ROBERT NOZICK

NEWCOMB'S PROBLEM AND TWO PRINCIPLES OF CHOICE*

> Both it and its opposite must involve no mere artificial illusion such as at once vanishes upon detection, but a natural and unavoidable illusion, which even after it has ceased to beguile still continues to delude though not to deceive us, and which though thus capable of being rendered harmless can never be eradicated.
>
> IMMANUEL KANT, *Critique of Pure Reason*, A422, B450

I

Suppose a being in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, etc.) You know that this being has often correctly predicted your choices in the past (and has never, so far as you know, made an incorrect prediction about your choices), and furthermore you know that this being has often correctly predicted the choices of other people, many of whom are similar to you, in the particular situation to be described below. One might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct.

There are two boxes, (B1) and (B2). (B1) contains $1000. (B2) contains either $1000000 ($M), or nothing. What the content of (B2) depends upon will be described in a moment.

$$\text{(B1) } \{\$1000\} \qquad \text{(B2) } \left\{ \begin{array}{c} \$M \\ \text{or} \\ \$0 \end{array} \right\}$$

You have a choice between two actions:

(1) taking what is in both boxes

(2) taking only what is in the second box.

Furthermore, and you know this, the being knows that you know this, and so on:

(I) If the being predicts you will take what is in both boxes, he does not put the $M in the second box.

(II) If the being predicts you will take only what is in the second box, he does put the $M in the second box.[1]

The situation is as follows. First the being makes its prediction. Then it puts the $M in the second box, or does not, depending upon what it has predicted. Then you make your choice. What do you do?

There are two plausible looking and highly intuitive arguments which require different decisions. The problem is to explain why one of them is not legitimately applied to this choice situation. You might reason as follows:

First Argument: If I take what is in both boxes, the being, almost certainly, will have predicted this and will not have put the $M in the second box, and so I will, almost certainly, get only $1000. If I take only what is in the second box, the being, almost certainly, will have predicted this and will have put the $M in the second box, and so I will, almost certainly, get $M. Thus, if I take what is in both boxes, I, almost certainly, will get $1000. If I take only what is in the second box, I, almost certainly, will get $M. Therefore I should take only what is in the second box.

Second Argument: The being has already made his prediction, and has already either put the $M in the second box, or has not. The $M is either already sitting in the second box, or it is not, and which situation obtains is already fixed and determined. If the being has already put the $M in the second box, and I take what is in both boxes I get $M+$1000, whereas if I take only what is in the second box, I get only $M. If the being has not put the $M in the second box, and I take what is in both boxes I get $1000, whereas if I take only what is in the second box, I get no money. Therefore, whether the money is there or not, and which it is already fixed and determined, I get $1000 more by taking what is in both boxes rather than taking only what is in the second box. So I should take what is in both boxes.

Let me say a bit more to emphasize the pull of each of these arguments:

The First: You know that many persons like yourself, philosophy

teachers and students, etc., have gone through this experiment. All those who took only what was in the second box, included those who knew of the second argument but did not follow it, ended up with $\$M$. And you know that all the shrewdies, all those who followed the second argument and took what was in both boxes, ended up with only $1000. You have no reason to believe that you are any different, *vis-à-vis* predictability, than they are. Furthermore, since you know that I have all of the preceding information, you know that I would bet, giving high odds, and be rational in doing so, that if you were to take both boxes you would get only $1000. And if you were to irrevocably take both boxes, and there were some delay in the results being announced, would not it be rational for you to then bet with some third party, giving high odds, that you will get only $1000 from the previous transaction? Whereas if you were to take only what is in the second box, would not it be rational for you to make a side bet with some third party that you will get $\$M$ from the previous transaction? Knowing all this (though no one is actually available to bet with) do you really want to take what is in both boxes, acting against what you would rationally want to bet on?

The Second: The being has already made his prediction, placed the $\$M$ in the second box or not, and then left. This happened one week ago; this happened one year ago. Box (B1) is transparent. You can see the $1000 sitting there. The $\$M$ is already either in the box (B2) or not (though you cannot see which). Are you going to take only what is in (B2)? To emphasize further, from your side, you cannot see through (B2), but from the other side it is transparent. I have been sitting on the other side of (B2), looking in and seeing what is there. Either I have already been looking at the $\$M$ for a week or I have already been looking at an empty box for a week. If the money is already there, it will stay there whatever you choose. It is not going to disappear. If it is not already there, if I am looking at an empty box, it is not going to suddenly appear if you choose only what is in the second box. Are you going to take only what is in the second box, passing up the additional $1000 which you can plainly see? Furthermore, I have been sitting there looking at the boxes, hoping that you will perform a particular action. Internally, I am giving you advice. And, of course, you already know which advice I am silently giving to you. In either case (whether or not I see the $\$M$ in the second box) I am hoping that you will take what is in both boxes. You

know that the person sitting and watching it all hopes that you will take the contents of both boxes. Are you going to take only what is in the second box, passing up the additional $1000 which you can plainly see, and ignoring my internally given hope that you take both? Of course, my presence makes no difference. You are sitting there alone, but you know that if some friend having your interests at heart *were* observing from the other side, looking into both boxes, he *would* be hoping that you would take both. So will you take only what is in the second box, passing up the additional $1000 which you can plainly see?

I should add that I have put this problem to a large number of people, both friends and students in class. To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly.[2]

Given two such compelling opposing arguments, it will not do to rest content with one's belief that one knows what to do. Nor will it do to just repeat one of the arguments, loudly and slowly. One must also disarm the opposing argument; explain away its force while showing it due respect.

Now for an unusual suggestion. It might be a good idea for the reader to stop reading this paper at the end of this section (but do, please, return and finish it), mull over the problem for a while (several hours, days) and then return. It is not that I claim to solve the problem, and do not want you to miss the joy of puzzling over an unsolved problem. It is that I want you to understand my thrashing about.

II

My strategy in attacking this problem is ostrich-like; that is, I shall begin by ignoring it completely (except in occasional notes) and proceed to discuss contemporary decision theory. Though the problem is not, at first, explicitly discussed, the course my discussion takes is influenced by my knowledge of the problem. Later in the paper, I shall remove my head from the sand, and face our problem directly, hopefully having advanced towards a solution, or at least having sharpened and isolated the problem.

Writers on decision theory state two principles to govern choices

among alternative actions.

> *Expected Utility Principle:* Among those actions available to a person, he should perform an action with maximal expected utility.

The expected utility of an action yielding the exclusive outcomes $O_1, \ldots, O_n$ with probabilities $p_1, \ldots, p_n$ respectively,

$$\left(\sum_{i=1}^{n} p_i = 1\right) \text{ is } p_1 \times u(O_1) + p_2 \times u(O_2) + \cdots + p_n \times u(O_n),$$

$$\text{i.e., } \sum_{i=1}^{n} p_i \times u(O_i).$$

> *Dominance Principle:* If there is a partition of states of the world such that relative to it, action A weakly dominates action B, then A should be performed rather than B.

Action A weakly dominates action B for person P iff, for each state of the world, P either prefers the consequence of A to the consequence of B, or is indifferent between the two consequences, and for some state of the world, P prefers the consequence of A to the consequence of B.

There are many interesting questions and problems about the framework used or assumed by these principles and the conditions governing preference, indifference, and probability which suffice to yield the utility measure, and the exact way the principles should be formulated.[3] The problem I want to begin with is raised by the fact that for some situations, one of the principles listed above requires that one choose one action whereas the other principle requires that one choose another action. Which should one follow?

Consider the following situation, where A and B are actions, S_1 and S_2 are states of the world, and the numerical entries give the utility of the consequences, results, effects, outcomes, upshots, events, states of affairs, etc., that obtain, happen, hold, etc., if the action is done and the state of the world obtains.

	S_1	S_2
A:	10	4
B:	8	3

According to the dominance principle, the person should do A rather than B. (In this situation A strongly dominates B, that is, for each state of nature the person prefers the consequence of A to the consequence of B.) But suppose the person believes it very likely that if he does A, S_2 will obtain, and if he does B, S_1 will obtain. Then he believes it very likely that if he does A he will get 4, and if he does B he will get 8.

	S_1	S_2
A:	10	4
B:	8	3

The expected utility of $A = \text{prob}(S_1/A)\,10 + \text{prob}(S_2/A)\,4$. The expected utility of $B = \text{prob}(S_1/B)\,8 + \text{prob}(S_2/B)\,3$. If, for example,

$$\text{prob}(S_1/A) = .2$$
$$\text{prob}(S_2/A) = .8$$
$$\text{prob}(S_1/B) = .9$$
$$\text{prob}(S_2/B) = .1,$$

then the expected utility of $A = 5.2$, and the expected utility of $B = 7.5$. Thus the expected utility principle requires the person to do B rather than A.[4]

The dominance principle as presented here speaks of dominance relative to a partition of the states of the world. This relativization is normally not made explicit, which perhaps accounts for the fact that writers did not mention that it may be that relative to one partition of the states of the world, one action A dominates another, whereas relative to another partition of the states of the world, it does not.

It will be helpful to have before us two facts:

First: Suppose a matrix is given, with states $S_1, \ldots, S_n$, in which action A does not dominate action B. If there is some rearrangement of the utility entries in the row for action A which gives a new row which dominates the row for action B, then there are states $T_1, \ldots, T_n$ such that in the matrix with these states, action A dominates action B.

Proof: I shall describe how one can get the appropriate states $T_1, \ldots, T_n$ in one case. It is obvious how this procedure can be used generally. Suppose that $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are utility numbers such that, for all i, $a_i \geqslant b_i$, and for some i, $a_i > b_i$. We may suppose that a_i is the entry

in the A row for the ith column, that is, for state S_i. We might, for example, have the following matrix:

	S_1	S_2	S_3	.....	S_n
A:	a_1	a_2	a_3	.....	a_n
B:	b_{12}	b_3	b_{19}	.....	b_6

Let

$T_1 = A \,\&\, S_{12}$ or $B \,\&\, S_1$[5]
$T_2 = A \,\&\, S_3$ or $B \,\&\, S_2$
$T_3 = A \,\&\, S_{19}$ or $B \,\&\, S_3$
$\vdots$
$T_n = A \,\&\, S_6$ or $B \,\&\, S_n$.

Thus we get the matrix,

	T_1	T_2	T_3	.....	T_n
A:	a_{12}	a_3	a_{19}	.....	a_6
B:	b_{12}	b_3	b_{19}	.....	b_6

In this matrix, action A dominates action B. Since the kind of procedure followed does not depend on any specific features of the example, the point is made.

Second: Suppose there is a matrix with states $S_1, \ldots, S_n$ such that action A dominates action B. If there is some rearrangement of the utility entries in the B row so that the rearranged row is not dominated by A, then there are states $T_1, \ldots, T_n$ such that if the matrix is set up with these states, B is not dominated by A.

Proof: Suppose that $a_i \geqslant b_i$, for all i; $a_i > b_i$ for some i; and that some B-row value is greater than some A-row value. (Given that there is some arrangement in which A dominates B, this last supposition follows from its being possible to rearrange the B row so that it is not dominated by the A row.) Suppose, without loss of generality that $b_{12} > a_2$. Thus we have the following matrix:

	S_1	S_2	S_3	.....	S_n
A:	a_1	a_2	a_3	.....	a_n
B:	b_1	b_2	b_3	.....	b_n

Let

$T_1 = S_1$
$T_2 = A \,\&\, S_2$ or $B \,\&\, S_{12}$
$T_3 = S_3$
$\vdots$
$T_{11} = S_{11}$
$T_{12} = A \,\&\, S_{12}$ or $B \,\&\, S_2$
$T_{13} = S_{13}$
$\vdots$
$T_n = S_n$.

Thus we get the following matrix:

	T_1	T_2	T_3	.....	T_{12}	.....	T_n
A:	a_1	a_2	a_3	.....	a_{12}	.....	a_n
B:	b_1	b_{12}	b_3	.....	b_2	.....	b_n

Since $b_{12} > a_2$, A does not dominate B.

It may seem that the states $T_1, \ldots, T_n$ defined in terms of the actions A and B, and the states $S_1, \ldots, S_n$ are contrived states, which some general condition could exclude. It should be noted that – since the states $S_1, \ldots, S_n$ can be defined in terms of the actions A and B and the states $T_1, \ldots, T_n$ (I will give some examples below) – attempts to show that $T_1, \ldots, T_n$ are contrived will face many of the problems encountered in ruling out Goodman-style predicates. Furthermore, as we shall see soon, there are cases where the S states and the T states which are interdefinable in this way, both seem perfectly natural and uncontrived.

The fact that whether one action dominates another or not may depend upon which particular partition of the states of the world is used would cause no difficulty if we were willing to apply the dominance principle to *any* partition of the states of the world. Since we are not, this raises the question of when the dominance principle is to be used. Let us look at some examples.

Suppose that I am about to bet on the outcome of a horserace in which only two horses, H_1 and H_2, are running. Let:

S_1 = Horse H_1 wins the race.
S_2 = Horse H_2 wins the race.
A_1 = I bet on horse H_1.
A_2 = I bet on horse H_2.

Suppose that I will definitely bet on one of the two horses, and can only bet on one of the two horses, and that the following matrix describes the situation. (I might have been offered the opportunity to enter this situation by a friend. Certainly no race track would offer it to me.)

	S_1	S_2
A_1:	I win \$50	I lose \$5
A_2:	I lose \$6	I win \$49

Suppose further that the personal probability for me that H_1 wins is .2, and the personal probability for me that H_2 wins is .8. Thus the expected utility of A_1 is $.2 \times u$(/I win \$50/)$+.8 \times u$(/I lose \$5/). The expected utility of A_2 is $.2 \times u$(/I lose \$6/)$+.8 \times u$(/I win \$49/). Given my utility assignment to these outcomes, the expected utility of A_2 is greater than that of A_1. Hence the expected utility principle would have me do A_2 rather than A_1.

However, we may set the matrix up differently. Let:

S_3 = I am lucky in my bet.
S_4 = I am unlucky in my bet.

(Given that I am betting on only one horse today, we could let S_3 = The only horse I bet on today wins. Similarly for S_4, with 'loses' substituted for 'wins'.) Thus we have the following matrix:

	S_3	S_4
A_1:	I win \$50	I lose \$5
A_2:	I win \$49	I lose \$6

But when set up in this way, A_1 dominates A_2. Therefore the dominance principle would have me do A_1 rather than A_2.[6]

In this example, the states are logically independent of which action I perform; from the fact that I perform $A_1(A_2)$ one cannot deduce which state obtains, and from the fact that S_1 (S_2, S_3, S_4) obtains one cannot deduce which action I perform. However one pair of states was not probabilistically independent of my doing the actions.[7] Assuming that S_1 and S_2 are each probabilistically independent of both A_1 and A_2, prob (S_3/I do A_1)=.2; prob (S_3/I do A_2)=.8; prob (S_4/I do A_1)=.8; prob (S_4/I do A_2)=.2. Thus neither of the states S_3 or S_4 is probabilistically independent of each of the actions A_1 and A_2.[8]

In this example, it is clear that one does not wish to follow the recommendation of the dominance principle. And the explanation seems to hinge on the fact that the states are not probabilistically independent of the actions. Even though one can set up the situation so that one action dominates another, I believe that if I do A_1, the consequence will probably be the italicized consequence in its row, and I believe that if I do A_2, the consequence will probably be the italicized consequence in A_2's row. And given my assignment of utilities in this case, and the probabilities I assign (the conditional probabilities of the states given the actions) it is clear why I prefer to do A_2, despite the fact that A_1 dominates A_2.

	S_3	S_4
A_1:	I win \$50	*I lose* \$5
A_2:	*I win* \$49	I lose \$6

Let us consider another example: Suppose that I am playing roulette on a rigged wheel, and that the owner of the casino offers me a chance to choose between actions A_1 and A_2 so that the following matrix describes the situation (where S_1 = black comes up on the next spin; S_2 = red comes up on the next spin):

	S_1	S_2
A_1:	I win \$10	I win \$100
A_2:	I win \$5	I win \$90

Finally suppose that I know that the owner's employee, who is overseeing the wheel and who I am confident is completely loyal to the owner, has been instructed to make black come up on the next spin if I choose A_1 and to make red come up on the next spin if I choose A_2. Clearly even though A_1 dominates A_2, given my knowledge of the situation I should choose A_2. I take it that this needs no argument. It seems that the reason that I should not be guided by dominance considerations is that the states S_1 and S_2 are not probabilistically independent of my actions A_1 and A_2. We can set up the situation so that the states are probabilistically independent of the actions. But when set up in this way, I am led, given my utility assignment to the outcomes, to do A_2.

Let S_3 = the fellow running the roulette wheel follows his boss' instructions; S_4 = the fellow running the roulette wheel disobeys his boss' instructions. (Note that $S_3 = A_1$ & S_1 or A_2 & S_2; $S_4 = A_1$ & S_2 or A_2 & S_1.)

We then have the following matrix:

	S_3	S_4
A_1:	I win \$10	I win \$100
A_2:	I win \$90	I win \$5

Even if I am not sure that S_3 is true, so long as the personal probability of S_3 for me is sufficiently high, I will be led to do A_2, given my utility assignment to the outcomes.

These examples suggest that one should not apply the dominance principle to a situation where the states are not probabilistically independent of the actions. One wishes instead to maximize the expected utility. However, the probabilities that are to be used in determining the expected utility of an action must now be the conditional probabilities of the states given that the action is done. (This is true generally. However when the states are probabilistically independent of the actions, the conditional probability of each state given that one of the actions is done will be equal to the probability of the state, so the latter may be used.) Thus in the roulette wheel example, we may still look at the first matrix given. However, one does not wish to apply the dominance principle but to find the expected utility of the actions, which in our example are:

$$\text{E.U.}(A_1) = \text{prob}(S_1/A_1) \times u(/\text{I win \$10}/) + \text{prob}(S_2/A_1) \times u(/\text{I win \$100}/)$$

$$\text{E.U.}(A_2) = \text{prob}(S_1/A_2) \times u(/\text{I win \$5}/) + \text{prob}(S_2/A_2) \times u(/\text{I win \$90}/).^{9}$$

The following position appropriately handles the examples given thus far (ignoring Newcomb's example with which the paper opens) and has intuitive appeal.[10]

(1) It is legitimate to apply dominance principles if and only if the states are probabilistically independent of the actions.

(2) If the states are not probabilistically independent of the actions, then apply the expected utility principle, using as the probability-weights the conditional probabilities of the states given the actions.

Thus in the following matrix, where the entries in the matrix are utility numbers,

	S_1	S_2	.....	S_n
A:	O_1	O_2	.....	O_n
B:	U_1	U_2	.....	U_n

the expected utility of A is $\sum_{i=1}^{n} \text{prob}(S_i/A)\, O_i$, and the expected utility of B is $\sum_{i=1}^{n} \text{prob}(S_i/B)\, U_i$.

III

Is this position satisfactory? Consider the following example: P knows that S or T is his father, but he does not know which one is. S died of some terrible inherited disease, and T did not. It is known that this disease is genetically dominant, and that P's mother did not have it, and that S did not have the recessive gene. If S is his father, P will die of this disease; if T is his father, P will not die of this disease. Furthermore, there is a well-confirmed theory available, let us imagine, about the genetic transmission of the tendency to decide to do acts which form part of an intellectual life. This tendency is genetically dominant. S had this tendency (and did not have the recessive gene), T did not, and P's mother did not. P is now deciding whether (a) to go to graduate school and then teach, or (b) to become a professional baseball player. He prefers (though not enormously) the life of an academic to that of a professional athlete.

	S is P's father	T is P's father
A:	x	y
B:	z	w

$x=P$ is an academic for a while, and then dies of the terrible disease; $z=P$ is a professional athlete for a while, and then dies of the terrible disease; $y=P$ is an academic and leads a normal academic life; $w=P$ is a professional athlete and leads the normal life of a professional athlete, though doing a bit more reading; and P prefers x to z, and y to w. However, the disease is so terrible that P greatly prefers w to x. The matrix might be as follows:

	S is P's father	T is P's father
A:	-20	100
B:	-25	95

Suppose that our well-confirmed theory tells us, and P, that if P chooses the academic life, then it is likely that he has the tendency to

choose it; if he does not choose the academic life, then it is likely that he does not have the tendency. Specifically

prob (P has the tendency/P decides to do A) = .9
prob (P does not have the tendency/P decides to do A) = .1
prob (P has the tendency/P decides to do B) = .1
prob (P does not have the tendency/P decides to do B) = .9.

Since P has the tendency iff S is P's father, we have

prob (S is P's father/P decides to do A) = .9
prob (T is P's father/P decides to do A) = .1
prob (S is P's father/P decides to do B) = .1
prob (T is P's father/P decides to do B) = .9.

The dominance principle tells P to do A rather than B. But according to the position we are now considering, in situations in which the states are not probabilistically independent of the actions, the dominance principle is not to be used, but rather one is to use the expected utility principle with the conditional probabilities as the weights. Using the above conditional probabilities and the above numerical assumptions about the utility values, we get:

The expected utility of $A = .9 \times -20 + .1 \times 100 = -8$
The expected utility of $B = .1 \times -25 + .9 \times 95 = 83$.

Since the expected utility of B is greater than that of A, the position we are considering would have P do B rather than A. But this recommendation is perfectly wild. Imagine P saying, 'I am doing B because if I do it it is less likely that I will die of the dread disease'. One wants to reply, 'It is true that you have got the conditional probabilities correct. If you do A it is likely that S is your father, and hence likely that you will die of the disease, and if you do B it is likely that T is your father and hence unlikely that you will die of the disease. But which one of them is your father is already fixed and determined, and has been for a long time. The action you perform legitimately affects our estimate of the probabilities of the two states, but which state obtains does not depend on your action at all. By doing B you are not *making* it less likely that S is your father, and by doing B you are not making it less likely that you will die of the disease'. I do not claim that this reply is without its

problems.[11] Before considering another example, let us first state a principle not under attack:

> The Dominance Principle is legitimately applicable to situations in which the states are probabilistically independent of the actions.[12]

If the states are not probabilistically independent of the actions, it *seems* intuitive that the expected utility principle is appropriate, and that it is not legitimate to use the dominance principle if it yields a different result from the expected utility principle. However, in situations in which the states, though not probabilistically independent of the actions, are already fixed and determined, where the actions do not affect whether or not the states obtain, then it *seems* that it is legitimate to use the dominance principle, and illegitimate to follow the recommendation of the expected utility principle if it differs from that of the dominance principle.

For such situations – where the states are not probabilistically independent of the actions, though which one obtains is already fixed and determined – persons may differ over what principle to use.

Of the twelve sorts of situation in which it is not the case both that none of the states are already fixed and determined and none of the states are probabilistically independent of the actions, I shall discuss only one; namely, where each of the states is already fixed and determined, and none of the states are probabilistically independent of the alternative actions.[13]

The question before us is: In this sort of situation, in which all of the states are already fixed and determined, and none of the states are probabilistically independent of the acts, and the dominance principle requires that one do one action, whereas the expected utility principle requires that one do another, should one follow the recommendation of the dominance principle or of the expected utility principle?

The question is difficult. Some may think one should follow the recommendation of the dominance principle; others may think one should follow the recommendation of the expected utility principle in such situations.

Now for the example which introduces a bit of reflexivity which I hope will soon serve us in good stead. Suppose that there are two inherited

tendencies ('tendencies' because there is some small probability that it would not be followed in a specific situation):

(1) an inherited tendency to think that the expected utility principle should be used in such situations. (If P has this tendency, he is in state S_1.)

(2) an inherited tendency to think that the dominance principle should be used in such situations. (If P has this tendency, he is in state S_2.)

It is known on the basis of *post mortem* genetic examinations that

(a) P's mother had two neutral genes. (A gene for either tendency genetically dominates a neutral gene. We need not here worry about the progeny who has a gene for each tendency.)

(b) One of the men who may be P's father, had two genes for the first tendency.

(c) The other man who may be P's father had two genes for the second tendency.

So it is known that P has one of the tendencies, but it is not known which one he has. P is faced with the following choice:

	S_1	S_2
A:	10	4
B:	8	3

The choice matrix might have arisen as follows. A deadly disease is going around, and there are two effective vaccines against it. (If both are given, the person dies.) For each person, the side effects of vaccine B are worse than that of vaccine A, and each vaccine has worse side effects on persons in S_2 than either does on persons in S_1.

Now suppose that the theory about the inherited tendencies to choice, tells us, and P knows this, that from a person's choice in *this* situation the probabilities of his having the two tendencies, given that he has one of the two, can be estimated, and in particular

$$\text{prob}\,(S_1/A) = .1$$
$$\text{prob}\,(S_2/A) = .9$$
$$\text{prob}\,(S_1/B) = .9$$
$$\text{prob}\,(S_2/B) = .1.$$

What should P do? What would you do in this situation?

P may reason as follows: if I do A, then very probably S_2 obtains, and I will get 4. If I do B, then very probably S_1 holds, and I will get 8. So I will do B rather than A.

One wants to reply: whether S_1 or S_2 obtains is already fixed and determined. What you decide to do would not bring about one or the other of them. To emphasize this, let us use the past tense. For you are in S_1 iff you were in S_1 yesterday; you are in S_2 iff you were in S_2 yesterday. But to reason 'If I do A then very probably I was in S_2 yesterday, and I will get 4. If I do B, then very probably, I was in S_1 yesterday, and I will get 8. So I will now do B rather than A' is absurd. What you decide to do does not affect which state you were in yesterday. For either state, over which you have no control, you are better off doing A rather than B. To do B for reasons such as the above is no less absurd than someone who has already taken vaccine B yesterday doing some other act C today because the prob (He was in S_1 yesterday/He does C today) is very high, and he wants the (delayed) side effects of the vaccine he has already taken to be less severe.

If an explanation runs from x to y, a correct explanatory theory will speak of the conditional probability prob (y/x). Thus the correct explanatory theory of P's choice in this situation will speak of

$$\text{prob}\,(P \text{ does } A/P \text{ is in } S_1)$$
$$\text{prob}\,(P \text{ does } A/P \text{ is in } S_2)$$
$$\text{prob}\,(P \text{ does } B/P \text{ is in } S_1)$$
$$\text{prob}\,(P \text{ does } B/P \text{ is in } S_2).$$

From these, the theory may enable us to determine

$$\text{prob}\,(P \text{ is in } S_1/P \text{ does } A)$$
$$\text{prob}\,(P \text{ is in } S_2/P \text{ does } A)$$
$$\text{prob}\,(P \text{ is in } S_1/P \text{ does } B)$$
$$\text{prob}\,(P \text{ is in } S_2/P \text{ does } B)$$

but these would not be the basic explanatory probabilities. Supposing that probabilistic explanation is legitimate, we could explain why P does A by having among our antecedent conditions the statement that P is in S_2 ,but we cannot *explain* why P is in S_2 by having among our antecedent conditions the statement that P does A (though P's doing A may be our reason for believing he is in S_2). Given that when the explanatory line runs from x to y (x is part of the explanation of y) and not from y to x, the theory will speak of and somehow distinguish the conditional probabilities prob (y/x), then the probability prob (x/y) will be a *likeli-*

hood (as, I think, this term is used in the statistical literature). Looking at the likelihoods of the states given the actions may perhaps give one the illusion of control over the states. But I suggest that when the states are already fixed and determined, and the explanatory theory has the influence running from the states to the actions, so that the conditional probabilities of the states on the actions are likelihoods, then if the dominance principle applies, it should be applied.

If a state is part of the explanation of deciding to do an action (if the decision is made) and this state is already fixed and determined, then the decision, which has not yet been made, cannot be part of the explanation of the state's obtaining. So we need not consider the case where prob (state/action) is in the basic explanatory theory, for an already fixed state.[14] What other possibilities are there for already fixed and determined states? One possibility would be a situation in which the states are not part of the explanation of the decision, and the decision is not part of the explanation of which state obtains, but some third thing is part of the explanation of the states obtaining, and the decision's being made. Hence neither prob (state of the matrix obtaining/*P* does a specific action) nor prob (*P* does a specific action/state of the matrix obtains) would be part of the basic explanatory theory (which has conditional probabilities from antecedent to consequent going in the direction of explanation).

Let us consider a case like this, whose matrix exemplifies the structure of the prisoners' dilemma situation, much discussed by game theorists.[15] There are two people, (I) and (II) and the following matrix describes their situation (where the first entry in each box represents the payoff to person (I) and the second entry represents the payoff to person (II)). The situation arises just once, and the persons cannot get together to agree upon a joint plan of action.

		(II)	
		C	*D*
(I)	*A*:	10, 3	4, 4
	B:	8, 8	3, 10

Notice that for person (I), action *A* dominates action *B*, and for person (II), action *D* dominates action *C*. Hence if each performs his dominant action, each ends up with 4. But if each performs the non-dominant action, each ends up with 8. So, in this situation, both persons' following the dominance principle leaves each worse off than if both did not follow the dominance principle.

People may differ over what should be done in this situation. Let us, once again, suppose that there are two inherited tendencies, one to perform the dominant action in this situation, and one to perform the other action. Either tendency is genetically dominant over a possible third inherited trait. Persons (I) and (II) are identical twins, who care only about their own payoffs as represented in this matrix, and knows that their mother had the neutral gene, one of their two possible fathers had only the gene to perform the dominant action, and the other had only the gene not to perform the dominant action. Neither knows which man was their father, nor which of the genes they have. Each knows, given the genetic theory, that it is almost certain that if he performs the dominant (dominated) action his brother will also. We must also suppose that the theory tells us and them that given all this information upon which they base their choice, the correlation between their actions holds as almost certain, and also given *this* additional information, it holds as almost certain, etc.

I do not wish here to discuss whether one should or should not perform the dominant action in Prisoners' Dilemma situations. I wish merely to consider the following argument for not performing the dominant action in the situation I have just described. Suppose brother I argues: 'If I perform the dominant action then it is almost certain$_1$ that I have that gene, and therefore that my brother does also, and so it is almost certain$_2$[16] that he will also perform the dominant action and so it is almost certain$_2$ that I will get 4. Whereas if I perform the dominated action, for similar reasons, it is almost certain that my brother will also, and hence it is almost certain that I will get 8. So I should perform the dominated action'.

Here one surely wishes to reply that *this* argument is not a good argument for performing the dominated action. For what this brother does will not affect what the other brother does. (To emphasize this, suppose that brother II has already acted, though brother I does not yet know what he has done.) Perhaps in prisoners' dilemma situations one should perform the dominated action, but *this* argument does not show that one should in this situation.

The examples thus far considered lead me to believe that if the actions

or decisions to do the actions do not affect, help bring about, influence, etc., *which* state obtains, then whatever the conditional probabilities (so long as they do not indicate an influence), one should perform the dominant action.

If the considerations thus far adduced are convincing, then it is clear that one should also choose the dominant action in the following situations, having the same structure (matrix) as Newcomb's, and differing only in that:

(1) The being makes his prediction and sets the process going whereby the $\$M$ gets placed in the second box, or not. You then make your choice, and *after* you do, the (long) process terminates and the $\$M$ gets in the box, or not. So while you are deciding, the $\$M$ is not already there, though at this time he has already decided whether it will be or not.

(2) The being gathers his data on the basis of which he makes his prediction. You make your choice (e.g., press one of two buttons which will open one or both boxes later by delayed action), and he then makes his prediction, on the basis of the data previously gathered, and puts the $\$M$ in, or not.

This suggests that the crucial fact is *not* whether the states are already fixed and determined but whether the actions *influence* or *affect* which state obtains.

Setting up a simple matrix,[17] we have the following possibilities (with the matrix entries being recommended decision policies for the situation).

	A dominant action is available	No dominant action is available
The actions influence which state obtains. The conditional probabilities differ.	(I) Maximize Expected Utility	(II) Maximize Expected Utility
No influence of actions on states. However conditional probabilities differ.	(III)	(IV)
No influence of actions on states. The conditional probabilities are all the same.	(V) Do dominant action (or, equivalently, Maximize Expected Utility)	(VI) Maximize Expected Utility

The standard theories make the recommendations in (V) and (VI).

They do not consider (I) and (II), but (ignoring other difficulties there might be with the policy) Maximizing Expected Utility seems reasonable here. The difficulties come in the middle row. (III) is the situation exemplified by Newcomb's situation and the other examples we have listed from the person choosing whether to lead the academic life, onwards. I have argued that in these situations, one should choose the dominant action and ignore the conditional probabilities which do not indicate an influence. What then should one do in situation (IV), where which action is done does not influence which state obtains, where the conditional probabilities of the states given the actions differ, and where *no* dominant action is available. If the lesson of case (III) is that one should ignore conditional probabilities which do not indicate an influence, must not one ignore them completely in case (IV) as well?

Not exactly. What one should do, in a choice between two actions A and B, is the following.[18] Let $p_1,\ldots,p_n$ be the conditional probability distribution of action A over the n states; let $q_1,\ldots,q_n$ be the conditional probability distribution of action B over the n states. A probability distribution $r_1,\ldots,r_n$, summing to 1, is between $p_1,\ldots,p_n$ and $q_1,\ldots,q_n$ iff for each i, r_i is in the closed interval $[p_i, q_i]$ or $[q_i, p_i]$. (Note that according to this account, $p_1,\ldots,p_n$, and $q_1,\ldots,q_n$ are each between $p_1,\ldots,p_n$ and $q_1,\ldots,q_n$.) Now for a recommendation: If relative to each probability distribution between $p_1,\ldots,p_n$ and $q_1,\ldots,q_n$, action A has a higher expected utility than action B, then do action A. The expected utility of A and B is computed with respect to the same probability distribution. It will not, of course, be the case that relative to every possible probability distribution A has a higher expected utility than B. For, by hypothesis, A does not dominate B. However it may be that relative to each probability distribution between $p_1,\ldots,p_n$ and $q_1,\ldots,q_n$, A has a higher expected utility than B. If, on the other hand, it is not the case that relative to each probability distribution between $p_1,\ldots,p_n$ and $q_1,\ldots,q_n$, A has a higher expected utility than B (and it is not the case that relative to each, B has a higher expected utility than A), then we are faced with a problem of decision under constrained uncertainty (the constraints being the end probability distributions), on which kind of problem there is not, so far as I know, agreement in the literature.[19] Since consideration of the issues raised by such problems would take us far afield, we thankfully leave them.

To talk more objectively than some would like, though more intuitively than we otherwise could, since the actions do not affect or influence which state obtains, there is some one probability distribution, which we do not know, relative to which we would like to compare the action *A* and *B*. Since we do not know the distribution, we cannot proceed as in cases (V) and (VI). But since there is *one* unknown correct distribution 'out there', unaffected by what we do, we must, in the procedure we use, compare each action with respect to the *same* distribution. Thus it is, at this point, an irrelevant fact that one action's expected utility computed with respect to one probability distribution is higher than another action's expected utility computed with respect to *another* probability distribution. It may seem strange that for case (IV) we bring in the probabilities in some way (even though they do not indicate an influence) whereas in case (III) we do not. This difference is only apparent, since we could bring in the probabilities in case (III) in exactly the same way. The reason why we need not do this, and need only note that *A* dominates *B*, is that if *A* dominates *B*, then relative to each probability distribution (and therefore for each one between the conditional ones established by the two actions) *A* has a higher expected utility than *B*.[20]

Now, at last, to return to Newcomb's example of the predictor. If one believes, for this case, that there is backwards causality, that your choice causes the money to be there or not, that it causes him to have made the prediction that he made, then there is no problem. One takes only what is in the second box. Or if one believes that the way the predictor works is by looking into the future; he, in some sense, sees what you are doing, and hence is no more likely to be wrong about what you do than someone else who is standing there at the time and watching you, and would normally see you, say, open only one box, then there is no problem. You take only what is in the second box. But suppose we establish or take as given that there is no backwards causality, that what you actually decide to do does not affect what he did in the past, that what you actually decide to do is not part of the explanation of why he made the prediction he made. So let us agree that the predictor works as follows: He observes you sometime before you are faced with the choice, examines you with complicated apparatus, etc., and then uses his theory to predict on the basis of this state you were in, what choice you would make later when faced with the choice. Your deciding to do as you do is not part of the explanation of why he makes the prediction he does, though your being in a certain state earlier, is part of the explanation of why he makes the prediction he does, and why you decide as you do.

I believe that one should take what is in both boxes. I fear that the considerations I have adduced thus far will not convince those proponents of taking only what is in the second box. Furthermore I suspect that an adequate solution to this problem will go much deeper than I have yet gone or shall go in this paper. So I want to pose one question. I assume that it is clear that in the vaccine example, the person should not be convinced by the probability argument, and should choose the dominant action. I assume also that it is clear that in the case of the two brothers, the brother should not be convinced by the probability argument offered. The question I should like to put to proponents of taking only what is in the second box in Newcomb's example (and hence not performing the dominant action) is: what is the difference between Newcomb's example and the other two examples which make the difference between not following the dominance principle, and following it?

If no such difference is produced, one should not rush to conclude that one should perform the dominant action in Newcomb's example. For it must be granted that, at the very least, it is not *as clear* that one should perform the dominant action in Newcomb's example, as in the other two examples. And one should be wary of attempting to force a decision in an unclear case by producing a similar case where the decision is clear, and challenging one to find a difference between the cases which makes a difference to the decision. For suppose the undecided person, or the proponent of another decision, cannot find such a difference. Does not the forcer, now, have to find a difference between the cases which explains why one is clear, and the other is not? And might not *this* difference then be produced by the other person as that which perhaps should yield different decisions in the two cases? Sometimes this will be implausible; e.g., if the difference is that one case is relatively simple, and the other has much additional detail, individually irrelevant, which prevent the other case from being taken in as a whole. But it does seem that someone arguing as I do about a case must not only: (a) describe a similar case which is clear, and challenge the other to state a difference between them which should make a difference to how they are handled, but must also, (b) describe a difference between the cases which explains

why though one case is clear, the other is not, or one is tempted to handle the other case differently. And, assuming that all accept the difference stated in (b) as explaining what it is supposed to explain,

(I) The simplest situation is that in which all agree that the difference mentioned in (b) is not a reason for different decisions in the two cases.

(II) However, if the forcer says it is not a reason for different decisions in the two cases, and the other person says it is or may be, difficult questions arise about upon whom, if anyone, the burden of further argument falls.

What then is the difference that makes some cases clear and Newcomb's example unclear, yet does not make a difference to how the cases should be decided? Given my account of what the crucial factors are (influence, etc.) my answer to this question will have to claim that the clear cases are clear cases of no influence (or, to recall the cases which we considered at the beginning, of influence), and that in Newcomb's example there is the *illusion* of influence. The task is to explain in a sufficiently forceful way what gives rise to this illusion so that, even as we experience it, we will not be deceived by it.

I have said that if the action is referred to in an explanation of the state's obtaining, so that the doing of the action affects or influences which state obtains, then the Dominance Principle should not be applied. And if the explanation of the states' obtaining does not make reference to the action, the action does not influence which state obtains, does not (partly) bring it about that a state obtains, then the Dominance Principle should be applied to such situations where a dominant action is available. But if this is so, where is there room for unclarity about a case? What other possibility is there? Either the action is referred to in the explanation of the state's obtaining, or it is not. How does the temptation to take only what is in the second box arise in the Newcomb example, and why does it linger?

The possibility to which I wish to call attention can be described differently, depending upon other views which one holds. (I describe the possibility specifically with Newcomb's example in mind.) (1) The action *is* referred to in the explanation of the state's obtaining, but the term which refers to the action occurs in the explanation, in a nonextensional belief context. Thus it does not follow from the fact that the action is referred to, in this way, in the explanation of the state's obtaining, that the doing of the action affects which state obtains. (2) The action is not referred to in the explanation of the state's obtaining. What is brought in by the explanation of the state's obtaining is some being's well-founded beliefs about the person's doing the action. Since the person's doing the action is not part of the explanation of the state's obtaining, it does not affect or influence which state obtains.

In Newcomb's example, the predictor makes his prediction on the basis of determining what state the person is in, and what his theory tells him about what such a person will do in the choice situation. Believing his theory accurate, he puts the money in or not, according to his belief about the person's future actions, where this belief depends upon his finding out what initial state the person is in, and what his theory tells him about a person in such an initial state. Thus, if the predictor puts the $\$M$ in the second box, part of the explanation of this is his belief that the person will take only what is in the second box. If he does not put the $\$M$ in the second box, part of the explanation of this is his belief that the person will take what is in both boxes. Thus the explanation of the money's being in the second box (or not) refers to the person's action only in a nonextensional belief context (or does not refer to it at all but only to the predictor's beliefs about it).

It is apparently a persistent temptation for people to believe, when an explanation of something x brings in terms referring to y in a nonextensional belief context (or brings in beliefs about y), that y, in some way, influences or affects x. Thus one finds writers on teleological explanation having to state that in the simple case where someone goes to the refrigerator to get an apple, it is not the apple's being there when he gets there which caused him to go, or which (partly) explains his actions, but rather his beliefs about an apple's being there. But this cannot be the whole story about Newcomb's example. For there are many persons not at all tempted to say that the apple's being there when he gets there influenced his action of going there, who do want to or are tempted to take only what is in the second box.

Let us return to the writers on teleology. To show that the apple's being there does not influence the person's actions, but rather it is his beliefs about the apple's being there that do, they usually argue that even if the apple were not there, so long as the person had the beliefs, he would

act in the same way. The relevant feature of nonextensional belief contexts here is that from P believes that $\ldots x \ldots$, it does not follow that x exists, from P believes that p, it does not follow that p is true. So, the argument runs, he *could* have his beliefs without there being an apple there, and this shows that the apple does not influence his actions in this case. And surely the explanation of his action should be the same, in the case where the apple is in the refrigerator, as in the case where it is not though he believes it is. The parallel argument for Newcomb's example would run: The predictor could believe that you will take only the second even if you do not. This shows that your action does not influence whether or not the money is there, but rather the predictor's beliefs about your action has this influence. But by the conditions of the problem, the predictor is almost certain to predict correctly, so that it is not clear that the predictor could believe that you will take only the second even it you do not. Thus, the condition of the problem which has the predictor's predictions almost certainly being correct tends to get us to treat the predictor's beliefs as though they do not have these nonextensional features. For if his predictions are almost certainly correct, then almost certainly: if he believes you will do A then you will do A.

One further thing should be mentioned. It is a reasonably intuitive principle that if R brings it about that p, and if p if and only if q (for some 'iff' stronger than the material biconditional), then R brings it about that q. Or, if it is up to R whether p, and p iff q (for some strong 'iff'), then it is up to R whether q. Thus one finds writers arguing that if there are necessary and sufficient causal conditions for our actions, which conditions go back to a time before we were born, then what we do is not up to us. For, so the argument runs, those conditions obtaining before we were born clearly were not up to us, and so what they are necessary and sufficient for is not up to us either. I do not wish here to discuss whether this principle is correct. All that is required for my purposes is that the principle have intuitive appeal, and be a hard one to escape.

This would also reinforce the feeling that as choosers in Newcomb's example, we can, somehow, influence what the predictor did. For, one might argue, Newcomb's problem is a problem for the chooser only if what he does is up to him. And if one assumes this, and the principle is operating, then it will be difficult to escape the feeling that what the predictor did is up to you, the chooser.

I do not claim that this last principle alone creates the problem. For the problem does not arise in e.g., the vaccine case.[21] But it does, I believe, contribute to it.

Thus I wish to claim that Newcomb's example is less clear than the others because

(a) in it the explanation of the state's obtaining refers to the action (though this reference occurs in a nonextensional belief-context)

and that

(b) the conditions of the problem prevent one obvious way of refuting the teleologist's view, in this case, which view depends upon the truth that generally if y is part of the explanation of x, then y influences x).

This leads to the feeling that, somehow, you as chooser can influence what the predictor did, and this feeling is perhaps reinforced by the operation of the intuitive principle. All this leads to the lurking feeling that one can now choose to take only what is in the second box, and so make oneself the sort of person who does so, and so, somehow, influence what the predictor did. I hope you find this explanation of why some cases are clear and Newcomb's is not, acceptable, and that it is clear that this difference between the cases should not make a difference to how they are decided.[22]

At this point one perhaps wants to say, 'If you produce a case having the features you say distinguish Newcomb's example from the others, where it is clear that the dominant action should be performed, then I will be convinced that the dominant action should be performed in Newcomb's example. But not until'. If I am right about the role of similar examples, then this cannot be done; an answer to Newcomb's example cannot be forced in this way. Or rather, if it can be done, then it will show that I have not picked out the right difference. For if one case that fits my description is clear, and another which fits it is not clear, then we still have to produce features to explain why one is clear and the other is not. And perhaps *those* features should make a difference between the decisions in the two cases. At some point, given an acceptable explanation of why one case is clear and another is not, one just has to see that the explanatory features do not make a difference to what should be decided in the two cases. Or, at any rate, the point that the explanatory features do not make a difference to what should be decided can itself be forced

by a clear case only at the cost of the claim that those very features explain why some cases are clear and others are not.

In closing this paper, I must muddy up the waters a bit (more?).

(1) Though Newcomb's example suggests much about when to apply the dominance principle, and when to apply the expected utility principle (and hence is relevant to formal decision theory), it is not the expected utility principle which leads some people to choose only what is in the second box. For suppose the probability of the being's predicting correctly was just .6.

> Then the expected utility of taking what is in both boxes = prob (he predicts correctly /I take both) × u(I receive \$1000) + prob (he predicts correctly/ I take only second) × u(I receive \$1001000) = .6 × u(\$1000) + .4 × u(\$1001000).

The expected utility of taking only what is in the second box = .6 × u(\$1000000) + .4 × u(\$0).

And given the utility I assume each of my readers assigns to obtaining these various monetary amounts, the expected utility of taking only what is in the second box is greater than the expected utility of taking what is in both boxes. Yet, I presume, if the probability of the beings predicting correctly were only .6, each of us would choose to take what is in both boxes.

So it is not (just) the expected utility argument that operates here to create the problem in Newcomb's example. It is crucial that the predictor is almost certain to be correct. I refrain from asking a proponent of taking only what is in the second box in Newcomb's example: if .6 is not a high enough probability to lead you to take only what is in the second box, and almost certainty of correct predictions leads you to take only the second, what is the minimum probability of correct prediction which leads you to take only what is in the second box? I refrain from asking this question because I am very unsure about the force of drawing-the-line arguments, and also because the person who wishes to take what is in both boxes may also face a problem of drawing the line, as we shall see in a moment.

(2) If the fact that it is almost certain that the predictor will be correct is crucial to Newcomb's example, this suggests that we consider the case where it *is* certain, where you know the prediction is correct (though you

do not know what the prediction is. Here one naturally argues: I know that if I take both, I will get \$1000. I know that if I take only what is in the second, I get \$$M$. So, of course, I will take only what is in the second. And does a proponent of taking what is in both boxes in Newcomb's example, (e.g., me) really wish to argue that it is the probability, however, minute, of the predictor's being mistaken which make the difference? Does he really wish to argue that if he knows the prediction will be correct, he will take only the second, but that if he knows someone using the predictor's theory will be wrong once in every 20 billion cases, he will take what is in both boxes? Could the difference between one in n, and none in n, for arbitrarily large finite n, make this difference? And how exactly does the fact that the predictor is certain to have been correct dissolve the force of the dominance argument?

To get the mind to really boggle, consider the following.

	S_1	S_2
A:	10	4
B:	8	3

Suppose that you know that either S_1 or S_2 already obtains, but you do not know which, and you know that S_1 will cause you to do B, and S_2 will cause you to do A. Now choose! ('Choose?')

To connect up again with a causalized version of Newcomb's example, suppose you know that there are two boxes, (B1) and (B2). (B1) contains \$1000. (B2) contains either a valuable diamond or nothing. You have to choose between taking what is in both boxes, and taking only what is in the second. You know that there are two states: S_1 and S_2. You do not know which obtains, but you know that whichever does, it has obtained for the past week. If S_2 obtains, it causes you to take only what is in the second, and it has already caused a diamond to be produced in box (B2). If S_1 obtains, it causes you to take what is in both boxes, and does not cause a diamond to be produced in the second box. You know all this. What do you choose to do?

While we are at it, consider the following case where what you decide (and why) either (1) does affect which future state will obtain, upon which consequences depend, or (though this would not be the same problem for the view I have proposed, it might be for yours) even if it does not affect which state obtains, the conditional probabilities of the

states, given what you do and why, differ.

	S_1	S_2
A:	live	die
B:	die	live

(1) Apart from your decisions (if you do not know of this matrix, or know of it and cannot reach a decision), prob S_1 > prob S_2

(2) prob (S_1/do A with (1) as reason) < prob (S_2/do A with (1) as reason)

(3) prob (S_1/do B with (2) as reason) > prob (S_2/do B with (2) as reason)

⋮

even (n) prob (S_1/do A with $n-1$ as reason) < prob (S_2/do A with $n-1$ as reason)

odd (n) prob (S_1/do B with $n-1$ as reason) > prob (S_2/do B with $n-1$ as reason)

⋮

Also: prob (S_1/you do what you do because indifferent between A and B) > prob (S_2/you do what you do because indifferent between A and B)

prob (S_1/doing A with all of the above as reason) <
prob (S_2/doing A with all of the above as reason)
and
prob (S_1/doing B with all of the above as reason) >
prob (S_2/doing B with all of the above as reason).

Finally, where 'all this' refers to all of what is above this place, and reflexively, to the next two, in which it appears:

prob (S_1/doing A with all this as reason) <
prob (S_2/doing A with all this as reason)
and
prob (S_1/doing B with all this as reason) >
prob (S_2/doing B with all this as reason).

What do you do?

Harvard University

REFERENCES

* It is not clear that I am entitled to present this paper. For the problem of choice which concerns me was constructed by someone else, and I am not satisfied with my attempts to work through the problem. But since I believe that the problem will interest and intrigue Peter Hempel and his many friends, and since its publication may call forth a solution which will enable me to stop returning, periodically, to it, here it is. It was constructed by a physicist, Dr. William Newcomb, of the Livermore Radiation Laboratories in California. I first heard the problem, in 1963, from his friend Professor Martin David Kruskal of the Princeton University Department of Astrophysical Sciences. I have benefitted from discussions, in 1963, with William Newcomb, Martin David Kruskal, and Paul Benacerraf. Since then, on and off, I have discussed the problem with many other friends whose attempts to grapple with it have encouraged me to publish my own. It is a beautiful problem. I wish it were mine.

[1] If the being predicts that you will consciously randomize your choice, e.g., flip a coin, or decide to do one of the actions if the next object you happen to see is blue, and otherwise do the other action, then he does not put the $\$M$ in the second box.

[2] Try it on your friends or students and see for yourself. Perhaps some psychologists will investigate whether responses to the problem are correlated with some other interesting psychological variable that they know of.

[3] If the questions and problems are handled as I believe they should be, then some of the ensuing discussion would have to be formulated differently. But there is no point to introducing detail extraneous to the central problem of this paper here.

[4] This divergence between the dominance principle and the expected utility principle is pointed out in Robert Nozick. *The Normative Theory of Individual Choice*, unpublished doctoral dissertation, Princeton University, Princeton, 1963, and in Richard Jeffrey, *The Logic of Decision*, McGraw-Hill, New York, 1965.

[5] This is shorthand for: action A is done and state S_{12} obtains or action B is done and state S_1 obtains. The 'or' is the exclusive or.

[6] Note that

$$S_1 = A_1 \,\&\, S_3 \text{ or } A_2 \,\&\, S_4$$
$$S_2 = A_1 \,\&\, S_4 \text{ or } A_2 \,\&\, S_3$$
$$S_3 = A_1 \,\&\, S_1 \text{ or } A_2 \,\&\, S_2$$
$$S_4 = A_1 \,\&\, S_2 \text{ or } A_2 \,\&\, S_1$$

Similarly, the above identities hold for Newcomb's example, with which I began, if one lets

S_1 = The money is in the second box.
S_2 = The money is not in the second box.
S_3 = The being predicts your choice correctly.
S_4 = The being incorrectly predicts your choice.
A_1 = You take only what is in the second box.
A_2 = You take what is in both boxes.

[7] State S is not probabilistically independent of actions A and B if prob (S obtains/A is done) $\neq$ prob (S obtains/B is done).

[8] In Newcomb's predictor example, assuming that 'He predicts correctly' and 'He predicts incorrectly' are each probabilistically independent of my actions, then it is

not the case that 'He puts the money in' and 'He does not put the money in' are each probabilistically independent of my actions.

Usually it will be the case that if the members of the set of exhaustive and exclusive states are each probabilistically independent of the actions A_1 and A_2, then it will not be the case that the states equivalent to our contrived states are each probabilistically independent of both A_1 and A_2. For example, suppose prob (S_1/A_1) = prob (S_1/A_2) = = prob (S_1); prob (S_2/A_2) = prob (S_2/A_1) = prob (S_2). Let:

$$S_3 = A_1 \,\&\, S_1 \text{ or } A_2 \,\&\, S_2$$
$$S_4 = A_1 \,\&\, S_2 \text{ or } A_2 \,\&\, S_1$$

If prob $(S_1) \neq$ prob (S_2), then S_3 and S_4 are not probabilistically independent of A_1 and A_2. For prob (S_3/A_1) = prob (S_1/A_1) = prob (S_1), and prob (S_3/A_2) = prob (S_2/A_2) = prob (S_2). Therefore if prob $(S_1) \neq$ prob (S_2), then prob $(S_3/A_1) \neq$ prob (S_3/A_2). If prob (S_1) = prob $(S_2) = 1/2$, then the possibility of describing the states as we have will not matter. For if, for example, A_1 can be shifted around so as to dominate A_2, then before the shifting it will have a higher expected utility than A_2. Generally, if the members of the set of exclusive and exhaustive states are probabilistically independent of both A_1 and A_2, then the members of the contrived set of states will be probabilistically independent of both A_1 and A_2 only if the probabilities of the original states which are components of the contrived states are identical. And in this case it will not matter which way one sets up the situation.

[9] Note that this procedure seems to work quite well for situations in which the states are not only not probabilistically independent of the actions, but are not logically independent either. Suppose that a person is asked whether he prefers doing A to doing B, where the outcome of A is /p if S_1 and r if S_2/ and the outcome of B is /q if S_2 and r if S_1/. And suppose that he prefers p to q to r, and that S_1 = I do B, and S_2 = I do A. The person realizes that if he does A, S_2 will be the case and the outcome will be r, and he realizes that if he does B, S_1 will be the case and the outcome will be r. Since the outcome will be r in any case, he is indifferent between doing A and doing B. So let us suppose he flips a coin in order to decide which to do. But given that the coin is fair, it is now the case that the probability of $S_1 = 1/2$ and the probability of $S_2 = 1/2$. If we mechanically started to compute the expected utility of A, and of B, we would find that A has a higher expected utility than does B. For mechanically computing the expected utilities, it would turn out that the expected utility of A = $= 1/2 \times u(p) + 1/2 \times u(r)$, and the expected utility of $B = 1/2 \times u(q) + 1/2 \times u(r)$. If, however, we use the conditional probabilities, then the expected utility of A = prob $(S_1/A) \times u(p)$ + prob $(S_2/A) \times u(r) = 0 \times u(p) + 1 \times u(r) = u(r)$. And the expected utility of B = prob $(S_2/B) \times u(q)$ + prob $(S_1/B) \times u(r) = 0 \times u(q) + 1 \times u(r) = u(r)$. Thus the expected utilities of A and B are equal, as one would wish.

[10] This position was suggested, with some reservations due to Newcomb's example, in Robert Nozick, *The Normative Theory of Individual Choice*, *op. cit.* It was also suggested in Richard Jeffrey, *The Logic of Decision*, *op. cit.*

[11] I should mention, what the reader has no doubt noticed, that the previous *example* is not fully satisfactory. For it seems that preferring the academic life to the athlete's life should be as strong evidence for the tendency as is choosing the academic life. And hence P's choosing the athlete's life, though he prefers the academic life, on expected utility grounds does not seem to make it likely that he does not have the tendency. What the example seems to require is an inherited tendency to decide to do A which is such that (1) The probability of its presence cannot be estimated on the basis of the person's preferences, but only on the basis of knowing the genetic make-up of his parents, or knowing his actual decisions; and (2) The theory about how the tendency operates yields the result that it is unlikely that it is present if the person decides not to do A in the example-situation, even though he makes this decision on the basis of the stated expected utility grounds. It is not clear how, for this example, the details are to be coherently worked out.

[12] That is, the Dominance Principle is legitimately applicable to situations in which $\sim (\exists S)\,(\exists A)\,(\exists B)$ [prob (S obtains/A is done) $\neq$ prob (S obtains/B is done)].

[13] The other eleven possibilities about the states are:

	Already fixed and determined		Not already fixed and determined	
	probabilistically independent of the actions	not probabilistically independent of the actions	prob. ind. of the actions	not prob. ind. of the actions
(1)	some	some	some	some
(2)	some	some	some	none
(3)	some	some	none	some
(4)	some	some	none	none
(5)	some	none	some	some
(6)	some	none	some	none
(7)	some	none	none	some
(8)	all	none	none	none
(9)	none	some	some	some
(10)	none	some	some	none
(11)	none	some	none	some

[14] Unless it is possible that there be causality or influence backwards in time. I shall not here consider this possibility, though it may be that only on its basis can one defend, for some choice situations, the refusal to use the dominance principle. I try to explain later why, for some situations, even if one grants that there is no influence back in time, one may not escape the feeling that, somehow, there is.

[15] Cf. R. Duncan Luce and Howard Raiffa, *Games and Decisions*, John Wiley & Sons, New York, 1957, pp. 94–102.

[16] Almost certainty$_1$ > almost certainty$_2$, since almost certainty$_2$ is some function of the probability that brother I has the dominant action gene given that he performs the dominant action (= almost certainty$_1$), and of the probability that brother II does the dominant action given that he has the dominant action gene.

[17] In choosing the headings for the rows, I have ignored more complicated possibilities, which must be investigated for a fuller theory, e.g., some actions influence which state obtains and others do not.

[18] I here consider only the case of two actions. Obvious and messy problems for the kind of policy about to be proposed are raised by the situation in which more than two actions are available (e.g., under what conditions do pairwise comparisons lead to a linear order), whose consideration is best postponed for another occasion.

[19] See R. Duncan Luce and Howard Raiffa, *op. cit.*, pp. 275–298 and the references therein; Daniel Ellsberg, 'Risk, Ambiguity, and the Savage Axioms', *Quarterly Journal of Economics* **75** (1961), 643–669, and the articles by his fellow symposiasts Howard Raiffa and William Feller.

[20] If the distinctions I have drawn are correct, then some of the existing literature is in need of revision. Many of the writers might be willing to just draw the distinctions we have adumbrated. But for the specific theories offered by some personal probability theorists, it is not clear how this is to be done. For example, L. J. Savage in *The Foundations of Statistics*, John Wiley & Sons, New York, 1954, recommends unrestricted use of dominance principles (his postulate *P2*), which would not do in case (I). And Savage seems explicitly to wish to deny himself the means of distinguishing case (I) from the others. (For further discussion, some of which must be revised in the light of this paper, of Savage's important and ingenious work, see Robert Nozick, *op. cit.*, Chapter V.) And Richard Jeffrey, *The Logic of Decision*, *op. cit.*, recommends universal use of maximizing expected utility relative to the conditional probabilities of the states given the actions (see footnote 10 above). This will not do, I have argued, in cases (III) and (IV). But Jeffrey also sees it as a special virtue of this theory that it does not utilize certain notions, and these notions look like they might well be required to draw the distinctions between the different kinds of cases.

While on the subject of how to distinguish the cases, let me (be the first to) say that I have used without explanation, and in this paper often interchangeably, the notions of influency, affecting, etc. I have felt free to use them without paying them much attention because even such unreflective use serves to open a whole area of concern. A detailed consideration of the different possible cases with many actions, some influencing, and in different degrees, some not influencing, combined with an attempt to state detailed principles using precise 'influence' notions undoubtedly would bring forth many intricate and difficult problems. These would show, I think, that my quick general statements about influence and what distinguishes the cases, are not, strictly speaking, correct. But going into these details would necessitate going into these details. So I will not.

[21] Though perhaps it explains why I *momentarily* felt I had succeeded too well in constructing the vaccine case, and that perhaps one *should* perform the non-dominant action there.

[22] But it also seems relevant that in Newcomb's example not only is the action referred to in the explanation of which state obtains (though in a nonextensional belief context), but also there is another explanatory tie between the action and the state; namely, that both the state's obtaining, and your actually performing the action are both partly explained in terms of some third thing (your being in a certain initial state earlier). A fuller investigation would have to pursue yet more complicated examples which incorporated this.

THE MEANING OF TIME*

I. INTRODUCTION

Studies of time by scientists have often been concerned with the multifaceted problems of measuring time intervals in atomic, geophysical, biological, and astronomical contexts. It has been claimed that in addition to exhibiting measurable intervals, time is characterized by a *transiency* of the present, which has often been called 'flux' or 'passage'.

Indeed, it has been maintained that '*the passage of time* ... is the very essence of the concept'.[1] I therefore wish to focus my concern with the meaning of time on the credentials which this transiency of the present can claim from the point of view of current physical theories.

In the common-sense view of the world, it is of the very essence of time that events occur now, or are past, or future. Furthermore, events are held to change with respect to belonging to the future or the present. Our commonplace use of tenses codifies our experience that any particular present is superseded by another whose event-content thereby 'comes into being'. It is this occurring *now* or coming into being of previously future events and their subsequent belonging to the past which is called 'becoming' or 'passage'. Thus, by involving reference to *present* occurrence, becoming involves more than mere occurrence at various serially ordered clock times. The past and the future can be characterized as respectively before and after the present. Hence I shall center my account of becoming on the status of the present or now as an attribute of events which is encountered in *perceptual* awareness.

II. THE ISSUE OF THE MIND-DEPENDENCE OF BECOMING

Granted that becoming is a prominent feature of our temporal awareness, I ask: *must* becoming therefore also be a feature of the order of physical events *independently* of our awareness of them, as the common-sense view

N. Rescher et al. (eds.), Essays in Honor of Carl G. Hempel.