ S,S$.

Now degrees of belief and utilities are measures which in our theory are derivative from the primitive notions of conditional and unconditional preference. But we do have some intuitive grasp of these measures, and it may be used to help us clarify our intuitions about preference. In doing so I attempt to informally explicate preference, without contradicting the official derivation of degrees of belief and utilities from preference. I think the intuition to which these examples appeal is that an unconditional preference is perturbed by a conditional hypothesis to the extent that the hypothesis carries information which makes a difference to the agent's estimate of P's *value* or *utility*. But since to hypothesize is not to acquire news, preference is not perturbed by way of alterations in the *degree of belief* in the information contained in P. So a preference for P,P need not be equal to the preferences T,P or T (where T is the valid proposition). And a preference for $P,\neg P$ need not be a preference for $\perp,P$ or for $\perp$ (where $\perp$ is the contradictory proposition; $\perp$ will *not* appear in this theory's preference rankings, nor will any $\perp,P$'s or $P,\perp$'s). And in ordinary situations such as example E_1 (and perhaps example E_2), the hypothesis that a proposition is true or the hypothesis that it is false does not affect the utility the agent attaches to it. But sometimes the hypothesis that the actual world is a P-world may perturb the value P has in the actual world, since the hypothesis may carry information about states (the agent believes are) correlated with P whose occurrence (he believes) influence the value of P: The hypothesis that I take both boxes carries with it an increased chance of the actual world's being a world in which the opaque box is empty, decreasing the value of taking it.[17] The hypothesis that I smoke the cigarette carries an increased chance of the actual world's being a world in which I have the cancer gene, and therefore an increased chance that the consequence of smoking will be enjoyment plus cancer.[18] In the present theory all of this is perfectly acceptable, and in fact valuable. We will see below that the

[17] It is true that what I am describing as perturbed value under a conditional hypothesis involves influences of the hypothesis on degrees of belief in the states. That is, the agent's degrees of belief do not change, but his evaluation of P under the condition employs his conditional degrees of belief in the states given the condition, rather than his unconditional degrees of belief in the states. See the discussion of our utility rule and the comparison with CEU theory below.

[18] It might also be argued that the hypothesis that I do go swimming today (example E_2) carries with it an increased chance of favorable conditions for swimming (supposing that I believe my swimming is correlated with favorable conditions), and so raises my preference for swimming, contrary to what is said above. This kind of reasoning may be applicable to many cases of conditional preferences A,A where A is a proposition describing a possible action, when such correlation is believed, for whatever reason. This poses no particular problem for the present theory, but it does violate the sufficient conditions given below for the agreement of the present theory with CEU theory. K-expectation decision theory can and should be applied to these cases, I believe. Propositions describing the favorable or unfavorable states are likely to satisfy the conditions for suitable K-partitions given below. Of course, the K-expectation and CEU evaluations may still agree even though the sufficient conditions for agreement are violated. Notice that in the extreme case where the correlation is so strong that performance of A is statistically independent of the consequences C_i's, and where those consequence descriptions are rich and detailed enough so that $A,C_i \sim A \,\&\, C_i$, we can again show agreement between CEU theory and the present theory: By (9), $U(A) = \Sigma_i P(C_i)\, U(A, C_i) = \Sigma_i P(C_i /A)\, U(A \,\&\, C_i)$. Finally, we should notice that this correlation between A and favorable states for A which figures in the suggestion that $A \nsim A,A$ is sensitive to differences in descriptions of actions which are sometimes glossed over: the correlation is less plausible if my deliberation and choice occurs some time before my performing A (and the occurrence of the favorable or unfavorable states), than if my choice and the states occur at the same time.

failure of preference for an act proposition A to be equal to preference for A,A is an indication that use of the K-expectation utility rule, rather than a simpler special case of it, is the appropriate way to evaluate the action A. (End of informal discussion of preference intuitions by use of probability and utility intuitions.)

One further example of conditional preference may be interesting here. It is intended to reinforce the idea that conditional preferences such as A,A and $A,\neg A$ make sense. It is drawn from Gibbard & Harper (1976):

> Consider the story of the man who met death in Damascus. Death looked surprised, but then recovered his ghastly composure and said, "I am coming for you tomorrow". The terrified man that night bought a camel and rode to Aleppo. The next day, death knocked on the door of the room where he was hiding and said, "I have come for you".
>
> "But I thought you would be looking for me in Damascus," said the man.
>
> "Not at all," said death "that is why I was surprised to see you yesterday. I knew that today I was to find you in Aleppo."
>
> Now suppose the man knows the following. Death works from an appointment book which states the time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo.
>
> If the man ascribes himself equal probabilities of going to Aleppo and staying in Damascus, he has equal grounds for thinking that death intends to seek him in Damascus and that death intends to seek him in Aleppo. If, however, he decides to go to Aleppo, he then has strong grounds for expecting that Aleppo is where death already expects him to be, and hence it is rational for him to prefer staying in Damascus. Similarly, deciding to stay in Damascus would give him strong grounds for thinking that he ought to go to Aleppo: once he knows he will stay in Damascus, he can be almost sure that death already expects him in Damascus, and hence that if he had gone to Aleppo, death would have sought him in vain.

If the story is filled in as Gibbard and Harper describe, and if we adopt the (inessential) fiction that going to Aleppo and remaining in Damascus are the man's only alternatives, then it is clear that for him $A,\neg A \succ A \succ A,A$ where A is "I go to Aleppo." And similarly for $\neg A$: $\neg A,A \succ \neg A \succ \neg A,\neg A$.

The above remarks clearly amount to much less than a complete account of conditional preference, but at this point I will proceed with the description of the theory's preference systems and representation theorem. I claim that the assumptions this theory makes about conditional preference are actually not very strong, and that whatever correct account of conditional preferences emerges from the intuitions about them I have described will be consistent with them. One further remark: for convenience in notation, unconditional preferences for P and for m will be formally treated as conditional preferences P,T and m,T where T is the valid proposition.

The examples so far given of conditional preferences have been conditional preferences for propositions, but we assume the agent to have conditional preferences for mixtures or lotteries as well. Conditional hypotheses are always propositions; and under a particular hypothesis H the agent's preferences form a mixture set $\boldsymbol{M_H}$. Note that mixtures are only formed under single conditional hypotheses; the agent is *not* assumed to have preferences like $.6(R,C) + .4(\neg R,W)$. (Luce & Krantz (1971) have such a theory with disjunctions, at least, and Balch & Fishburn (1974) present one with mixtures, but I regard such combinations as intuitively problematic and formally unnecessary.) All the agent's preferences are assumed comparable; the union of all the mixture sets is the set $\boldsymbol{X}$ which is ordered by $\gtrsim$ (interpreted 'is at least as preferable as').

4. Derivation of *K*-expectation CDT.

The axioms for rational preference systems and the representation theorem will be given in section **5** . We will look ahead in this section and show how the results stated there provide a foundation for *K*-expectation CDT. The theorem stated below, Fishburn's 1973 representation theorem, shows that a preference system $<\boldsymbol{X}, \gtrsim>$ which satisfies the axioms to be stated can be represented by a probability measure P and utility function U. U is linear and order-preserving, P is unique, U is unique up to positive linear transformation, and the following utility rule holds for all x in $\boldsymbol{X}$:

(7) $U(x,B \text{ v } C) = P(B\,/B \text{ v } C)\; U(x,B) + P(C\,/B \text{ v } C)\; U(x,C)$

whenever x,B and x,C are in $\boldsymbol{X}$, and B and C are incompatible.[19] And when $B_1,\ldots, B_n$ is a partition of propositions such that x,B_i is in $\boldsymbol{X}$ for each i, iterated application of (7) gives us the rule

(8) $U(x) = U(x,T) = \Sigma_i\, P(B_i)\; U(x, B_i)$.

If x is a proposition A, (8) becomes

(9) $U(A) = \Sigma_i\, P(B_i)\; U(A, B_i)$.

If U is a utility function and P the probability measure which represent my preferences, (9) says that, for example, the utility I attach to swimming this afternoon $U(S)$ is equal to my degree of belief in rain $P(R)$ times the utility of swimming under the hypothesis that it rains, plus $P(\neg R)$ times the utility of S under the hypothesis $\neg R$. In the Newcomb problem, the utility I attach to taking both boxes A_2 is equal to my degree of belief that the opaque box is filled $P(K_2)$ times the utility of A_2 under the hypothesis K_2, plus $P(K_1)$ times $U(A_2, K_1)$. If, as we supposed earlier, these utilities are measured by the money, then (9) tells us that $U(A_2)$ is the weighted average $1{,}000\ P(K_1) + 1001{,}000\ P(K_2)$. This agrees with CDT's evaluation of A_2, since the weights are *not* $P(K_1 / A_2)$ and $P(K_2 / A_2)$ as in CEU theory.

[19] In the precise statement of this rule in section **5**, $P(B\,/B \text{ v } C)$ is written $P_{BvC}\,(B)$. The representation theorem actually finds probability measures P_Q on subsets of $\mathscr{E}$ for every proposition Q in $\mathscr{E}'$. It establishes that these probabilities behave as conditional probabilities; see clause (*iii*) of the theorem. The probability measure P in the text is the measure P_T.

To see that $P_{BvC}\,(B) = P_T\,(B\,/B \text{ v } C)$ when $P_T\,(B \text{ v } C) > 0$, note that $B \Rightarrow B \text{ v } C \Rightarrow T$, so by clause (*iii*), $P_T\,(B) = P_T\,(B \text{ v } C)\, P_{BvC}\,(B)$. So, if $P_T\,(B \text{ v } C) > 0$,

$$P_{BvC}\,(B) = P_T\,(B) \,/\, P_T\,(B \text{ v } C) = P_T\,[B \;\&\; (B \text{ v } C)] \,/\, P_T\,(B \text{ v } C) = P_T\,(B\,/B \text{ v } C).$$

In cases where $B \text{ v } C = T$, the probability weights in (7) are $P(B\,/T)$ and $P(C\,/T)$, or simply $P(B)$ and $P(C)$.

When we compare (9) with the CEU utility rules (1) or (3) it is clear that they agree when for each i, $P(B_i) = P(B_i /A)$ and $U(A,B_i) = U(A \& B_i)$. Since the function U is order-preserving, the latter conditions hold when I am indifferent between A,B_i and $A \& B_i$. So if a) I believe the states B_i are statistically independent of A, and b) my conditional preferences for A under the hypotheses B_i are the same as my unconditional preferences for the conjunctions $A \& B_i$, then evaluation of A using CEU theory will agree with the present theory's evaluation of A.

If A is an act proposition, however, we are likely to want to evaluate it in terms of its possible *consequences*, as in (1) for CEU theory, rather than to use some other partition of states. (Or we want to use a partition of state descriptions which include descriptions of the possible results of the action.) If our descriptions of the possible consequences form a partition $C_1,..., C_n$ then (9) becomes

(10) $U(A) = \Sigma_i P(C_i) U(A, C_i)$.

And if the conditions (a) and (b) given above hold, then (10) will agree with CEU's (1). But in almost any interesting decision problem, of course, condition (a) will not hold—I will believe that A's possible consequences are influenced by whether or not I do A, *i.e.,* that they are *not* statistically independent of A. In such cases, we may yet have agreement between (10) and (1), but their agreement cannot be demonstrated as we did above.

There is another direction we may take, though, in seeking cases where the two rules agree. When I am indifferent between A and A,A then $U(A) = U(A,A)$ since U is order-preserving. So by iterated application of (7), putting A for x and $A \& C_i$ for the disjuncts, we have

(11) $U(A) = U(A,A) = \Sigma_i P(A \& C_i /A) U(A,A \& C_i)$.

This agrees with (1) if for all i, $U(A, A \& C_i) = U(A \& C_i)$, since $P(A \& C_i /A) = P(C_i /A)$. So if $A \sim A,A$ and for all i, $A,(A \& C_i) \sim A \& C_i$, I may use the CEU rule (1) to evaluate A and I will get the correct value for $U(A)$. These conditions are satisfied, I believe, in the ordinary decision situations where we expect CEU theory to give the correct evaluations of the available acts. (And this theory, which agrees with CEU theory then, also gives the correct answers for these situations.) The first says that my hypothesis that I do A does not perturb my preference for A, and it was discussed above. The second says that my preference for A, under my hypothesis that I do A and consequence C_i results, is equal to my unconditional preference for the conjunction of A and result C_i. This simply requires that the C_i's be accurate descriptions of the possible consequences of doing A that I care about. (Actually, since I will be comparing $U(A)$ to $U(\neg A)$ or $U(B)$, etc. I might consider C_i's which describe possible consequences of any of the alternatives.)

The following two sets of sufficient conditions for agreement of CEU theory and this theory have emerged from our discussion above:

(I) If $\{B_1,..., B_n\}$ is a partition of propositions such that for all i, $A,B_i \sim A \& B_i$ and $P(B_i) = P(B_i /A)$, then $U(A) = \Sigma_i P(B_i) U(A,B_i) = \Sigma_i P(B_i /A) U(A \& B_i)$.

(II) If $\{C_1,..., C_n\}$ is a partition of propositions such that for all i, $A,(A \& C_i) \sim A \& C_i$ and if $A \sim A,A$, then $U(A) = U(A,A) = \Sigma_i P_A(A \& C_i) U(A, A \& C_i) = \Sigma_i P(C_i /A) U(A \& C_i)$.

When do (7) and (9) take the form of rule (6), the K-expectation utility rule? Well, they will if they agree with CEU theory, and if CEU theory in turn agrees with K-expectation utility theory (*i.e.*, with (6) for some appropriate K-partition). We discussed the latter agreement in section **1** above; it occurs when the K_j's are believed to be statistically independent of A. So in cases where that is true and where $A \sim A,A$, (9) yields the K-expectation rule which agrees (as it should) with CEU theory.

What about cases in which $A \nsim A,A$? Such cases include, as I have claimed above, the causal counterexamples to CEU theory. I argued in section **1** that in these cases A is correctly evaluated by the K-expectation utility rule when appropriate K-partitions can be found. How does (9) yield the K-expectation rule (6)? If we suppose that the agent is not indifferent between A and A,A, then for most state descriptions B_j we have little reason to suppose he will be indifferent between A,B_j and $A,(A \,\&\, B_j)$. But some state descriptions B_j may align those preferences after all: among the descriptions which do this are descriptions which make up appropriate K-partitions. For example, in the smoking gene story my hypothesis that I smoke the cigarette A_s devalues my preference for A_s. But consider my preferences A_s,K_1 and $A_s,(A_s \,\&\, K_1)$, where K_1 is "I have the gene." My preference for A_s, under my hypothesis that I have the gene, is *not* devalued by the additional hypothesis that I smoke. My preference for A_s was already devalued by the hypothesis that K_1 and (given the story assumed in this example) is not further affected by supposing I light up. Similarly for K_2, "I do not have the gene." Having or not having the gene is what matters in this decision problem, and once I suppose that I do or that I do not, my preference for A_s is fixed, with respect to the additional hypothesis A_s. This is true in general of propositions which are members of adequate K-partitions—the hypothesis that one of them holds in the actual world fixes the value of the act to the extent that whether or not the additional hypothesis that the act is done is added, the agent's preference for it remains the same.[20] (Keep in mind that the adequacy of a K-partition is relative to individual decision contexts.)

So if we return to (9), impose the condition $A,B_j \sim A,(A \,\&\, B_j)$ for each state description B_j, and then, in recognition of the special character of this partition of state descriptions, rewrite them as K_j's, we have

$$(12) \qquad U(A) = \textstyle\sum_j P(K_j)\, U(A, A \,\&\, K_j)\ .$$

This will generate the K-expectation rule (6) if we can find a partition of consequence descriptions which for each K_j satisfies our condition given above on well-selected C_i's: If there is a partition $C_1,\ldots, C_n$ such that $A,(A \,\&\, C_i \,\&\, K_j) \sim A \,\&\, C_i \,\&\, K_j$ for all i,j, then we can use (7) to analyze $U(A, A \,\&\, K_j)$ in terms of the possible consequences:

$$(13) \qquad U(A, A \,\&\, K_j) = \textstyle\sum_i P(C_i / A \,\&\, K_j)\, U(A, A \,\&\, C_i \,\&\, K_j)$$

$$= \textstyle\sum_i P(C_i / A \,\&\, K_j)\, U(A \,\&\, C_i \,\&\, K_j).$$

[20] This behavior of adequate K-descriptions in the preference ordering is quite similar to the behavior of propositions which satisfy Salmon's screening off relation in a belief system. C is said to screen off B from A iff $P(B/C) = P(B/A \,\&\, C) \neq P(B/A)$. And the hypothesis K_j might be said to screen off the influence of the hypothesis A on the agent's preference for A. Salmon (1973) has said that the screening off relation can be generally exploited to distinguish causal influence from symptomatic correlation. I believe that this kind of screening off by the K_j's of influences on preferences often results from the agent's beliefs about causal connections among the states the propositions describe, but I do not think this need always be so.

And so by substitution into (12)

(14) $U(A) = \Sigma_j\, P(K_j)\, \Sigma_i\, P(C_i\, /A\ \&\ K_j)\, U(A\ \&\ C_i\ \&\ K_j)$.

This is the K-expectation rule (6) and shows that when $A \nsim A,A$ our theory endorses that rule for partitions $K_1,..., K_m$ and $C_1,..., C_n$ which satisfy

$A,K_j \sim A,(A\ \&\ K_j)$

and

$A,(A\ \&\ C_i\ \&\ K_j) \sim A\ \&\ C_i\ \&\ K_j$

for all i and j. This is a nice result since these are exactly the conditions which we should expect appropriate K and C descriptions to satisfy. In general, the appropriate K-partitions as described by Skyrms will satisfy these conditions, when accompanied by well-selected consequence partitions. Skyrms describes the appropriate K-propositions as maximally specific descriptions of the factors the agent believes are outside his influence and causally relevant to those outcomes of his available alternatives which matter to him. When such a description K_j is made a conditional hypothesis, it may of course perturb the agent's preference for one of his alternatives A: $A_2,K_2 \succ A_2$ in the Newcomb problem. But once that hypothesis is made, the uncertain states of the world which may influence the outcome of the action (including the states which are correlated with but not caused by the action) are fixed, and the additional hypothesis that A is done should not further perturb the preference for A. The condition $A, K_j \sim A,(A\ \&\ K_j)$ captures the idea behind Skyrms' appropriate K-partitions, and it does so in the desirable way mentioned in section **2**: adequate K-partitions are picked out by reference to the way they behave in the agent's preference system.

5. The axioms for rational preference and the representation theorem.

The structure which is interpreted as the agent's preference ranking is a collection of mixture sets each of whose basic elements are members of a Boolean algebra $\mathscr{E}$ of propositions. Each mixture set corresponds to the agent's preferences under the hypothesis that some one element of $\mathscr{E}$ is true. We start with our set of propositions which includes descriptions of states, acts, and consequences. After deleting the contradictory proposition, we construct a set $\boldsymbol{M}$ of all mixtures or gambles on propositions in $\mathscr{E}'$ ($\mathscr{E}' = \mathscr{E} - \perp$). $\boldsymbol{M}$ contains every proposition in $\mathscr{E}'$, and for any n members of $\boldsymbol{M}$, $m_1,..., m_n$, and any n non-negative real coefficients $\alpha_1,..., \alpha_n$ which sum to 1, $\Sigma\alpha_i m_i$ is a member of $\boldsymbol{M}$. (The n-ary mixtures are obtained by iteration of binary mixing; the axiom stating closure under mixing will only mention binary mixtures.) We interpret these mixtures as the agent's preferences for lotteries on the mixed propositions, where the lotteries' odds are given by the coefficients. $\boldsymbol{M}$ is a large set containing all such lotteries, and $\boldsymbol{M}$ is the set of all the agent's unconditional preferences. (Actually the agent's unconditional preferences will be regarded as the ordered pairs which are elements of $\boldsymbol{M} \times T$, where T is the valid proposition.) The first two axioms describe $\mathscr{E}$ and $\boldsymbol{M}$.[21]

(1) $\mathscr{E}$ is a Boolean algebra of propositions. And $\mathscr{E}' = \mathscr{E} - \perp$.

(2) $\boldsymbol{M}$ is the mixture set formed from $\mathscr{E}'$, *i.e.*, $\mathscr{E}' \subset \boldsymbol{M}$ and for all $m_1, m_2 \in \boldsymbol{M}$ and all $\alpha \in [0,1]$:

[21] See note 14.

(a) Closure under mixing: $\alpha m_1 + (1 - \alpha)m_2 \in \boldsymbol{M}$;

(b) $1(m_1) + 0(m_2) = m_1$;

(c) Commutation. $\alpha m_1 + (1 - \alpha)m_2 = (1 - \alpha)m_2 + \alpha m_1$;

(d) Linearity. For all $\beta \in [0,1]$, $\alpha[\beta m_1 + (1 - \beta)m_2] + (1 - \alpha)m_2 = \alpha\beta m_1 + (1 - \alpha\beta)\, m_2$.

The set $\boldsymbol{X}$ is the set which contains all the agent's preferences, conditional and unconditional. It contains all the elements of $\boldsymbol{M}$ x T together with the elements of $\boldsymbol{M_P}$ for every proposition P in $\mathscr{E}'$. For each P in $\mathscr{E}'$, $\boldsymbol{M_P}$ is interpreted as the set of the agent's preferences for the elements of $\boldsymbol{M}$ under the hypothesis P. It is a (possibly improper) subset of $\boldsymbol{M}$ x P. It is assumed to be a mixture set. We also assume that if m,P_1 and m,P_2 are formed, so is $m,(P_1 \text{ v } P_2)$, for all incompatible P_1 and P_2. $\boldsymbol{X}$ is the union of all the $\boldsymbol{M_P}$'s and is ordered by the relation $\succsim$ (interpreted "is at least as preferable as"). Notice that mixtures of conditional preferences are only allowed when they are formed under the same hypothesis: $\boldsymbol{X}$ contains $[\alpha P + (1 - \alpha)Q],R$ but not $\alpha P,R + (1 - \alpha)Q,S$. Of course, it makes sense to compare any two conditional preferences m_1,P and m_2,Q since $\boldsymbol{X}$ is ordered. Also note that the ordering $\succsim$ of $\boldsymbol{X}$ yields orderings of each mixture set $\boldsymbol{M_P}$. Axioms 3 and 4 describe the $\boldsymbol{M_P}$'s and $\boldsymbol{X}$:

(3) For all $P \in \mathscr{E}'$, $\boldsymbol{M_P} \subset \boldsymbol{M}$ x P is a mixture set (see Axiom 2). Also, $\boldsymbol{M_T} = \boldsymbol{M}$ x T.

(4) The set $\boldsymbol{X}$ is the union of all the $\boldsymbol{M_P}$'s and is ordered by $\succsim$.

We now come to a pair of axioms which are Herstein-Milnor's axioms for mixture set utility functions generalized to $\boldsymbol{X}$ (which is a collection of mixture sets). Axiom 5 is a continuity (and Archimedean) axiom stated in terms of the topological properties of sets of mixing coefficients. In the company of the other axioms, it has the consequence that for any three elements of $\boldsymbol{X}$, x,A, y,A, and z,B such that $x,A \succ z,B \succ y,A$, there is some mixing coefficient α such that the lottery $[\alpha x + (1 - \alpha)y]$, A is indifferent to z,B. Axiom 6 implies a more general principle: if $x,A \sim z,B$ and $y,A \sim w,B$ then for *every* mixing coefficient α, $[\alpha x + (1 - \alpha)y],A \sim [\alpha z + (1 - \alpha)w],B$. The idea is that if I am indifferent between the members of two pairs of things, I am indifferent between equivalent lotteries (same odds) each constructed with one member from each pair.

(5) For all $A, B \in \mathscr{E}'$, all x,A , $y,A \in \boldsymbol{M_A}$, and all $z,B \in \boldsymbol{M_B}$, $\{\alpha\colon (\alpha x + (1 - \alpha)y),A \succsim z,B\}$ and $\{\alpha\colon z,B \succsim (\alpha x + (1 - \alpha)y),A\}$ are closed in the usual topology for $[0,1]$.

(6) For all $A, B \in \mathscr{E}'$, for all x,A , $y,A \in \boldsymbol{M_A}$, and for all z,B , $w,B \in \boldsymbol{M_B}$, if $x,A \sim z,B$ and $y,A \sim w,B$, then $[(1/2)x + (1/2)y],A \sim [(1/2)\mathrm{z} + (1/2)w],B$.

Setting $A = B$, these axioms imply that Herstein-Milnor's axioms are satisfied by the elements of mixture set $\boldsymbol{M_A}$ for every A. So they imply (via the Herstein-Milnor theorem) that for every $\boldsymbol{M_A}$ there exists a real-valued utility function U_A which is linear and order-preserving over $\boldsymbol{M_A}$, and is unique up to positive linear transformation. The U_A's will be used to construct the utility function U over all of $\boldsymbol{X}$.

Herstein-Milnor Theorem (1953)

> For each $P \in \mathscr{E}'$, there is a real-valued function U_P on $\boldsymbol{M_P}$ such that U_P is linear and order-preserving on $\boldsymbol{M_P}$, *i.e.*
>
> (i) $U_P(\sum_i \alpha_i\, m_i,P) = \sum_i \alpha_i U_P(m_i,P)$, for all mixtures $\sum_i \alpha_i\, m_i,P$ such that $m_i,P \in \boldsymbol{M_P}$;
>
> (ii) $m_i,P \succ m_j,P$ iff $U_P(m_i,P) > U_P(m_j,P)$, for all m_i,P and $m_j,P \in \boldsymbol{M_P}$;
>
> and U_P is unique up to positive linear transformation.

Axiom 7 guarantees that whenever A & $B = \perp$ and x,A and x,B exist, then so does $x,(A \text{ v } B)$. Axiom 8 then states a kind of averaging principle. It says, for example, that if going swimming on a sunny holiday is at least as preferable as going swimming on a cloudy holiday, then swimming on a sunny holiday is at least as preferable as swimming on a holiday which may be either cloudy or sunny, and the latter is at least as preferable as swimming on a cloudy holiday.

(7) For all $A, B \in \mathscr{E}'$ such that A & $B = \perp$, and for all $x \in \boldsymbol{M}$, if $x,A \in \boldsymbol{M_A}$ and $x,B \in \boldsymbol{M_B}$, then $x,(A \text{ v } B) \in \boldsymbol{M_{A\text{v}B}}$.

(8) For all x,A and x,B in $\boldsymbol{X}$, if A & $B = \perp$ and $x,A \gtrsim x,B$, then $x,A \gtrsim x,(A \text{ v } B) \gtrsim \text{x},B$.

Axiom 9 is a non-triviality condition. It simply denies that I am indifferent between all members of $\boldsymbol{X}$:

(9) For some $x, y \in \boldsymbol{X}$, $x \succ y$.

Axiom 10 is required to generate the comprehensive utility function U on $\boldsymbol{X}$ from the many U_P's on the $\boldsymbol{M_P}$'s. From the x and y whose existence it asserts, a gamble z on x and y will be found such that $z,A \sim z,B$. This axiom denies that there are two incompatible hypotheses A and B such that the preference for every element of $\boldsymbol{M_A}$ is greater than for the corresponding element of $\boldsymbol{M_B}$, and it is a fairly strong structural condition:

(10) For all $A, B \in \mathscr{E}'$ such that A & $B = \perp$, there exist $x, y \in \boldsymbol{M}$ such that $x,A \succ x,B$ and $y,B \succ y,A$.

Axioms (1)-(10) imply the existence of a utility function U on $\boldsymbol{X}$ which is linear, order-preserving, and unique up to positive linear transformation, and also the existence for each incompatible A and B in $\mathscr{E}'$ of unique non-negative real numbers which sum to 1, $P_{A\text{v}B}(A)$ and $P_{A\text{v}B}(B)$, such that

$$U(x,A \text{ v } B) = P_{A\text{v}B}(A)\ U(x,A) + P_{A\text{v}B}(B)\ U(x,B),$$

for all x,A, $x,B \in \boldsymbol{X}$ (Fishburn, Theorem 3). Axiom 11 is required to guarantee the additivity of the probabilities P_A and the chain condition: $P_C(A) = P_C(B)P\,(A)$, whenever $A \Rightarrow B \Rightarrow C$:

(11) For all $A, B, C \in \mathscr{E}'$ which are pairwise incompatible, if there is an $x \in \boldsymbol{M}$ such that $x,A \sim x,B$, then there is a $y \in \boldsymbol{M}$ such that exactly two of y,A, y,B, and y,C are indifferent.

For example, if I am considering having lunch with a friend who is to pick the location from among restaurants ***a***, ***b***, and ***c***, and if I am indifferent between lunch at restaurant ***a*** (meeting my friend for lunch under the hypothesis that he picks ***a***) and lunch at restaurant ***b***, then either I am *not* indifferent between lunch at restaurant ***a*** (or ***b***) and lunch at restaurant ***c***, or if I am, there is some *other* group of preferences I have, perhaps those for walking to lunch at restaurant **a** (walking to meet my friend for lunch under the hypothesis that he picks ***a***), walking to lunch at restaurant ***b***, and walking to lunch at restaurant ***c***, which are such that I am indifferent between exactly two of them. This axiom may at first seem implausible, but it seems less restrictive if we realize that we might use one of our hypotheses to find the required *y*. For example, if I am indifferent between lunch at any of the three restaurants, it is plausible that I am indifferent between *y*,(my friend picks ***a***) and *y*,(my friend picks ***b***) when *y* is 'meet my friend for lunch and pay his check the next time we eat at restaurant ***c***,' but that I am *not* indifferent between *y* if he picks ***a*** and *y* if he picks ***c***.

Axioms (1)-(11) imply the representation theorem:

THEOREM (Fishburn, 1973). If $\mathscr{E}$, $\boldsymbol{X}$, and $\succsim$ satisfy Axioms (1)-(11) above, then there is a real-valued function U on $\boldsymbol{X}$ and a finitely-additive probability measure P_A on $\{A \ \& \ B\colon B \in \mathscr{E}\}$ for each $A \in \mathscr{E}'$ such that:

(*i*) $x,A \succ y,B$ iff $U(x,A) > U(y,B)$, for all x,A and y,B in $\boldsymbol{X}$;

(*ii*) $U(x,A)$ is linear (as a function on $\boldsymbol{M_A}$) for each A in $\mathscr{E}'$;

(*iii*) $P_C(A) = P_C(\mathrm{B})\, P_B(A)$ whenever $A \Rightarrow B \Rightarrow C$, $A \in \mathscr{E}$, and $B, C \in \mathscr{E}'$;

(*iv*) $U(x,A \text{ v } B) = P_{A\text{v}B}(A)\, U(x,A) + P_{A\text{v}B}(B)\, U(x,B)$ whenever x,A, $x,B \in \boldsymbol{X}$ and $A \ \& \ B = \bot$;

furthermore, the P_A's are unique and U is unique up to positive linear transformation.

Fishburn produces examples which show that axioms (10) and (11) are required for the representation and uniqueness theorems in the sense that deletion of either axiom may lead to failure of some parts of the theorems to hold for some preference systems (*e.g.*, *P* additivity may fail when axiom (11) is deleted, and the uniqueness conditions for *U* and *P* may fail when axiom (10) is deleted). Both axioms place structural conditions (not derivable backwards from the representation theorem) on preference systems, as does the unobjectionable axiom (9). Axioms (4)-(8) are necessary.

Clause (*iii*) of this theorem provides measures P_B even when $P_T(B) = 0$. As note 19 shows, whenever $A \Rightarrow \mathrm{B}$ and $P_T(B) > 0$, $P_B(A) = P_T(A \,/B)$. Clause (*iv*) of the theorem is the general form of our theory's rule (7) relating the utilities of a proposition or mixture under different conditional hypotheses.

K-EXPECTATION UTILITY RULE. If for a given act proposition *A* there are finite sets of propositions $\{K_1,\ldots, K_m\}$ and $\{C_1,\ldots, C_n\}$ such that:

(1) $A \ \& \ K_j$ and $\neg A \ \& \ K_j$ are in $\mathscr{E}'$ for each K_j; and for each K_j and C_i either $A,(A \ \& \ K_j \ \& \ C_i) \in \boldsymbol{X}$ or $(A \ \& \ K_j \ \& \ C_i) = \bot$;

(2) $\{K_1,\ldots, K_m\}$ and $\{C_1,\ldots, C_n\}$ are both exhaustive sets of pairwise incompatible propositions;

(3) $A,K_j \sim A,(A \ \& \ K_j)$, for all K_j;

(4) $A,(A \ \& \ K_j \ \& \ C_i) \sim (A \ \& \ K_j \ \& \ C_i)$, for all K_j, C_i ;

Then if $P(A \,\&\, K_j) > 0$, $U(A) = \Sigma_j \Sigma_i P(K_j) P(C_i / A \,\&\, K_j) U(A \,\&\, K_j \,\&\, C_i)$.

Proof. By clause (*iv*) of the representation theorem,

$$U(A, A \,\&\, K_j) = \Sigma_i P_{A\&Kj}(A \,\&\, K_j \,\&\, C_i)\, U(A, A \,\&\, K_j \,\&\, C_i) .$$

(Unless there is a k such that $(A \,\&\, K_j \,\&\, C_k) = \bot$, in which case the above summation and those below should be for $i \neq k$; see comment (2) below.) So by clause (*iii*), if $P_T(A \,\&\, K_j) > 0$,

$$U(A, A \,\&\, K_j) = \Sigma_i P_T(C_i / A \,\&\, K_j)\, U(A, A \,\&\, K_j \,\&\, C_i).$$

And by condition (4), since U is order-preserving, $U(A, A \,\&\, K_j \,\&\, C_i) = U(A \,\&\, K_j \,\&\, C_i)$.

So $\quad U(A, A \,\&\, K_j) = \Sigma_i P(C_i / A \,\&\, K_j)\, U(A \,\&\, K_j \,\&\, C_i)$.

Finally, since (*iv*) also gives us $U(A) = \Sigma_j P(K_j)U(A, K_j)$, we combine this with the fact (from condition 3, and U's order preservation) that $U(A, K_j) = U(A, A \,\&\, K_j)$ and with the above analysis of $U(A, A \,\&\, K_j)$ to get

$$U(A) = \Sigma_j \Sigma_i P(K_j) P(C_i / A \,\&\, K_j)\, U(A \,\&\, K_j \,\&\, C_i).$$

Comments

(1) The probability measure without subscripts P is P_T.

(2) The reason why $(A \,\&\, K_j \,\&\, C_i) = \bot$ is allowed in condition (1), and why the modification of the K-expectation rule mentioned in the parenthetical remark in necessary is that it would be natural to seek sets of K_j's and C_i's which are useful for evaluating A's competitors ($\neg A$ or B or D...). To do so, we would want conditions (1)-(4) to hold when $\neg A$ or B or ... is substituted for A. It might be that the most useful sets of K_j's and C_i's would be such that not every conjunction of act, state, and consequence is consistent.

(3) Condition (3) above could have instead been the equivalent condition that $A,(A \,\&\, K_j) \sim A,(\neg A \,\&\, K_j)$: By clause (*iv*),

$$U(A, K_j) = P_{Kj}(A \,\&\, K_j)\, U(A, A \,\&\, K_j) + P_{Kj}(\neg A \,\&\, K_j)\, U(A, \neg A \,\&\, K_j).$$

So if $A,K_j \sim A,(A \,\&\, K_j)$, then

$$[1 - P_{Kj}(A \,\&\, K_j)]\, U(A, A \,\&\, K_j) = P_{Kj}(\neg A \,\&\, K_j)\, U(A,\neg A \,\&\, K_j).$$

$$P_{Kj}(\neg A \,\&\, K_j)\, U(A, A \,\&\, K_j) = P_{Kj}(\neg A \,\&\, K_j)\, U(A, \neg A \,\&\, K_j).$$

So $U(A, A \,\&\, K_j) = U(A, \neg A \,\&\, K_j)$, if $P_{Kj}(\neg A \,\&\, K_j) > 0$.

And for the converse, if $A,(A \,\&\, K_j) \sim A,(\neg A \,\&\, K_j)$, then by substitution into the above line from clause (*iv*),

$$U(A, K_j) = [P_{Kj}(A \,\&\, K_j) + P_{Kj}(\neg A \,\&\, K_j)]\, U(A, A \,\&\, K_j) = 1 \cdot U(A, A \,\&\, K_j) = U(A, A \,\&\, K_j).$$

(4) Condition (4) is a requirement that the value attached to act A is entirely carried by the conjunctions $A \,\&\, K_j \,\&\, C_i$; the C_i's are therefore assumed to be fully enough specified to make this so. We might assume that the C_i's are so fully specified that they carry *all* the value of A (so that $A,(A \,\&\, K_j \,\&\, C_i) \sim A \,\&\, K_j \,\&\, C_i \sim C_i$), which simplifies the utility rule in

that $U(C_i)$ may be used rather than $U(A \;\&\; K_j \;\&\; C_i)$. This will work if *A* has no inherent value, or if *A* has inherent value and a description of which act is done is included in each C_i, but in the latter case most of the conjunctions $A \;\&\; K_j \;\&\; C_i$ will be inconsistent.

6. Summary.

I have argued that CDT is a good rational decision theory; its use leads to the correct recommendations in problems where V-maximization goes wrong. Objections to *K*-expectation CDT were mentioned in section **2**, and I claimed that they would be answered by the foundation presented above. One of the objections was simply that CDT lacks the foundation that a representation theorem provides; this objection has now been met. Of course, we really seek a *good* foundation. If we are convinced that the theorem does follow from the axioms, that the uniqueness portion of the theorem is as strong as we would like it to be, and that the utility rule (7) is correct, the relevant question is: Are the conditions placed on rational preference systems by the axioms understandable, plausible, and not overly restrictive? The answer is *yes*, qualified by the acknowledgement that there are no doubt improvements to be found. Full justification of that answer requires a general discussion of formal, idealized treatments of rational preference and representation theorems for rational decision theory. Such a discussion appears in Armendt (1983) and Armendt (forthcoming). It also requires careful assessment of Axioms (1)-(11) given above; this appears in Armendt (1983).

The other objections to *K*-expectation CDT mentioned in section **2** were directed toward the problem of the selection of appropriate *K*-partitions. These objections have been answered by the statement of the sufficient conditions for appropriate *K*'s given in sections **4** and **5**. (Notice that a statement of interesting necessary conditions is likely to be difficult; for a particular decision problem a partition of state descriptions might by coincidence yield an accurate evaluation of action *A* when used in rule (6) even though the partition fails to satisfy any intuitively correct conditions for appropriate *K*'s.) Those conditions describe, as they ought, the behavior of the *K* propositions in the agent's preferences. And the conditions are simple enough, clear enough, that partitions which satisfy them are readily found in decision problems which require their use.

References

Armendt, B.: 1983, *Rational Decision Theory: the Foundations of Causal Decision Theory*, Ph.D. dissertation for the Department of Philosophy, University of Illinois at Chicago.

Armendt, B.: forthcoming, ‘Conditional preference and causal expected utility’, to appear in proceedings of the Conference on Probability and Causation held at the University of California, Irvine in July 1985.

Balch, M. and P. Fishburn: 1974, ‘Subjective expected utility for conditional primitives’, in *Essays on Economic Behavior under Uncertainty*, M. Balch, D. McFadden, S. Wu (eds.), North-Holland, Amsterdam.

Eells, E.: 1982, *Rational Decision and Causality*, Cambridge University Press, Cambridge.

Fishburn, P.: 1973, ‘A mixture-set axiomatization of conditional subjective expected utility’, *Econometrica* **41**, 1-25.

Fishburn, P.: 1974, ‘On the foundations of decision making under uncertainty’, in Balch, McFadden, and Wu (eds.), *op.cit.*

Gibbard, A. and W. Harper: 1976, ‘Counterfactuals and two kinds of expected utility’, in *Ifs*, Harper, Stalnaker, and Pearce (eds.), Reidel, Dordrecht.

Herstein, I. and J. Milnor: 1953, ‘An axiomatic approach to measurable utility’, *Econometrica* **21**, 291-297.

Jeffrey, R.: 1965, *The Logic of Decision*, Wiley, New York.

Jeffrey, R.: 1981, ‘The logic of decision defended’, *Synthese* **48**, 473-92.

Jeffrey, R.: 1983, *The Logic of Decision,* 2nd edition, University of Chicago Press, Chicago.

Lewis, D.: 1969, *Convention: A Philosophical Study*, Harvard University Press, Cambridge.

Lewis, D.: 1976, ‘Convention: a reply to Jamieson’, *Canadian Journal of Philosophy* **6**, 113-120.

Lewis, D.: 1981, ‘Causal decision theory’, *Australasian Journal of Philosophy* **59**, 5-30.

Luce, R.D. and D. Krantz: 1971, ‘Conditional expected utility’, *Econometrica* **39**, 253-271.

Salmon, W.: 1973, ‘Reply to Lehman,’ *Philosophy of Science* **40**, 397-402.

Savage, Leonard J.: 1954, *The Foundations of Statistics*, Wiley, New York.

Skyrms, B.: 1979, *Causal Necessity*, Yale University Press, New Haven.

Skyrms, B.: 1982, ‘Causal decision theory,’ *Journal of Philosophy* **79**, 695-711.

Sobel, J. Howard: 1978, *Probability, Chance, and Choice: A Theory of Rational Agency*, unpublished.

Stalnaker, R.: 1972, ‘Letter to David Lewis,’ in *Ifs*, Harper, Stalnaker, Pearce (eds.), Reidel, Dordrecht.

von Neumann, J. and O. Morgenstern: 1947, *Theory of Games and Economic Behavior,* 2nd edition, Princeton University Press, Princeton.