Gibbard, Allan; Harper, William L

**Working Paper**

# Counterfactuals and Two Kinds of Expected Utility

Discussion Paper, No. 194

**Provided in Cooperation with:**
Kellogg School of Management - Center for Mathematical Studies in Economics and Management Science, Northwestern University

*Suggested Citation:* Gibbard, Allan; Harper, William L (1976) : Counterfactuals and Two Kinds of Expected Utility, Discussion Paper, No. 194, Northwestern University, Kellogg School of Management, Center for Mathematical Studies in Economics and Management Science, Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/220553

Discussion Paper No. 194

# COUNTERFACTUALS AND TWO KINDS OF EXPECTED UTILITY

by

ALLAN GIBBARD and WILLIAM L. HARPER

January, 1976*

Departments of Philosophy, University of Pittsburgh and University of Western Ontario respectively

*This is a preliminary draft based on discussions between the authors, and there is no guarantee that both authors agree with everything in it. A much shorter version was presented to the 5th International Congress of Logic, Methodology, and Philosophy of Science, London, Ontario, August 1975. There, and at the earlier University of Western Ontario research colloquium on Foundations and Applications of Decision Theory, we benefited from discussions with many people; in particular we should mention Richard Jeffrey, Isaac Levi, and Barry O'Neill.

# 1. INTRODUCTION

We begin with a rough theory of rational decision-making. In the first place, rational decision-making involves conditional propositions: when a person weighs a major decision, it is rational for him to ask, for each act he considers, what would happen if he performed that act. It is rational, then, for him to consider propositions of the form "If I were to do a, then c would happen." Such a proposition we shall call a counterfactual, and we shall form counterfactuals with a connective '$\Box\!\!\rightarrow$' on this pattern: "If I were to do a, then c would happen" is to be written 'I do a $\Box\!\!\rightarrow$ c happens."

Now ordinarily, of course, a person does not know everything that would happen if he performed a given act. He must resort to probabilities: he must ascribe a probability to each pertinent counterfactual "I do a $\Box\!\!\rightarrow$ c happens." He can then use these probabilities, along with the desirabilities he ascribes to the various things that might happen if he did a given act, to reckon the expected utility of a. If a has possible outcomes $o_1, \ldots, o_n$, the expected utility of a is the weighted sum

$$\Sigma_i \, \underline{\text{prob}}(\text{I do a } \Box\!\!\rightarrow \, o_i \text{ obtains}) \mathcal{D} o_i,$$

where $\mathcal{D} o_i$ is the desirability of $o_i$. On the view we are sketching, then, the probabilities to be used in calculating expected utility are the probabilities of certain counterfactuals.

That is not the story told in familiar Bayesian accounts of rational decision; those accounts make no overt mention of counterfactuals. We shall discuss later how Savage's account (1972) does without counterfactuals; consider first an account given by Jeffrey (1965, pp. 5-6).

> A formal Bayesian decision problem is specified by two rectangular arrays (matrices) of numbers which represent probability and desirability assignments to the act-condition pairs. The columns represent a set of incompatible conditions, an unknown one of which actually obtains. Each row of the desirability matrix,
>
> $$d_1 \quad d_2 \quad \cdots \quad d_n$$
>
> represents the desirabilities that the agent attributes to the $n$ conditions described by the column headings, on the assumption that he is about to perform the act described by the row heading; and the corresponding row of the probability matrix,
>
> $$p_1 \quad p_2 \quad \cdots \quad p_n$$
>
> represents the probabilities that the agent attributes to the same $n$ conditions, still on the assumption that he is about to perform the act described by the row heading. To compute the expected desirability of the act, multiply the corresponding probabilities and desirabilities, and add:
>
> $$p_1d_1 + p_2d_2 + \cdots + p_nd_n$$

On the Bayesian model as presented by Jeffrey, then, the probabilities to be used in calculating "expected desirability" are "probabilities that the agent attributes" to certain conditions "on the assumption that he is about to perform" a given act. These, then, are conditional probabilities; they take the form prob(S/A), where A is the proposition that the agent is about to perform a given act and S is the proposition that a given condition holds.

On the account Jeffrey gives, then, the probabilities to be used in decision problems are not the unconditional probabilities of certain counterfactuals, but are instead certain conditional probabilities. They take the form prob(S / A), whereas on the view we sketched at the outset, they should take the form prob(A $\Box\!\!\rightarrow$ S). Now perhaps, for all we have said so far, the difference between these accounts is merely one of presentation. Perhaps for every appropriate A and S, we have

$$\text{prob}(A \Box\!\!\rightarrow S) = \text{prob}(S / A); \qquad (1)$$

the probability of a counterfactual $A \Box\!\!\rightarrow S$ always equals the corresponding conditional probability. That would be so if (1) is a logical truth. David Lewis, however, has shown (1975) that on certain very weak and plausible assumptions, (1) is not a logical truth: it does not hold in general for arbitrary propositions A and S. That leaves the possibility that (1) holds at least in all decision contexts: that it holds whenever A is an act an agent can perform and prob gives that agent's probability ascriptions at the time.

In Section 3, we shall state a condition that guarantees the truth of (1) in decision contexts. We shall argue, however, that there are decision contexts in which this condition is violated. The context we shall use as an example is patterned after one given by Stalnaker. We shall follow Stalnaker in arguing that in such contexts, (1) indeed fails, and it is probabilities of counterfactuals rather than conditional probabilities that should be used in calculations of expected utility. The rest of the paper takes up the ramifications for decision theory of the two ways of calculating expected utility. In particular, the two opposing answers to Newcomb's problem (Nozick, 1969) are supported respectively by the two kinds of expected utility maximization we are discussing.

We are working in this paper within the Bayesian tradition in decision theory, in that the probabilities we are using are subjective probabilities, and we suppose an agent to ascribe values to all probabilities needed in calculations of expected utilities. It is not our purpose here to defend this general tradition, but rather to work within it, and to consider two divergent ways of developing it.

## 2. COUNTERFACTUALS

What we shall be saying requires little in the way of an elaborate theory of counterfactuals. We do suppose that counterfactuals are genuine propositions. For a proposition to be a counterfactual, we do not require that its antecedant be false: on the view we are considering, a rational agent entertains

counterfactuals of the form "I do a $\Box\!\!\rightarrow$ S" both for the act he will turn out to perform and for acts he will turn out not to perform. To say A $\Box\!\!\rightarrow$ S is not to say that A's holding would **bring about** S's holding: A $\Box\!\!\rightarrow$ S is indeed true if A's holding would bring about S's holding, but A $\Box\!\!\rightarrow$ S is true also if S would hold regardless of whether A held.

These comments by no means constitute a full theory of counterfactuals. In what follows, we shall appeal not to a theory of counterfactuals, but to the reader's intuitions about them—asking the reader to bear clearly in mind that "I do a $\Box\!\!\rightarrow$ S" is to be read "If I were to do a, then S would hold."

It may nevertheless be useful to sketch a theory that would support what we shall be saying; the theory we sketch here is somewhat like that of Stalnaker and Thomason (Stalnaker, 1968; Stalnaker and Thomason, 1970). Let a be an act which I might decide at time t to perform. An a-world will be a possible world which is like the actual world before t, in which I decide to do a at t and do it, and which obeys physical laws from time t on. Let $W_a$ be the a-world which, at t, is most like the actual world at t. Thus $W_a$ is a possible world which unfolds after t in accordance with physical law, and whose initial conditions at time t are minimally different from conditions in the actual world at t in such a way that 'I do a' is true in $W_a$. The differences in initial conditions should be entirely within the agent's decision-making apparatus. Then "I do a $\Box\!\!\rightarrow$ S" is true iff S is true in $W_a$.

Two axioms that hold on this theory will be useful in later arguments. In the first place, if I actually do a, then the

a-world which, at t, is most like the actual world will be the actual world itself. Thus "I do a $\Box\!\!\rightarrow$ S" will be true **iff** S is true in the actual world.

Axiom 1. $A \supset [(A \Box\!\!\rightarrow S) \equiv S]$.

In the second place, since "I do a $\Box\!\!\rightarrow$ S" is true iff S holds in $W_a$ and "I do a $\Box\!\!\rightarrow \bar{S}$" is true iff $\bar{S}$ holds in $W_a$, we can say the following.

Axiom 2. $(A \Box\!\!\rightarrow S) \not\equiv (A \Box\!\!\rightarrow \bar{S})$.

We shall invoke both these axioms, not because they follow from the rough theory we have given, and not because we regard them as intuitively self-evident, but because they simplify matters. The rationales just given for these two axioms are shaky. The rationale we gave for Axiom 1 fails if the world is indeterministic, so that a physically possible world exactly like the actual world at time t could later diverge from the actual world. The rationale we gave for Axiom 2 fails if there is no unique a-world which is closest at t to the actual world. (See Lewis, 1973, pp. 79-83). Circumstances in which these axioms may fail, however, involve complications which it would be best to ignore in preliminary work. If the axioms fail, then what we are saying should be regarded as an approximation which ignores these complications: in accepting Axiom 1, we ignore complications that would arise from physical indeterminacy, and in accepting Axiom 2, we ignore states of affairs that neither definitely would hold nor definitely would fail to hold if an act were performed.

Our appeals to these Axioms will be rare and explicit. For the most part in treating counterfactuals we shall simply depend on a normal understanding of the way counterfactuals apply to the situations we discuss.

## 3. TWO KINDS OF EXPECTED UTILITY

We have spoken on the one hand of expected utility calculated from the probabilities of counterfactuals, and on the other hand of expected utility calculated from conditional probabilities. In what follows, we shall not distinguish between an act an agent can perform and the proposition that says that he is about to perform it; acts will be expressed by capital letters early in the alphabet. An outcome O of an act is a proposition which may, for all the agent knows, express all the consequences of an act which he cares about. The agent, we shall assume, ascribes a magnitude $\mathcal{D}O$ to each outcome O. He knows that if he performed the act, one and only one of its outcomes would obtain, although he does not ordinarily know which of its outcomes that would be.

Let $O_1, \ldots O_m$ be the outcomes of act A. The expected utility of A calculated from probabilities of counterfactuals we shall call $\mathcal{U}(A)$; it is given by the formula

$$\mathcal{U}(A) = \Sigma_j \, \underline{prob}(A \mathrel{\square\!\!\rightarrow} O_j) \, \mathcal{D}O_j .$$

The expected utility of A calculated from conditional probabilities we shall call $\mathcal{V}(A)$; it is given by the formula

$$\mathcal{V}(A) = \Sigma_j \, \underline{prob}(O_j/A) \, \mathcal{D}O_j .$$

Perhaps the best mnemonic for distinguishing $\mathcal{U}$ from $\mathcal{V}$ is this:

we shall be advocating the use of counterfactuals in calculating expected utility, and we shall claim that $\mathcal{U}(A)$ is the genuine expected utility of A. $\mathcal{V}(A)$, we shall claim, measures instead the welcomeness of the news that one is about to perform A. Remember $\mathcal{V}(A)$, then, as the value of A as news, and remember $\mathcal{U}(A)$ as what the authors regard as the genuine expected utility of A.

Now clearly $\mathcal{U}(A)$ and $\mathcal{V}(A)$ will be the same if

$$\underline{\text{prob}}\ (A \mathbin{\square\!\!\rightarrow} O_j) = \underline{\text{prob}}\ (O_j/A) \qquad (2)$$

for each outcome $O_j$. Unless (2) holds for every $O_j$ such that $\mathcal{D}O_j \neq 0$, $\mathcal{U}(A)$ and $\mathcal{V}(A)$ will be the same only by coincidence. We know from Lewis's work (1975) that (2) does not hold for all propositions A and $O_j$; can we expect that (2) will hold for the appropriate propositions?

One assumption, together with the logical truth of Axiom 1, will guarantee that (2) holds for an act and its outcomes. Here and throughout, we suppose that the function prob gives the probability ascriptions of an agent who can immediately perform the act in question, and that $\underline{\text{prob}}\ \varphi = 1$ for any logical truth $\varphi$.

Condition 1 on act A and outcome $O_i$. The counterfactual $A \mathbin{\square\!\!\rightarrow} O_i$ is stochastically independent of the act A. That is to say,

$$\underline{\text{prob}}\ (A \mathbin{\square\!\!\rightarrow} O_i/A) = \underline{\text{prob}}\ (A \mathbin{\square\!\!\rightarrow} O_i).$$

Assertion 1. Suppose Axiom 1 is a logical truth. If A and $O_i$ satisfy Condition 1, and $\underline{\text{prob}}\ A > 0$,

$$\underline{\text{prob}}\ (A \mathbin{\square\!\!\rightarrow} O_i) = \underline{\text{prob}}\ (O_i/A).^{1}$$

Proof: Since Axiom 1 is a logical truth, for any propositions P and Q ,

$$\underline{\text{prob}}\ (P \supset [(P \boxdot\!\!\rightarrow Q) \equiv Q]) = 1.$$

Hence if $\underline{\text{prob}}\ P > 0$, then

$$\underline{\text{prob}}\ ([P \boxdot\!\!\rightarrow Q] \equiv Q/P) = 1 \text{ and}$$

$$\underline{\text{prob}}\ (P \boxdot\!\!\rightarrow Q/P) = \underline{\text{prob}}\ (Q/P).$$

From this general truth we have

$$\underline{\text{prob}}\ (A \boxdot\!\!\rightarrow O_i/A) = \underline{\text{prob}}\ (O_i/A),$$

and from this and Condition 1, it follows that

$$\underline{\text{prob}}\ (A \boxdot\!\!\rightarrow O_i) = \underline{\text{prob}}\ (O_i/A).$$

That proves the Assertion.

Condition 1 is that the counterfactuals relevant to decision be stochastically independent of the acts contemplated. Stochastic independence is the same as epistemic independence. For $\underline{\text{prob}}\ (A \boxdot\!\!\rightarrow O_i/A)$ is the probability it would be rational for the agent to ascribe to the counterfactual $A \boxdot\!\!\rightarrow O_i$ on learning A and nothing else—on learning that he was about to perform that act. Thus to say that $\underline{\text{prob}}\ (A \boxdot\!\!\rightarrow O_i/A) = \underline{\text{prob}}\ (A \boxdot\!\!\rightarrow O_i)$ is to say that learning that one was about to perform the act would not change the probability one ascribes to the proposition that if one were to perform the act, outcome $O_i$ would obtain. We shall use the terms 'stochastic independence' and 'epistemic independence' interchangeably.

The two kinds of expected utility $\mathcal{U}$ and $\mathcal{V}$ can also be characterized in a way suggested by Jeffrey's account of the Bayesian model. Let acts $A_1, \ldots, A_m$ be open to the agent. Let states $S_1, \ldots, S_n$ partition the possibilities in the following

sense. For any propositions $S_1$, ..., $S_n$, the truth-function aut ($S_1$, ..., $S_n$) will be their exclusive disjunction: aut ($S_1$, ..., $S_n$) holds in and only in circumstances where exactly one of $S_1$, ..., $S_n$ is true. Let the agent know aut ($S_1$, ..., $S_n$). For each act $A_i$ and state $S_j$, let him know that if he did $A_i$ and $S_j$ obtained, the outcome would be $O_{ij}$. Let him ascribe each outcome $O_{ij}$ a desirability $\mathcal{D}O_{ij}$. This will be a matrix formulation of a decision problem; its defining features are that the agent knows that $S_1$, ..., $S_n$ partition the possibilities, and in each of these states $S_1$, ..., $S_n$, each act open to the agent has a unique outcome. A set $\{S_1, \ldots, S_n\}$ of states which satisfy these conditions will be called the states of a matrix formulation of the decision problem in question.

Both $\mathcal{U}$ and $\mathcal{V}$ can be characterized in terms of a matrix formulation:

$$\mathcal{U}(A_i) = \Sigma_j \text{ prob } (A_i \Box\!\!\rightarrow S_j)\, \mathcal{D}O_{ij};$$

$$\mathcal{V}(A_i) = \Sigma_j \text{ prob } (S_j/A_i)\, \mathcal{D}O_{ij}.$$

If $\mathcal{D}O_{ij}$ can be regarded as the desirability the agent attributes to $S_j$ "on the assumption that" he will do $A_i$, then $\mathcal{V}(A_i)$ is the desirability of $A_i$ as characterized in the account we quoted from Jeffrey.

On the basis of these matrix characterizations of $\mathcal{U}$ and $\mathcal{V}$, we can state another sufficient condition for the $\mathcal{U}$-utility and $\mathcal{V}$-utility of an act to be the same.

Condition 2 on act $A_i$, states $S_1$, ..., $S_n$, and the function

prob. For each $A_i$ and $S_j$,

$$\underline{\text{prob}}\ (A_i \Box\!\!\rightarrow S_j/A_i) = \underline{\text{prob}}\ (A_i \Box\!\!\rightarrow S_j).$$

Assertion 2. Suppose Axiom 1 is a logical truth. If a decision problem satisfies Condition 2 for act $A_i$, then $\mathcal{U}(A_i) = \mathcal{V}(A_i)$. The proof is like that of Assertion 1.

## 4. ACT-DEPENDENT STATES IN THE SAVAGE FRAMEWORK

Savage's representation of decision problems (1972) is roughtly the matrix formulation just discussed. Ignorance is represented as ignorance about which of a number of states of the world obtains. These states are mutually exclusive, and as specific as the problem requires (p. 15). The agent ascribes desirability to "consequences," or what we are calling outcomes. For each act open to the agent, he knows what outcome obtains for each state of the world; if he does not, the problem must be reformulated so that he does. Savage indeed defines an act as a function from states to outcomes.

It is a consequence of the axioms Savage gives that a rational agent is disposed to choose as if he ascribed a numerical desirability to each outcome and a numerical probability to each state, and then acted to maximize expected utility, where the expected utility of an act A is

$$\Sigma_S\ \underline{\text{prob}}\ (S)\ \mathcal{L}O(A,S). \qquad (3)$$

(Here O(A,S) is the outcome of act A in state S). Another consequence of Savage's axioms is the principle of dominance: If

for every state S, the outcome of act A in S is more desirable than the outcome of B in S, then A is preferable to B.

Consider this misuse of the Savage apparatus; it is of a kind discussed by Jeffrey (1965, pp. 8-10).

Case 1. David wants Bathsheba, but since she is the wife of Uriah, he fears that summoning her to him would provoke a revolt. He reasons to himself as follows: "There are two possibilities: R, that there will be a revolt, and $\bar{R}$, that there won't be. The outcomes and their desirabilities are given in Matrix 1, where B is that I take Bathsheba and A is that I abstain from her. Whether or not there is a revolt, I prefer having Bathsheba to not having her, and so taking Bathsheba dominates over abstaining from her.

| | R | $\bar{R}$ |
|---|---|---|
| A | $R\bar{B}(0)$ | $\bar{R}\bar{B}(9)$ |
| B | $RB(1)$ | $\bar{R}B(10)$ |

Matrix 1

This argument is of course fallacious: dominance requires that the states in question be independent of the acts contemplated, whereas taking Bathsheba may provoke revolt. To apply the Savage framework to a decision problem, one must find states of the world which are in some sense act-independent.

We now pursue a suggestion by Jeffrey[2] on how to deal with states that are act-dependent. Construct four new states:

$S_{00}$: There would be no revolt whatever I did.

$S_{01}$: A would not elicit revolt, whereas B would.

$S_{10}$: A would elicit revolt, whereas B would not.

$S_{11}$: There would be a revolt whatever I did.

| | $S_{00}$ | $S_{01}$ | $S_{10}$ | $S_{11}$ |
|---|---|---|---|---|
| A | $\bar{R}\bar{B}(9)$ | $\bar{R}\bar{B}(9)$ | $R\bar{B}(0)$ | $R\bar{B}(0)$ |
| B | $\bar{R}B(10)$ | $RB(1)$ | $\bar{R}B(10)$ | $RB(1)$ |

Matrix 2

These states hold independently of A and B, and so we can work from Matrix 2 without fallacy. Since in Matrix 2 neither row dominates, the decision must be made on the basis of probabilities ascribed to the states $S_{00}, \ldots, S_{11}$.

What should the probabilities of these states be? One possible answer would be this: Each of the four states $S_{00}, \ldots, S_{11}$ can be expressed as a conjunction of counterfactuals. $S_{01}$, for instance, is the proposition $(A \mathbin{\square\!\!\rightarrow} \bar{R}) \,\&\, (B \mathbin{\square\!\!\rightarrow} R)$. The probability of $S_{01}$, then, is simply the probability of this proposition, $\underline{\text{prob}}\ ([A \mathbin{\square\!\!\rightarrow} \bar{R}] \,\&\, [B \mathbin{\square\!\!\rightarrow} R])$.

The expected utility of an act can now be calculated in the standard way given by (3). The expected utility of A, for instance, will be

$$\sum_S \underline{\text{prob}}\ (S)\, \mathcal{O}(A,S), \qquad (4)$$

where the summation is over the new states S, and $O(A,S)$ is the outcome of A in state S.

Does this procedure give the correct expected utility for the act? What it gives as the expected utility of A, we can show, is $\mathcal{U}(A)$ —at least that is what it gives if Axiom 2 is part of the logic of counterfactuals. For (4) expands to

$$\underline{\text{prob}}\ (S_{00})\mathcal{L}O(A,S_{00}) + \underline{\text{prob}}\ (S_{01})\mathcal{L}O(A,S_{01}) + \underline{\text{prob}}\ (S_{10})\mathcal{L}O(A,S_{10}) + \underline{\text{prob}}\ (S_{11})\mathcal{L}O(A,S_{11}).$$

We have

$$O(A,S_{00}) = O(A,S_{01}) = \bar{R}\bar{B};$$

$$O(A,S_{10}) = O(A,S_{11}) = R\bar{B}.$$

Thus since $S_{00}$ and $S_{01}$ are mutually exclusive, (4) becomes

$$\underline{\text{prob}}\ (S_{00 \vee} S_{01})\mathcal{L}\bar{R}\bar{B} + \underline{\text{prob}}\ (S_{10 \vee} S_{11})\mathcal{L}R\bar{B}.$$

Now $S_{00 \vee} S_{01}$ is $([A \mathrel{\square\!\!\rightarrow} \bar{R}] \ \& \ [B \mathrel{\square\!\!\rightarrow} R])_{\vee}([A \mathrel{\square\!\!\rightarrow} \bar{R}] \ \& \ [B \mathrel{\square\!\!\rightarrow} R])$,

and in virtue of the logical truth of Axiom 2, this is $A \mathrel{\square\!\!\rightarrow} \bar{R}$. Similarly, $S_{10 \vee} S_{11}$ is $A \mathrel{\square\!\!\rightarrow} R$. Thus (4) becomes

$$\underline{\text{prob}}\ (A \mathrel{\square\!\!\rightarrow} \bar{R})\mathcal{L}\bar{R}\bar{B} + \underline{\text{prob}}\ (A \mathrel{\square\!\!\rightarrow} R)\ \mathcal{L}R\bar{B},$$

which is $\mathcal{U}(A)$. This proof can of course be generalized.

We have considered one way to construct act-independent states from act-dependent states; it is a way that makes use of counterfactuals. Suppose, though, we want to avoid the use of counterfactuals and rely instead on conditional probabilities. Jeffrey, as we understand him, suggests the following: ascribe to each new, act-independent state the product of the pertinent conditional probabilities. We shall call this probability prob*; thus, for instance,

$$\underline{\text{prob}}^*(S_{01}) = \underline{\text{prob}}\ (\bar{R}/A)\ \underline{\text{prob}}\ (R/B),$$

and corresponding formulas hold for the other new states $S_{00}$, $S_{10}$, and $S_{11}$.

Using prob*, we can again calculate expected utility in the standard way given by (3). The expected utility of A, for instance,

will be

$$\Sigma_S \underline{\text{prob}}^* (S) \mathcal{O}O(A,S), \qquad (5)$$

where again the summation is over the new states $S_{00}$, $S_{01}$, $S_{10}$, and $S_{11}$. Now (5), it can be shown, has the value $\mathcal{V}(J)$. For (5) is the sum of terms

$$\underline{\text{prob}}^*(S_{00})\mathcal{O}O(A,S_{00}) = \underline{\text{prob}}\ (\bar{R}/A)\ \text{prob}\ (\bar{R}/B)\mathcal{O}\bar{R}\bar{B},$$

$$\underline{\text{prob}}^*(S_{01})\mathcal{O}O(A,S_{01}) = \underline{\text{prob}}\ (\bar{R}/A)\ \underline{\text{prob}}\ (R/B)\mathcal{O}\bar{R}\bar{B},$$

$$\underline{\text{prob}}^*(S_{10})\mathcal{O}O(A,S_{10}) = \underline{\text{prob}}\ (R/A)\ \underline{\text{prob}}\ (\bar{R}/B)\mathcal{O}R\bar{B},$$

$$\underline{\text{prob}}^*(S_{11})\mathcal{O}O(A,S_{11}) = \underline{\text{prob}}\ (R/A)\ \underline{\text{prob}}\ (R/B)\mathcal{O}R\bar{B}.$$

Thus (5) equals

$$[\underline{\text{prob}}(\bar{R}/B) + \underline{\text{prob}}\ (R/B)]\ \underline{\text{prob}}(\bar{R}/A)\mathcal{O}\bar{R}\bar{B}$$

$$+ [\underline{\text{prob}}(\bar{R}/B) + \underline{\text{prob}}\ (R/B)]\ \underline{\text{prob}}\ (R/A)\mathcal{O}R\bar{B}$$

$$= \underline{\text{prob}}\ (\bar{R}/A)\mathcal{O}\bar{R}\bar{B} + \underline{\text{prob}}\ (R/A)\mathcal{O}R\bar{B},$$

and this is $\mathcal{V}(A)$.

Where, then, a decision problem is misformulated in the Savage framework with act-dependent states, we now have two ways of reformulating the problem with act-independent states. The first way is to express each act-independent state as a conjunction of counterfactuals. If the expected utility of an act A is then calculated in the standard manner and Axiom 2 holds, the result is $\mathcal{U}(A)$. The second way to reformulate the problem is to ascribe to each new act-independent state the product of the pertinent conditioned probabilities. If the expected utility of an act A is then calculated in the standard manner, the result is $\mathcal{V}(A)$. If Axiom 2 holds, then, the two reformulations yield respectively the two kinds of expected utility we have been discussing.

So far we have given the two reformulations only for an example. Here is the way the two methods of reformulation work in general. Let acts $A_1, \ldots, A_m$ be open to the agent, let states $S_1,\ldots,S_n$ not all be act-independent, and for each $A_i$ and $S_j$, let the outcome of act $A_i$ in $S_j$ be $O_{ij}$. For each possible sequence $T_1,\ldots,T_m$ consisting of states in $\{S_1,\ldots S_n\}$, there will be a new, act-independent state $S(T_1,\ldots,T_m)$. The outcome of an act $A_i$ in the new state $S(T_1,\ldots,T_m)$ will simply be the outcome of $A_i$ in the old state $T_i$. What has been said so far applies to both methods. Now, according to the first method of reformulation, this new state $S(T_1,\ldots,T_m)$ will be

$$(A_1 \boxdot\!\!\rightarrow T_1) \;\&\; \ldots \;\&\; (A_m \boxdot\!\!\rightarrow T_m),$$

and hence, of course, its probability will be the probability of this proposition. According to the second method of reformulation, the probability of new state $S(T_1,\ldots,T_m)$ will be

$$\underline{\text{prob}}\ (T_1/A_1) \times \ldots \times \underline{\text{prob}}\ (T_m/A_m).$$

Once the problem is reformulated, expected utility is to be calculated in the standard way by formula (3).

Are these two ways of reformulating a decision problem equivalent or distinct? They are, of course, equivalent if Axiom 2 holds and $\mathcal{U}(A_i) = \mathcal{V}(A_i)$ for each act $A_i$, since the first method yields $\mathcal{U}(A_i)$ if Axiom 2 holds and the second method yields $\mathcal{V}(A_i)$. We already know that Condition 2 and the logical truth of Axiom 1 guarantee that $\mathcal{U}(A_i) = \mathcal{V}(A_i)$. Therefore, we may conclude that if Condition 2 holds and Axioms 1 and 2 are logical truths, the two reformulations are equivalent. Condition 2, recall, is that the counterfactuals $A_i \boxdot\!\!\rightarrow S_j$ are epistemically

act-independent: that for each of the old, act-dependent states in terms of which the problem is formulated, learning that one is about to perform a given act will not change the probability one ascribes to the proposition that if one were to perform that act, that state would obtain.

The upshot of the discussion is this. For the Savage apparatus to apply to a decision problem, the states of the decision matrix must be independent of the acts. We have considered two ways of dealing with a problem stated in terms of act-dependent states; both ways involve reformulating the problem in terms of new states which are act-independent. Given the logical truth of Axioms 1 and 2, a sufficient condition for the equivalence of the two reformulations is that the counterfactuals $A_i \mathbin{\Box\!\!\rightarrow} S_j$ be epistemically act-independent.

## 5. ACT-DEPENDENT COUNTERFACTUALS

Should we expect Condition 2 to hold? In the case of David, it seems that we should. Suppose David somehow learned that he was about to send for Bathsheba; that would give him no reason to change the probability he ascribes to the proposition "If I were to send for Bathsheba, there would be a revolt." Similarly, if David learned that he was about to abstain from Bathsheba, that would give him no reason to change the probability he ascribes to the proposition "If I were to abstain from Bathsheba, there would be a revolt." In the case of David, it seems, the pertinent counterfactuals are epistemically act-independent, and hence for each act he can perform, the $\mathcal{U}$-utility and the $\mathcal{V}$-utility are the same.

When, however, a common factor is believed to affect both behavior and outcome, Condition 2 may fail, and $\mathcal{U}$-utility may diverge from $\mathcal{V}$-utility. The following case is patterned after an example used by Stalmaker to make the same point.[3]

Case 2. Solomon faces a situation like David's, but he, unlike David, has studied works on psychology and political science which teach him the following: Kings have two basic personality types, charismatic and uncharismatic. A king's degree of charisma depends on his genetic make-up and early childhood experiences, and cannot be changed in adulthood. Now charistmatic kings tend to act justly and uncharismatic kings unjustly. Successful revolts against charismatic kings are rare, whereas successful revolts against uncharismatic kings are frequent. Unjust acts themselves, though, do not cause successful revolts; the reason that uncharismatic kings are prone to successful revolts is that they have a sneaky, ignoble bearing. Solomon does not know whether or not he is charismatic; he does know that it is unjust to send for another man's wife.

Now in this case, Condition 2 fails for states R and $\bar{R}$. The counterfactual $B \mathrel{\square\!\!\rightarrow} R$ is not epistemically independent of B: we have

$$\underline{\text{prob}}\ (B \mathrel{\square\!\!\rightarrow} R/B) > \underline{\text{prob}}\ (B \mathrel{\square\!\!\rightarrow} R).$$

For the conditional probability of anything on B is the probability Solomon would rationally ascribe to it if he learned that B. Since he knows that B's holding would in no way tend to bring about R's holding, he always ascribes the same probability to $B \mathrel{\square\!\!\rightarrow} R$ as to R. Hence both $\underline{\text{prob}}\ (B \mathrel{\square\!\!\rightarrow} R) = \underline{\text{prob}}\ (R)$ and $\underline{\text{prob}}\ (B \mathrel{\square\!\!\rightarrow} R/B) = \underline{\text{prob}}\ (R/B)$. Now if Solomon learned that B, he would have reason

to think that he was uncharismatic, and thus revolt-prone. Hence $\underline{prob}\ (R/B) > \underline{prob}\ (R)$, and therefore

$$\underline{prob}(B \boxdot\!\!\rightarrow R/B) = \underline{prob}(R/B) > \underline{prob}(R) = \underline{prob}\ (B \boxdot\!\!\rightarrow R). \quad (6)$$

Here, then, the counterfactual is not epistemically act-independent.

(6) states also that $\underline{prob}\ (B \boxdot\!\!\rightarrow R) < \underline{prob}\ (R/B)$, so that in this case, the probability of the counterfactual does not equal the corresponding conditional probability. By a similar argument we could show that $\underline{prob}\ (A \boxdot\!\!\rightarrow R) > \underline{prob}\ (R/A)$. Indeed in this case a $\mathcal{U}$-maximizer will choose to send for his neighbor's wife whereas a $\mathcal{V}$-maximizer will choose to abstain from her—although we shall need to stipulate the case in more detail to prove the latter.

Consider first $\mathcal{U}$-maxmization. We have that

$$\mathcal{U}(B) = \underline{prob}(B \boxdot\!\!\rightarrow \bar{R})\mathcal{L}\bar{R}B + \underline{prob}(B \boxdot\!\!\rightarrow R)\mathcal{L}RB;$$

$$\mathcal{U}(A) = \underline{prob}(A \boxdot\!\!\rightarrow \bar{R})\mathcal{L}\bar{R}\bar{B} + \underline{prob}(A \boxdot\!\!\rightarrow R)\mathcal{L}R\bar{B}.$$

We have argued that $\underline{prob}\ (B \boxdot\!\!\rightarrow R) = \underline{prob}\ (R)$. Similarly, $\underline{prob}\ (A \boxdot\!\!\rightarrow R) = \underline{prob}\ (R)$, and so $\underline{prob}(A \boxdot\!\!\rightarrow R) = \underline{prob}(B \boxdot\!\!\rightarrow R)$. Likewise $\underline{prob}\ (A \boxdot\!\!\rightarrow \bar{R}) = \underline{prob}\ (B \boxdot\!\!\rightarrow \bar{R})$.

We know that $\mathcal{L}\bar{R}B > \mathcal{L}\bar{R}\bar{B}$ and $\mathcal{L}RB > \mathcal{L}R\bar{B}$. Therefore $\mathcal{U}(B) > \mathcal{U}(A)$. This is in effect an argument from dominance, as we shall discuss in Section 8.

Now consider $\mathcal{V}$-maximization. Learning that A would give Solomon reason to think he was charismatic and thus not-revolt prone, whereas learning that B would give him reason to think that he was uncharismatic and revolt-prone. Thus $\text{prob}\ (R/B) > \underline{prob}\ (R/A)$. **Suppose the difference between these probabilities** is greater than 1/9, so that where $\underline{prob}\ (R/A) = \alpha$ and $\text{prob}\ (R/B) = \alpha + \epsilon$, we have $\epsilon > 1/9$. From Matrix 1, we have

an act may be welcome or unwelcome: an act may be welcome because its being performed is an indication that the world is in a desired state. Solomon, for instance, would welcome the news that he was about to abstain from his neighbor's wife, but he would welcome it not because he thought just acts anymore likely to have desirable consequences than unjust acts, but because he takes just acts to be a sign of charisma, and he thinks that charisma may bring about a desired outcome.

$\mathcal{U}$-utility, in contrast, is a measure of the expected efficacy of an act in bringing about states of affairs the agent desires; it measures the expected value of the consequences of an act. That can be seen in the case of Solomon. The $\mathcal{U}$-utility of sending for his neighbor's wife is greater than that of abstaining, and that is because he knows that sending for her will bring about a consequence he desires—having the woman—and he knows that it will not bring about any consequences he wishes to avoid: in particular, he knows that it will not bring about a revolt.

What is it for an act to bring about a consequence? Here are two possible answers, both formulated in terms of counterfactuals.

In the first place, roughly following Sobel (1970, p. 400) we may say that act A brings about state S if $A \Box\!\!\rightarrow S$ holds, and for some alternative A* to A, $A^* \Box\!\!\rightarrow S$ does not hold. (An alternative to A is another act open to the agent on the same occasion). Now on this analysis, the $\mathcal{U}$-utility of an act as we have defined it is the sum of the expected value of its consequences plus a term which is the same for all acts open to the agent on the occasion in question; this latter term is the expected value of un-

avoidable outcomes. A state S is unavoidable iff for every act A* open to the agent, $A^* \Box\!\!\rightarrow S$ holds. Thus $A \Box\!\!\rightarrow S$ holds iff S is a consequence of A or S is unavoidable. Hence in particular, for any outcome O,

$$\underline{\text{prob}}\ (A \Box\!\!\rightarrow O) = \underline{\text{prob}}\ (O \text{ is a consequence of } A) + \underline{\text{prob}}\ (O \text{ is unavoidable}),$$

and so we have

$$\mathcal{U}(A) = \Sigma_O\ \underline{\text{prob}}\ (A \Box\!\!\rightarrow O)\ \mathcal{D}O$$
$$= \Sigma_O \underline{\text{prob}}\ (O \text{ is a consequence of } A)\ \mathcal{D}O$$
$$+ \Sigma_O\ \underline{\text{prob}}\ (O \text{ is unavoidable})\ \mathcal{D}O.$$

The first term is the expected value of the consequences of A, and the second term is the same for all acts open to the agent. Therefore on this analysis of the term 'consequence,' $\mathcal{U}$-utility is maximal for the act or acts whose consequences have maximal expected value.

Here is a second possible way of analyzing what it is to be a consequence. When an agent chooses between two acts A and B, what he really needs to know is not what the consequences of A are and what the consequences of B are, but rather what the consequences are of A as opposed to B and vice versa. Thus for purposes of decision-making, we can do without an analysis of the clause 'S is a consequence of A', and analyze instead the clause 'S is a consequence of A as opposed to B.' This we can analyze as

$$(A \Box\!\!\rightarrow S)\ \&\ \sim(B \Box\!\!\rightarrow S).$$

Now on this analysis, $\mathcal{U}(A) > \mathcal{U}(B)$ iff the expected value of the consequences of A as opposed to B exceeds the expected value of the consequences of B

as opposed to A. For for any state S, $A \boxrightarrow S$ holds iff either S is a consequence of A as opposed to B or $(A \boxrightarrow S)$ & $(B \boxrightarrow S)$ holds. Thus

$$\mathcal{U}(A) = \Sigma_0 \underline{\text{prob}}\ (A \boxrightarrow O)\ \mathcal{D}O$$

$$= \Sigma_0 \underline{\text{prob}}\ (O \text{ is a consequence of A as opposed to B})\ \mathcal{D}O$$

$$+ \Sigma_0 \underline{\text{prob}}\ ([A \boxrightarrow O] \ \&\ [B \boxrightarrow O])\ \mathcal{D}O$$

$$\mathcal{U}(B) = \Sigma_0 \underline{\text{prob}}\ (O \text{ is a consequence of B as opposed to A})\ \mathcal{D}O$$

$$+ \Sigma_0 \underline{\text{prob}}\ ([A \boxrightarrow O] \ \&\ [B \boxrightarrow O])\ \mathcal{D}O$$

The second term is the same in both cases, and so $\mathcal{U}(A) > \mathcal{U}(B)$ iff

$$\Sigma_0 \underline{\text{prob}}\ (O \text{ is a consequence of A as opposed to B})\ \mathcal{D}O >$$

$$\Sigma_0 \underline{\text{prob}}\ (O \text{ is a consequence of B as opposed to A})\ \mathcal{D}O.$$

The left side is the expected value of the consequences of A as opposed to B; the right side is the expected value of the consequences of B as opposed to A. Thus for any pair of alternatives, to prefer the one with the higher $\mathcal{U}$-utility is to prefer the one the consequences of which as opposed to the other have the greater expected value.

We can now ask whether $\mathcal{U}$ or $\mathcal{V}$ is more properly called the "utility" of an act. The answer seems clearly to be $\mathcal{U}$. The "utility" of an act should be its expected genuine efficacy in bringing about states of affairs the agent wants, not the degree to which news of the act ought to cheer the agent. Since $\mathcal{U}$-utility is a matter of what the act can be expected to bring about whereas $\mathcal{V}$-utility is a matter of the welcomeness of news, $\mathcal{U}$-utility seems best to capture the notion of utility.

Jeffrey (1965, pp. 73-4) writes, "If the agent is deliberating about performing act A or act B, and if AB is impossible, there

is no effective difference between asking whether he prefers A to B as a news item or as an act, for he makes the news." It should now be clear why it may sometimes be rational for an agent to choose an act B instead of an act A, even though he would welcome the news of A more than that of B. The news of an act may furnish evidence of a state of the world which the act itself is known not to produce. In that case, though the agent indeed makes the news of his act, he does not make all the news his act bespeaks.

## 7. TWO SURE THING PRINCIPLES

Case 3. Upon his accession to the throne, Reoboam wonders whether to announce that he will reign severely or to announce that he will reign leniently. He will be bound by what he announces. He slightly prefers a short severe reign to a short lenient reign, and he slightly prefers a long severe reign to a long lenient reign. He strongly prefers a long reign of any kind to a short reign of any kind. Where L is that he is lenient and D, that he is deposed early, his utilities are as in the Matrix 3.

| | D | $\bar{D}$ |
|---|---|---|
| L | 0 | 80 |
| $\bar{L}$ | 10 | 100 |

Matrix 3

The wise men of the kingdom give him these findings of behavioral science: There is no correlation between a king's severity and the length of his reign. Severity, nevertheless, often causes early deposition. The reason for the lack of cor-

relation between severity and early deposition is that on the one hand, charismatic kings tend to be severe, and on the other hand, lack of charisma tends to elicit revolts. A king's degree of charisma cannot be changed in adulthood. There is at present no indication of whether Reoboam is charismatic or not.

These findings were based on a sample of 100 kings, 48 of whom had their reigns cut short by revolt. On post mortem examination of the pineal gland, 50 were found to have been charismatic and 50 uncharismatic. 80% of the charismatic kings had been severe and 80% of the uncharismatic kings had been lenient. Of the charismatic kings, 40% of those who were severe were deposed whereas only 20% of those who were lenient were deposed. Of the uncharismatic kings, 80% of those who were severe were deposed whereas only 55% of those who were lenient were deposed. The totals were as in Table 1. This is Reoboam's total evidence on the subject.

TABLE 1

| | Charismatic | Uncharismatic | Total |
|---|---|---|---|
| severe | 16 deposed (40%)<br>24 long-reigned | 8 deposed (80%)<br>2 long-reigned | 24 deposed (48%)<br>26 long reigned |
| lenient | 2 deposed (20%)<br>8 long-reigned | 22 deposed (55%)<br>18 long-reigned | 24 deposed (48%)<br>26 long reigned |

Reoboam's older advisors argue from a sure thing principle. There are two possibilities, they say: that Reoboam is charismatic and that he is uncharismatic; what he does now will not affect

his degree of charisma. On the assumption that he is charismatic, it is rational to prefer lenience. For since 40% of severe charismatic kings are deposed, the expected utility of severity in that case would be

$$.4\mathcal{D}SD + .6\mathcal{D}S\bar{D} = .4 \times 10 + .6 \times 80 = 52,$$

whereas since only 20% of lenient charistmatic kings are deposed, the expected utility of lenience in that case would be

$$.2\mathcal{D}LD + .8\mathcal{D}L\bar{D} = .2 \times 0 + .8 \times 10 = 64.$$

On the assumption that he is uncharismatic, it is again rational to prefer lenience. For since 80% of severe uncharismatic kings are deposed, the expected utility of severity in this case would be

$$.8\mathcal{D}SD + .2\mathcal{D}S\bar{D} = .8 \times 10 + .2 \times 100 = 28,$$

whereas since only 55% of lenient uncharismatic kings are deposed, the expected utility of lenience in this case would be

$$.55\mathcal{D}LD + .45\mathcal{D}L\bar{D} = .55 \times 0 + .45 \times 80 = 36.$$

Thus in either case, lenience is to be preferred, and so by a sure thing principle, it is rational to prefer lenience in the actual case.

Reoboam's youthful friends argue that on the contrary, sure thing considerations prescribe severity. Severity is indeed the dominant strategy. There are two possibilities: D, that Reoboam will be deposed, and $\bar{D}$, that he will not be. These two states are stochastically independent of the acts contemplated: both prob (D/S) and prob (D/L) are .48. Therefore, his youthful friends urge, one can without fallacy use the states D and $\bar{D}$

in an argument from dominance. On the assumption that he will be deposed, he prefers to be severe, and likewise on the assumption that he will not be deposed, he prefers to be severe. Thus by dominance, it is rational for him to prefer severity.

Here, then, are two sure thing arguments which lead to contrary prescriptions. One argument appeals to the finding that charisma is causally independent of the acts contemplated; the other appeals to the finding that being deposed is stochastically independent of the acts. The old advisors and youthful companions are in effect appealing to different versions of a sure thing principle, one of which requires causal independence and the other of which requires stochastic independence. The two versions lead to incompatible conclusions.

The sure thing principle is this: if a rational agent knows aut ($S_1$, ..., $S_n$) and prefers A to B in each case, then he prefers A to B. If the propositions $S_1$, ..., $S_n$ are required to be states in a matrix formulation of the decision problem, so that each pair of state and act determine a unique outcome, the sure thing principle becomes the principle of dominance to be discussed in Section 8; the principle of dominance is thus a special case of the sure thing principle. Now the principle of dominance, we have said, requires a proviso that the states in question be act-independent. The sure thing principle should presumably include the same proviso. The sure thing principle, then, should be this: If a rational agent knows that precisely one of the propositions $S_1$, ..., $S_n$ holds and prefers act A to act B in each case, and if in addition the propositions $S_1$, ..., $S_n$

are independent of the acts A and B, then he prefers A to B.

The problem in the case of Reoboam is that his two groups of advisors appeal to different kinds of independence to reach opposing conclusions. The older advisors appeal to causal independence; they cite the finding that a king's degree of charisma is unaffected by his adult actions. His youthful companions appeal to stochastic independence; they cite the finding that there is no correlation between severity in kings and revolt. The two appeals yield opposite conclusions.

It seems, then, that the sure thing principle comes in two different versions, one of which requires that the propositions in question be causally independent of the acts, and the other of which requires the propositions to be stochastically independent of the acts.

The principle to which the youthful companions appeal can be put as follows.

Definition. Act A is sure against act B with stochastic independence of $S_1, \ldots, S_n$ iff the following hold. The agent knows that independently of the choice between A and B, propositions $S_1, \ldots, S_n$ partition the possibilities; that is to say, prob (aut $(S_1,\ldots,S_n)/A$) = 1 and prob (aut $(S_1,\ldots,S_n)/B$) = 1. The propositions $S_1,\ldots,S_n$ are epistemically independent of the choice between A and B, in the sense that for each, prob $(S_i/A)$ = prob $(S_i/B)$. Finally, for each of these propositions $S_i$ it would be rational to prefer A to B if it were known that $S_i$ held.

Sure-thing with Stochastic Independence. If act A is sure against act B with stochastic independence, then it is rational to

prefer A to B.

The principle to which the older advisors appeal will take longer to formulate. The proviso for this version will be that the propositions $S_1, \ldots, S_n$ be causally independent of the choice between A and B; this can be formulated in terms of counterfactuals. To say that a state $S_i$ is causally independent of the choice between A and B is to say that $S_i$ would hold **if** A were performed **iff** $S_i$ would hold if B were performed: $(A \mathbin{\square\!\!\rightarrow} S_i) \equiv (B \mathbin{\square\!\!\rightarrow} S_i)$. We now want to suppose that for each state $S_i$, A would be preferred to B <u>given</u>, in some sense, knowledge of $S_i$. This knowledge of $S_i$ should not simply be knowledge that $S_i$ holds, but knowledge that $S_i$ holds independently of the choice between A and B: that $(A \mathbin{\square\!\!\rightarrow} S_i)$ & $(B \mathbin{\square\!\!\rightarrow} S_i)$. We can now state the principle.

<u>Definition</u>. A <u>is sure against</u> B <u>with causal independence of</u> $S_1, \ldots, S_n$ **iff** the following hold. The agent knows

<u>aut</u> $(A \mathbin{\square\!\!\rightarrow} S_1, \ldots, A \mathbin{\square\!\!\rightarrow} S_n)$,

<u>aut</u> $(B \mathbin{\square\!\!\rightarrow} S_1, \ldots, B \mathbin{\square\!\!\rightarrow} S_n)$.

The states $S_1, \ldots S_n$ are each known to be causally independent of the choice between A and B, in the sense that for each of these $S_i$'s the agent knows that $(A \mathbin{\square\!\!\rightarrow} S_i) \equiv (B \mathbin{\square\!\!\rightarrow} S_i)$. For each $S_i$, it would be rational to prefer A to B if $S_i$ were known to hold independently of the choice between A and B—if, that is, it were known that $(A \mathbin{\square\!\!\rightarrow} S_i)$ & $(B \mathbin{\square\!\!\rightarrow} S_i)$.

<u>Sure-thing with Causal Independence</u>. If A is sure against B with causal independence, then it is rational to prefer A to B.

In the case of Reoboam, we have seen, Sure-thing with Stochastic Independence prescribes severity and Sure-thing with Causal Independence prescribes lenience. Now to us it seems clear that the only rational action in this case is that prescribed by Sure-thing with Causal Independence. It is rational for Reoboam to prefer lenience because severity tends to bring about deposition and he wants not to be deposed much more strongly than he wants to be severe. To be guided by Sure-thing with Stochastic Independence in this case is to ignore the finding that severity tends to bring about revolt—to ignore that finding simply because severity is not on balance a sign that revolt will occur. To choose to be severe is to act in a way that tends to bring about a dreaded consequence, simply because the act is not a sign of the consequence. That seems to us to be irrational.

The two versions of the sure thing principle we have discussed correspond to the two kinds of utility discussed earlier. Sure Thing with Stochastic Independence follows from the principle that an act is rationally preferred to another iff it has greater $\mathcal{V}$-utility, whereas Sure Thing with Causal Independence follows from the principle that an act is rationally preferred to another iff it maximizes $\mathcal{U}$-utility.

Assertion. Suppose that in any possible situation, it is rational to prefer an act A to an act B iff the $\mathcal{U}$-utility of A is greater than that of B. Then Sure Thing with Causal Independence holds.

Proof: Suppose A is sure against B with causal independence of $S_1, \ldots, S_n$, and that in any possible circumstance, it would be rational to prefer A to B iff A's $\mathcal{U}$-utility were greater than B's.

The Assertion will be proved if we show from these assumptions that $\mathcal{U}(A) > \mathcal{U}(B)$.

For each proposition $S_i$, let $S_i^*$ be the proposition

$(A \mathrel{\square\!\!\rightarrow} S_i)$ & $(B \mathrel{\square\!\!\rightarrow} S_i)$.

Now since A is sure against B with causal independence of $S_1, \ldots, S_n$, for each $S_i$ it would be rational to prefer A to B if $S_i^*$ were known to hold. Therefore if $S_i^*$ were known to hold, the $\mathcal{U}$-utility of A would be greater than that of B. Now the $\mathcal{U}$-utility that A would have if $S_i^*$ were known is

$$\Sigma_O \underline{\text{prob}} (A \mathrel{\square\!\!\rightarrow} O/S_i^*) \mathcal{U}O.$$

Call this $\mathcal{U}_i^*(A)$, and define $\mathcal{U}_i^*(B)$ in a like manner. We have supposed that for each $S_i$, $\mathcal{U}_i^*(A) > \mathcal{U}_i^*(B)$.

Now by definition of the function $\mathcal{U}$,

$$\mathcal{U}(A) = \Sigma_O \underline{\text{prob}}(A \mathrel{\square\!\!\rightarrow} O)\, \mathcal{U}O;$$

Since A is sure against B with causal independence of $S_1, \ldots, S_n$, it is known that $\underline{\text{aut}}(A \mathrel{\square\!\!\rightarrow} S_1, \ldots, A \mathrel{\square\!\!\rightarrow} S_n)$, and for each $S_i$ it is known that $(A \mathrel{\square\!\!\rightarrow} S_i) \equiv (B \mathrel{\square\!\!\rightarrow} S_i)$. Hence it is known that $\underline{\text{aut}}$ $(S_1^*, \ldots, S_n^*)$ holds. By the probability calculus, then, for each outcome O

$$\underline{\text{prob}}(A \mathrel{\square\!\!\rightarrow} O) = \Sigma_i \underline{\text{prob}} (A \mathrel{\square\!\!\rightarrow} O/S_i^*) \underline{\text{prob}}\ S_i^*.$$

Therefore

$$\begin{aligned}\mathcal{U}(A) &= \Sigma_O[\Sigma_i \underline{\text{prob}}(A \mathrel{\square\!\!\rightarrow} O/S_i^*) \underline{\text{prob}}\ S_i^*]\, \mathcal{U}O \\ &= \Sigma_i \underline{\text{prob}}\ S_i^*[\Sigma_O \underline{\text{prob}} (A \mathrel{\square\!\!\rightarrow} O/S_i^*)\, \mathcal{U}O], \\ &= \Sigma_i\, \mathcal{U}_i^*(A) \underline{\text{prob}}\ S_i^*.\end{aligned}$$

By a like argument,

$$\mathcal{U}(B) = \Sigma_i\, \mathcal{U}_i^*(B) \underline{\text{prob}}\ S_i^*.$$

Since for each $S_i$, $\mathcal{U}_i^*(A) > \mathcal{U}_i^*(B)$, it follows that $\mathcal{U}(A) > \mathcal{U}(B)$, and the Assertion is proved.

Assertion. Suppose that in any possible circumstance, it is rational to prefer an act A to an act B iff the $\gamma$-utility of A is greater than that of B. Then Sure Thing with Stochastic Independence holds.

Proof. Suppose A is sure against B with causal independence of $S_1, \ldots, S_n$, and that in any possible circumstances, it would be rational to prefer A to B iff A's $\gamma$-utility is greater than B's. The Assertion will be proved if we show from these assumptions that $\gamma(A) > \gamma(B)$.

Now since A is sure against B with stochastic independence of $S_1, \ldots, S_n$, for each $S_i$ it would be rational to prefer A to B if $S_i$ were known to hold. Therefore, if $S_i$ were known to hold, then the $\gamma$-utility of A would be greater than that of B. Now the $\gamma$-utility that A would have if $S_i$ were known to hold is

$$\Sigma_O \underline{\text{prob}}\ (O/AS_i)\ \beta O.$$

Call this $\gamma_i^*(A)$, and define $\gamma_i^*(B)$ correspondingly. We have that for each $S_i$, $\gamma_i^*(A) > \gamma_i^*(B)$.

Now by definition of the function $\gamma$,

$$\gamma(A) = \Sigma_O \underline{\text{prob}}\ (O/A)\ \beta O.$$

Since A is sure against B with stochastic independence of $S_1, \ldots S_n$, we have $\underline{\text{prob}}\ (\underline{\text{aut}}\ (S_1, \ldots, S_n)/A) = 1$, and so by the probability calculus, for each O,

$$\underline{\text{prob}}\ (O/A) = \Sigma_i \underline{\text{prob}}\ (O/AS_i)\ \underline{\text{prob}}\ (S_i/A).$$

Hence

$$\begin{aligned}\gamma(A) &= \Sigma_O[\Sigma_i \underline{\text{prob}}\ (O/AS_i)\ \underline{\text{prob}}\ (S_i/A)]\ \beta O \\ &= \Sigma_i \underline{\text{prob}}\ (S_i/A)[\Sigma_O \underline{\text{prob}}\ (O/AS_i)\ \beta O] \\ &= \Sigma_i \underline{\text{prob}}\ (S_i/A)\ \gamma_i^*(A).\end{aligned}$$

Although these two principles are respective consequences of two principles of expected utility maximization which may conflict, they cannot themselves conflict. For suppose A strongly dominates B with respect to some set of states $S_1$, ... $S_n$. Then the worst outcome of A is more desirable than some outcome of B. For the worst outcome of A is the outcome of A in some state $S_i$, and since A strongly dominates B with respect to $S_1$, ..., $S_n$, the outcome of A in $S_i$ is more desirable than the outcome of B in $S_i$. Thus the worst outcome of A is more desirable than the worst outcome of B. It cannot be the case, then, that B strongly dominates A with respect to some other set of states $T_1$, ..., $T_n$. For if that indeed were the case, then, we have seen, the worst outcome of B would be more desirable than the worst outcome of A. We have seen that if A strongly dominates B with respect to a set of states, then there is no set of states with respect to which B strongly dominates A. For that reason, the two principles of dominance we have stated will never yield conflicting prescriptions for a simple decision problem.

In a weaker form, however, dominance indeed can be exploited to yield conflicting prescriptions.

Definition. Let $S_1$, ..., $S_n$ be the states of a standard decision matrix, and let A and B be acts. A weakly dominates B with respect to $S_1$, ..., $S_n$ iff for each state $S_i$, the outcome of A in $S_i$ is at least as desirable as the outcome of B in $S_i$, and for some state $S_i$ with prob $(S_i) > 0$, the outcome of A in $S_i$ is more desirable than the outcome of B in $S_i$.

We now get two Principles of Weak Dominance by substituting 'weakly dominates' for 'strongly dominates' in the two Principles

Although these two principles are respective consequences of two principles of expected utility maximization which may conflict, they cannot themselves conflict. For suppose A strongly dominates B with respect to some set of states $S_1, \ldots S_n$. Then the worst outcome of A is more desirable than some outcome of B. For the worst outcome of A is the outcome of A in some state $S_i$, and since A strongly dominates B with respect to $S_1, \ldots, S_n$, the outcome of A in $S_i$ is more desirable than the outcome of B in $S_i$. Thus the worst outcome of A is more desirable than the worst outcome of B. It cannot be the case, then, that B strongly dominates A with respect to some other set of states $T_1, \ldots, T_n$. For if that indeed were the case, then, we have seen, the worst outcome of B would be more desirable than the worst outcome of A. We have seen that if A strongly dominates B with respect to a set of states, then there is no set of states with respect to which B strongly dominates A. For that reason, the two principles of dominance we have stated will never yield conflicting prescriptions for a simple decision problem.

In a weaker form, however, dominance indeed can be exploited to yield conflicting prescriptions.

Definition. Let $S_1, \ldots, S_n$ be the states of a standard decision matrix, and let A and B be acts. A weakly dominates B with respect to $S_1, \ldots, S_n$ iff for each state $S_i$, the outcome of A in $S_i$ is at least as desirable as the outcome of B in $S_i$, and for some state $S_i$ with prob $(S_i) > 0$, the outcome of A in $S_i$ is more desirable than the outcome of B in $S_i$.

We now get two Principles of Weak Dominance by substituting 'weakly dominates' for 'strongly dominates' in the two Principles

of Dominance stated above.

Case 4. A subject is presented with two boxes, one to the left and one to the right. He must choose between two acts:

$A_L$ Take the box on the left.

$A_R$ Take the box on the right.

The experimenter has already done one of the following.

$M_{11}$ Place a million dollars in each box.

$M_{01}$ Place a million dollars in the box on the right and nothing in the box on the left.

$M_{00}$ Place nothing in either box.

He has definitely not placed money in the left box without placing money in the right box. Now the experimenter has predicted the behavior of the subject, and before making his prediction, he has used a random device to select one of the following three strategies.

(i) Reward choice of left box: $M_{11}$ if $A_L$ is predicted; $M_{00}$ if $A_R$ is predicted.

(ii) Ensure payment: $M_{11}$ if $A_L$ is predicted; $M_{01}$ if $A_R$ is predicted.

(iii) Ensure non-payment: $M_{01}$ if $A_L$ is predicted; $M_{00}$ if $A_R$ is predicted.

**The subject knows all this, and believes in** the accuracy of the experimenter's predictions with complete certainty.

The Principle of Weak Dominance with Causal Independence prescribes taking the box on the right. The three states $M_{11}$, $M_{01}$, and $M_{00}$ are causally independent of the act the subject performs. The possible outcomes are shown in the table, where

1 is getting the million dollars and 0 is not getting it.

| | $M_{11}$ | $M_{01}$ | $M_{00}$ |
|---|---|---|---|
| $A_L$ | 1 | 0 | 0 |
| $A_R$ | 1 | 1 | 0 |

$M_{01}$ has non-zero probability, since if $A_L$ was predicted it would result from the experimenter's using strategy (iii) and if $A_R$ was predicted, it would result from the experimenter's using strategy (ii). Thus $A_R$ weakly dominates $A_L$ with respect to $M_{11}$, $M_{01}$, $M_{00}$, and the Principle of Weak Dominance with Causal Independence prescribes taking the box on the right.

The Principle of Weak Dominance with Stochastic Independence, in contrast, prescribes taking the box on the left.

The possibilities can be partitioned as follows:

$S_1$ the experimenter predicts correctly and follows strategy (i).

$S_2$ $S_1$ does not hold and the subject wins a million dollars.

$S_3$ $S_1$ does not hold and the subject wins nothing.

The payoffs are given in the table.

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| $A_L$ | 1 | 1 | 0 |
| $A_R$ | 0 | 1 | 0 |

Now <u>prob</u> $(S_1) \neq 0$, and hence $A_L$ weakly dominates $A_R$ with respect

to $S_1$, $S_2$, $S_3$. Moreover, the states $S_1$, $S_2$, and $S_3$ are stochastically independent of $A_L$ and $A_R$. For the subject knows that the experimenter has selected his strategy independently of his prediction, by means of a random device; hence learning that he was about to perform $A_L$, say, would not affect the probability he ascribes to the experimenter's having had any given strategy. By the subject's probability function, then, which strategy the experimenter has used is stochastically independent of the subject's act. Now the subject believes that the experimenter has predicted correctly and used strategy (i), (ii), or (iii). Hence he thinks that $S_1$ holds iff the experimenter has used strategy (i), that $S_2$ holds iff the experimenter has used strategy (ii), and that $S_3$ holds iff the experimenter has used strategy (iii). Hence under his probability function, states $S_1$, $S_2$, and $S_3$ are stochastically independent of $A_L$ and $A_R$. Thus the Principle of Weak Dominance with Stochastic Independence applies, and it prescribes taking the box on the left.

Some readers may object in Case 4 to the subject's complete certainty that the experimenter has predicted correctly. It is possible to construct a conflict between the two principles of weak dominance without requiring such certainty, but the example becomes more complicated.

Case 5. Same as Case 4, except for the following.

The subject ascribes a probability of .8 to the experimenter's having predicted correctly, and this probability is independent of the subject's choice of $A_L$ or $A_R$. Thus where C is "The experimenter has predicted correctly," prob $(C/A_L) = .8$ and

prob $(C/A_R) = .8$.

The experimenter has chosen among the following three strategies by means of a random device.

(i) $M_{11}$ if $A_L$ is predicted; $M_{00}$ if $A_R$ is predicted.

(ii*) $M_{11}$ if $A_L$ is predicted; $M_{01}$ or $M_{00}$, with equal probability, if $A_R$ is predicted.

(iii*) $M_{11}$ or $M_{01}$, with equal probability, if $A_L$ is predicted; $M_{00}$ if $A_R$ is predicted.

He has followed (i) with a probability .5, (ii*) with a probability .25, and (iii*) with a probability .25.

In Case 5, as in Case 4, the states $M_{11}$, $M_{01}$, and $M_{00}$ are causally independent of the acts $A_R$ and $A_L$, and from the Principle of Weak Dominance with Causal Independence and the facts of the case, it follows that it is rational to prefer $A_R$ to $A_L$.

Now let states $S_1$, $S_2$, and $S_3$ be as before: $S_1$ is that the experimenter predicts correctly and follows strategy (i); $S_2$ is that $S_1$ does not hold and the subject receives a million dollars; $S_3$ is that $S_1$ does not hold and the subject receives nothing. As in Case 4, if $S_1$, $S_2$, and $S_3$ are stochastically independent of $A_L$ and $A_R$, then from the Principle of Weak Dominance with Stochastic Independence and the facts of the case, it follows that it is rational to prefer $A_L$ to $A_R$. It is clear that $S_1$ is stochastically independent of the acts $A_L$ and $A_R$; we now show that $S_2$ and $S_3$ are as well: that prob $(S_2/A_L)$ = prob $(S_2/A_R)$ and prob $(S_3/A_L)$ = prob $(S_3/A_R)$.

There are two possible acts, two possible experimenter's predictions, and three possible experimenter's strategies, some of which may involve the flip of a coin. Call a combination of act, prediction, experimenter's strategy, and result of coin flip if it matters, a <u>case</u>. For each case, the Table 2 shows.

(1) The state $M_{11}$, $M_{01}$, or $M_{00}$ which would hold in that case.

(2) The conditional probability of the case given the act.

(3) The outcome in that case: 1 for getting the million dollars, 0 for not.

(4) The state $S_1$, $S_2$, or $S_3$ which holds in that case.

The conditional probability <u>prob</u> $(S_2/A_L)$ is then obtained by adding up the conditional probabilities given $A_L$ of cases in which $S_2$ holds; a like procedure gives <u>prob</u> $(S_3/A_L)$, <u>prob</u> $(S_2/A_R)$, and <u>prob</u> $(S_3/A_R)$.

The conclusion of Table 2 is that the states $S_1$, $S_2$, and $S_3$ are indeed epistemically independent of the acts $A_L$ and $A_R$. Since $A_L$ weakly dominates $A_R$ with respect to states $S_1$, $S_2$, and $S_3$, it follows that $A_L$ weakly dominates $A_R$ with respect to stochastically independent states. We already know that $A_R$ weakly dominates $A_L$ with respect to causally independent states $M_{11}$, $M_{01}$, and $M_{00}$. In Case 5, then, the two principles of weak dominance are in conflict.

Table 2

| | $A_L$ Performed | | $A_R$ Performed | |
|---|---|---|---|---|
| | $A_L$ Predicted .8 | $A_R$ Predicted .2 | $A_L$ Predicted .2 | $A_R$ Predicted .8 |
| Strategy (i) .5 | $M_{11}$ 1 .4 $S_1$ | $M_{00}$ 0 .1 $S_3$ | $M_{11}$ 1 .1 $S_2$ | $M_{00}$ 0 .4 $S_1$ |
| Strategy (ii*) .25 | $M_{11}$ 1 .2 $S_2$ | $M_{01}$ 0 .025 $S_3$<br>$M_{00}$ 0 .025 $S_3$ | $M_{11}$ 1 .05 $S_2$ | $M_{01}$ 1 .1 $S_2$<br>$M_{00}$ 0 .1 $S_3$ |
| Strategy (iii*) .25 | $M_{11}$ 1 .1 $S_2$<br>$M_{01}$ 0 .1 $S_3$ | $M_{00}$ 0 .05 $S_3$ | $M_{11}$ 1 .025 $S_2$<br>$M_{01}$ 1 .025 $S_2$ | $M_{00}$ 0 .2 $S_3$ |
| Totals | $\underline{\text{prob}}\ (S_2/A_L) = .3$<br>$\underline{\text{prob}}\ (S_3/A_L) = .3$ | | $\underline{\text{prob}}\ (S_2/A_R) = .3$<br>$\underline{\text{prob}}\ (S_3/A_R) = .3$ | |

## 9. ACT-INDEPENDENCE IN THE SAVAGE FORMULATION

In Section 4, we said that to apply the Savage framework to a decision problem, one must find states of the world that are in some sense act-independent. In the last section, we distinguished two kinds of independence, causal and epistemic. Which kind is needed in the Savage formulation of decision problems?

The answer is that the Savage formulation has both a $\mathcal{U}$-maximizing interpretation and a $\mathcal{V}$-maximizing interpretation. On the

$\mathcal{U}$-maximizing interpretation, the states must be causally independent of the acts, whereas on the $\mathcal{V}$-maximizing interpretation, the states must be epistemically independent of the acts. That is to say, if the states are causally act-independent, then utility as calculated by the Savage method is $\mathcal{U}$-utility, whereas if the states are epistemically act-independent, then utility as calculated by the Savage method is $\mathcal{V}$-utility. If the states are both causally and epistemically act-independent, then the $\mathcal{U}$-utility of each act equals its $\mathcal{V}$-utility. Thus the Savage formulation itself is not committed to either kind of utility: the kind of utility it yields depends on the way it is applied to decision problems.

The expected utility of an act A in the Savage theory is

$$\sum_S \underline{\text{prob}}\ (S)\ \mathcal{D}O(A,S). \qquad (3)$$

If the states S are all known to be causally independent of A, so that for each state S, the agent knows that $(A \mathrel{\square\!\!\rightarrow} S) \equiv S$, then for each S, we have $\underline{\text{prob}}\ (S) = \underline{\text{prob}}\ (A \mathrel{\square\!\!\rightarrow} S)$. (3) thus becomes

$$\sum_S \underline{\text{prob}}\ (A \mathrel{\square\!\!\rightarrow} S)\ \mathcal{D}O(A,S),$$

and this, we said in Section 3, is $\mathcal{U}(A)$. If, on the other hand, the states S are stochastically independent of A, so that for each S, $\underline{\text{prob}}\ (S) = \underline{\text{prob}}\ (S/A)$, then (3) becomes

$$\sum_S \underline{\text{prob}}\ (S/A)\ \mathcal{D}O(A,S),$$

which is $\mathcal{V}(A)$.

## 10. NEWCOMB'S PROBLEM

The Newcomb paradox discussed by Nozick (1969) has the same

structure as the case of Solomon discussed in Section 3. Nozick treats it as a conflict between the principle of expected utility maximization and the principle of dominance. On the views we have propounded in this paper, the problem is rather a conflict between two kinds of expected utility maximization. The problem is this. There are two boxes, transparent and opaque; the transparent box contains a thousand dollars. The agent can perform $A_1$, taking just the contents of the opaque box, or $A_2$, taking the contents of both boxes. A predictor has already placed a million dollars in the opaque box if he predicted $A_1$ and nothing if he predicted $A_2$. The agent knows all this, and he knows the predictor to be highly reliable in that both prob (he has predicted $A_1/A_1$) and prob (he has predicted $A_2/A_2$) are close to one.

To show how the expected utility calculations work, we must add detail to the specification of the situation. Suppose, somewhat unrealistically, that getting no money has a utility of zero, getting \$1000 a utility of 10, that getting \$1,000,000 has a utility of 100, and that getting \$1,001,000 has a utility of 101. Let M be "there are a million dollars in the opaque box," and suppose prob $(M/A_1) = .9$ and prob $(M/A_2) = .1$. The calculation of $\mathcal{V}(A_1)$ and $\mathcal{V}(A_2)$ is familiar.

$$\mathcal{V}(A_1) = \text{prob}\ (M/A_1)\mathcal{D}\$1{,}000{,}000 + \text{prob}\ (\bar{M}/A_1)\mathcal{D}\$0$$
$$= .9(100) + .1(0) = 90.$$
$$\mathcal{V}(A_2) = \text{prob}\ (M/A_2)\mathcal{D}\$1{,}001{,}000 + \text{prob}\ (\bar{M}/A_2)\mathcal{D}\$1000$$
$$= .1(101) + .9(10) = 19.1.$$

Maximization of $\mathcal{V}$, as is well known, prescribes taking only the contents of the opaque box.[5]

$\mathcal{U}(A_1)$ and $\mathcal{U}(A_2)$ depend on the probability of M, which in turn depends on the probabilities of $A_1$ and $A_2$. For any probability of M, though, we have $\mathcal{U}(A_2) > \mathcal{U}(A_1)$. For let the probability of M be $\mu$; then since M is causally act-independent, $\underline{\text{prob}}\ (A_1 \boxdot\!\!\rightarrow M) =$ and $\underline{\text{prob}}\ (A_2 \boxdot\!\!\rightarrow M) = \mu$. Therefore

$$\mathcal{U}(A_1) = \underline{\text{prob}}\ (A_1 \boxdot\!\!\rightarrow M) \mathcal{D} \$1{,}000{,}000 + \underline{\text{prob}}\ (A_1 \boxdot\!\!\rightarrow \bar{M}) \mathcal{D} \$0$$
$$= 100\mu + 0(1-\mu) = 100\mu.$$
$$\mathcal{U}(A_2) = \underline{\text{prob}}\ (A_2 \boxdot\!\!\rightarrow M) \mathcal{D} \$1{,}001{,}000 + \underline{\text{prob}}\ (A_2 \boxdot\!\!\rightarrow \bar{M}) \mathcal{D} \$1000$$
$$= 101\mu + 10(1-\mu) = 91\mu + 10.$$

Thus $\mathcal{U}(A_2) - \mathcal{U}(A_1) = 10 - 9\mu$, and since $\mu \leq 1$, this is always positive. Therefore whatever probability M may have, $\mathcal{U}(A_2) > \mathcal{U}(A_1)$, and $\mathcal{U}$-maximization prescribes taking both boxes.

To some people, this prescription seems irrational. One possible argument against it takes roughly the form "If you're so rational and you want to be rich, why ain't you?" $\mathcal{V}$-maximizers tend to leave the experiment millionaires whereas $\mathcal{U}$-maximizers do not. Both very much want to be millionaires, and the $\mathcal{V}$-maximizers usually succeed; hence it must be the $\mathcal{V}$-maximizers who are making the rational choice. We take the moral of the paradox to be something else: If someone is very good at predicting behavior and rewards predicted irrationality richly; then irrationality will be richly rewarded.

To see this, consider a variation on Newcomb's story: the subject of the experiment is to take the contents of the opaque box first and learn what it is; he then may choose either to take the thousand dollars in the second box or not to take it. The predictor has an excellent record, and a thoroughly accepted theory to back it up. Most people find nothing in the first box and then

take the contents of the second box. Of the million subjects tested, 1% have found a million dollars in the first box, and strangely enough only 1% of these—100 in 10000—have gone on to take the thousand dollars they could each see in the second box. When those who leave the thousand dollars are later asked why they do so, they say things like "If I were the sort of person who would take the thousand dollars in that situation, I wouldn't be a millionaire."

On both grounds of $\mathcal{U}$-maximization and of $\mathcal{V}$-maximization, these new millionaires have acted irrationally in failing to take the extra thousand dollars. They know for certain that they have the million dollars; therefore the $\mathcal{V}$-utility of taking the thousand as well is 101, whereas the $\mathcal{V}$-utility of not taking it is 100. Even on the view of $\mathcal{V}$-maximizers, then, this experiment will almost always make irrational people and only irrational people millionaires. Everyone knows so at the outset.

Return now to the unmodified Newcomb situation, where the subject must take or pass up the thousand dollars before he sees whether the opaque box is full or empty. What happens if the subject knows not merely that the predictor is highly reliable, but that he is infallible? The argument that the $\mathcal{U}$-utility of taking both boxes exceeds that of taking only one box goes through unchanged. To some people, however, it seems especially apparent in this case that it is rational to take only the opaque box and irrational to take both. For in this case the subject is certain that he will be a millionaire if and only if he takes only the opaque box. If in the case where the predictor is known to be

infallible it is irrational to take both boxes, then, $\mathcal{U}$-maximization is not always the rational policy.

We maintain that $\mathcal{U}$-maximization is rational even in the case where the predictor is known to be infalliable. True, where R is "I become a millionaire," the agent knows in this case that R holds iff $A_1$ holds: he knows the truth-functional proposition $R \equiv A_1$. From this proposition, however, it does not follow that he would be a millionaire if he did $A_1$, or that he would be a non-millionaire if he did $A_2$.

If the subject knows for sure that he will take just the opaque box, then he knows for sure that the million dollars is in the opaque box, and so he knows for sure that he will be a millionaire. But since he knows for sure that the million dollars is already in the opaque box, he knows for sure that even if he were to take both boxes, he would be a millionaire. If, on the other hand, the subject knows for sure that he will take both boxes, then he knows for sure that the opaque box is empty, and so he knows for sure that he will be a non-millionaire. But since in this case he knows for sure that the opaque box is empty, he knows for sure that even if he were to take just the opaque box, he would be a non-millionaire.

If the subject does not know what he will do, then what he knows is this: either he will take just the opaque box and be a millionaire, or he will take both boxes and be a non-millionaire. From this, however, it follows neither that (i) if he took just the opaque box, he would be a millionaire, nor that (ii) if he took both boxes he would be a non-millionaire. For (i), the subject knows, is true iff the opaque box is filled with a million dollars,

and (ii), the subject knows, is true iff the opaque box is empty. Thus, if (i) followed from what the agent knows, he could conclude for certain that the opaque box contains a million dollars, and if (ii) followed from what the agent knows, he could conclude that the opaque box is empty. Since the subject, we have supposed, does not know what he will do, he can conclude neither that the opaque box contains a million dollars nor that it is empty. Therefore neither (i) nor (ii) follows from what the subject knows.

Rational choice in Newcomb's situation, we maintain, depends on a comparison of what would happen if one took both boxes with what would happen if one took only the opaque box. What the agent knows for sure is this: if he took both boxes, he would get a thousand dollars more than he would if he took only the opaque box. That, on our view, makes it rational for someone who wants as much money as he can get to take both boxes, and irrational to take only one box.

Why, then, does it seem obvious to many people that if the predictor is known to be infallible, it is rational to take only the opaque box and irrational to take both boxes? We have three possible explanations. The first is that a person may have a tendency to want to bring about an indication of a desired state of the world, even if it is known that the act that brings about the indication in no way brings about the desired state itself. Taking just the opaque box would be a sure indication that it contained a million dollars, even though taking just the opaque box in no way brings it about that the box contains a million dollars.

The second possible explanation lies in the force of the argument "If you're so rational and you want to be rich, why

ain't you?" That argument, though, if it holds good, should apply equally well to the modified Newcomb situation, with a predictor who is known to be highly accurate but fallible. There the conclusion of the argument seems absurd: according to the argument, having already received the million dollars, one should pass up the additional thousand dollars one is free to take, on the grounds that those who are disposed to do so tend to end up millionaires. Since the argument leads to an absurd conclusion in one case, there must be something wrong with it.

The third possible explanation is a fallacious inference from

> Either I shall take one box and be a millionaire, or I shall take both boxes and be a non-millionaire

to the conclusion

> If I were to take one box, I would be a millionaire, and if I were to take both boxes, I would be a non-millionaire.

If, to someone who is free of fallacies, it is still intuitively apparent that the subject should take only the opaque box, we have no further arguments to give him. If in addition he thinks the subject should take only the opaque box even in the case where the predictor is known to be somewhat fallible, if he also thinks that in the modified Newcomb situation the subject, on receiving the extra million dollars, should take the extra thousand, if he also thinks that it is rational for Reoboam to be severe, and if he also thinks it is rational for Solomon to abstain from his neighbor's wife, then he may genuinely have the intuitions

of a $\mathcal{V}$-maximizer: $\mathcal{V}$-maximization then provides a systematic account of his intuitions. If he thinks some of these things but not all of them, then we leave it to him to provide a systematic account of his views. Our own views are systematically accounted for by $\mathcal{U}$-maximization.

## NOTES

1. This is stated by Lewis (1965, **note 10).**

2. This is our understanding of a proposal made by Jeffrey at the colloquium on Foundations and Applications of Decision Theory, University of Western Ontario, 1975.

3. Meeting of the Canadian Philosophical Association, 1972. Nozick gives a similar example (1969, p. 125).

4. Nozick (1969) in effect endorses the Principle of Dominance with Stochastic Independence (p. 127), but not $\mathcal{V}$-maximization: in cases of the kind we have been considering, he considers the recommendations of $\mathcal{V}$-maximization "perfectly wild" (p. 126).

5. For $\mathcal{V}$-maximizing treatments of Newcomb's problem, see Bar Hillel and Margalit (1972) and Levi (1975).

## REFERENCES

Bar Hillel, Maya and Margalit, Avishi, "Newcomb's Paradox Revisited," British Journal for the Philosophy of Science 23: 295-304 (1972).