Newcomb's Perfect Predictor

DON HUBIN
OHIO STATE UNIVERSITY
and GLENN ROSS
FRANKLIN AND MARSHALL COLLEGE

Since Robert Nozick (1970) first published Newcomb's problem, much has been written of it. Most of the literature has dealt primarily with the case in which the predictor is assumed to be highly reliable but not infallible. As the problem is typically formulated, traditional evidential decision theories, such as Jeffrey's (1965) recommend taking one box. Causal decision theories, such as those developed by Gibbard and Harper (1981), Skyrms (1980, 1982), Lewis (1981), and Sobel (1977) favor taking both boxes. Recently, refined evidential decision theories have been offered by Eells (1981) and Jeffrey (1981) which are intended to prescribe a two-box solution. We think it clear that one should take two boxes and prefer the causal decision theorists' justification of this choice. Our concern here is with the case in which the predictor is not merely reliable, but perfect. In this case, causal decision theories are less satisfactory, for they continue to recommend choosing both boxes, and this recommendation is not so intuitive.[1] The infallibility of the predictor changes intuitions, because it seems that taking one box guarantees that you become a millionaire, while taking both guarantees that you do not. Gibbard and Harper claim that this appearance is based on mistaking truth-functional conditionals for counterfactuals. We shall argue, though, that the counterfactuals leading to the one-box solution have independent support. As a result, the argument for taking one box is cogent in the infallible case. In some formulations of the puzzle, so is the argument for taking two. In these cases, the problem lies not in either of these competing arguments but in the puzzle itself.

Perfect predictor versions of Newcomb's problem can be interpreted in various ways, but on plausible interpretations the prob-

lem is impossible because it is over-constrained. Either answer can be shown to be correct by attending to certain constraints and ignoring others. Hence controversy between one-boxers and two-boxers persists because the reasoning of each side depends upon paying selective attention to certain constraints on the puzzle. In order to establish this claim, we must begin by examining the general constraints on a practical decision problem, and the way in which the specific constraints on Newcomb's problem function to delimit the range of possible solutions to the puzzle.

I.

When presented with a puzzle, one is asked to assume that certain conditions obtain. These conditions function as constraints upon what counts as an adequate solution. Newcomb's problem asks what one ought to do. We shall call such problems 'practical problems'. In contrast to the practical problems faced in everyday life, Newcomb's puzzle is merely hypothetical. Practical problems, generally, are to be distinguished from theoretical problems, which ask what is the best supported hypothesis given certain evidence. Theoretical problems too, can be actual or hypothetical.

There are constraints on decision problems which arise simply because they are practical in nature; others define the specific problem. We shall discuss generic constraints on decision problems first and then turn to the role which specific puzzle conditions play in a hypothetical practical problem.

Generic Constraints. In order to solve a practical decision problem, one must consider the truth of certain subjunctive conditionals. This is not surprising. When confronted by a practical problem, what could be more natural than to ask oneself, 'What would [or might] happen were I to do that?' and 'Would that still happen even were I to do this instead?' What is surprising is not that subjunctive reasoning is crucial to practical reasoning, but that until recently this fact has not been reflected in formal decision theories.

The subjunctive conditionals relevant to decision have antecedents asserting that an action is performed in a given state of the world and consequents asserting that certain outcomes result. We shall symbolize this conditional connective by '$>$'.

There are certain counterfactuals of this form that strike some as true but cannot serve as a basis for decision. Suppose one of your weak and cowardly friends boasts that were he now to fight the current heavyweight champion, he would win. Challenged, he

is quick to explain. He would, he claims, be psychologically incapable of climbing into the ring with a professional boxer unless he were certain that he could not lose. Even after hearing his reasoning, some might be inclined to reject his claim flat out; others (e.g., Lewis, 1979) might admit that his claim is true under some (deviant) interpretation. But even if this is correct, your friend's claim cannot provide a reason to fight.

There are two further constraints on the counterfactuals relevant to decision-making. First, the outcomes of the acts must be possible. This requirement entails that the states and the acts must be compossible, which in turn entails that the acts and the states must be possible. The fact that an act would have the best consequences were the world a way it could not be can hardly be a reason to prefer that act. Similarly, the actions considered must be possible.

The decision situation itself must be one in which there is at least one state of the world which has more than one action compatible with it. States of the world are assumed to be beyond an agent's control. Hence, those states with which only one action is compatible are irrelevant to one's choice; if one knows that a state is of this sort, one may be justifiably fatalistic about it. If *all* states are of this sort, there is no decision problem at all.

This last requirement is not entirely uncontroversial. One might substitute the epistemic requirement that the agent not *know* that only one action is compatible with each state of the world. But this seems to confuse having a decision problem with having reason to suppose that you have one. Consider Uri, who does not know that he cannot alter the course of Pluto by concentrated exertion of his will; though we are to suppose that he cannot. Uri may believe that the question of whether to alter the course of Pluto by mental effort poses a decision problem for him. Depending on his evidence, this may be a reasonable belief. Nevertheless, Uri does not have a genuine decision problem. He may decide not to alter Pluto's course, thinking that he can, but he really has no choice.[2] He could not alter the planetary course even were he to try. The appearance of a decision problem is illusory.

Specific Constraints. In addition to generic constraints there are constraints imposed by the particular features of a puzzle. These puzzle conditions influence our assessment of the relevant subjunctive conditionals. There are (at least) two views about the nature of this influence. The first view holds that the puzzle conditions simply describe a situation which we are to suppose does, but need not, obtain. On this view, the questions we are to ask

ourselves in solving the problem have the form: 'If the puzzle conditions were to obtain, then if I were to do so-and-so, what would happen?' We should view each of the relevant counterfactuals as being the consequent of a larger counterfactual, the antecedent of which consists of the puzzle conditions. Accordingly, we shall call this view of puzzle constraints 'the exported view'. Given a possible world semantics, this view requires world-hopping. Puzzle conditions direct us to the closest world(s) in which these conditions obtain. From these worlds, we evaluate the embedded counterfactuals by considering neighboring worlds in which the action is performed.

We could view puzzle conditions in another way, though. Perhaps the relevant question to ask is: 'If I were to do so-and-so when the puzzle conditions hold, what would happen?' Puzzle conditions would then be viewed as being conjoined to the antecedents of each of the relevant counterfactuals. Call this 'the imported view'. Characterized semantically, the relevant outcomes are those obtaining in the closest world(s) at which the puzzle conditions hold and the action is performed. As a consequence, for any action A, state S, outcome O, and puzzle condition C: if C and $(A \ \& \ S \ \& \ C) > O$, then '$(A \ \& \ S) > O$' can be assumed for purposes of solving the puzzle. (Italics abbreviate sentences stating that the action, state or outcome denoted by the corresponding letter in roman type obtains.)

It seldom matters which of these two views is adopted. Unless the action you are contemplating conflicts with the puzzle conditions, the two views will yield the same judgments. But in those cases where exportation fails for counterfactuals the two views pull apart.

If we are to decide which of these views is correct, we must look to the situations which divide them. The clearest cases of conflict arise when one of the puzzle conditions simply rules out one of the (normally possible) actions. Suppose you are given the following problem: You are the president of a superpower. Tomorrow (for some reason that the puzzle conditions may not state), you will press a button which is currently wired to launch a nuclear warhead at a hostile nation—an action which would lead to an undesirable war. What should you do?

There are various answers you could give. You could say that you ought to disconnect the button today and see that it is not reconnected tomorrow. Perhaps you should have the missile disarmed, just in case. There are many possibilities, but one response which will not do as an answer to the problem is that you should

not press the button tomorrow. True as this claim is, it does not solve the problem. Yet the exported view of puzzle conditions cannot rule this out. Consider whether the conditional, 'Were I to avoid pressing the button tomorrow, then the best consequences would result', is true at the closest world(s) in which you do press the button tomorrow. This certainly appears to be true. So, according to the exported view, the answer to the problem of what you should do, given that you are going to press the button tomorrow, is that you ought to avoid pressing the button tomorrow.

On the imported view, however, we consider as the outcome of an action what would happen were you to perform the action (in a given state) when the puzzle conditions obtain. Of course, it is not possible both to press the button and to avoid pressing it. Any conditional with an impossible antecedent is vacuously true. The set of consequents which would be true if such an antecedent were true is inconsistent. Since, on the imported view, this set describes the outcome of pressing the button, the outcome of pressing the button is not possible. This is why it is no answer to the problem to say 'I should not press the button tomorrow.'

The imported view is also preferable when an action *could* have been performed were the puzzle conditions to obtain, even though it would not have been. Suppose that you are merely told that someone will press the button tomorrow. The exported view still yields the answer that you should simply not press the button. For, given your privileged access to the button, in the closest worlds in which someone presses the button, it is you who presses it. In these worlds, were you not to press the button the best consequences would ensue. The imported view, however, does not recommend mere forbearance on your part. In the closest worlds in which someone other than you presses the button, the outcome is dire. Again, when the two views yield different answers, it is the imported view's that is correct.

If our concern were purely theoretical—if we merely wished to know if certain counterfactuals were true on certain assumptions—then the exported view of constraints would be correct. But our concern is practical, not theoretical. We are concerned with what ought to be done given certain conditions. Only the imported view of constraints yields the proper answer to the above (hypothetical) practical problems.

II.

Let us return to the perfect predictor version of Newcomb's problem. The puzzle conditions stipulate that the predictor never makes

a mistake. On the imported view, the outcome of any act considered is the outcome of performing that act when the predictor has correctly predicted it. As a consequence, it makes no difference whether the puzzle stipulates that the predictor's infallibility is coincidental or a matter of physical or logical necessity. Indeed, it only matters that the puzzle guarantee that the predictor is correct in the particular instance. In any case, one could reason: 'If I were to take both boxes when the predictor is not mistaken then I would receive but $1000. If I were to take only the one box when it was predicted that I would, I would receive a million dollars (M). Since I prefer a million dollars, I should take only the one box'. This argument has considerable appeal.

The argument can be clarified by considering the following matrix:

	S_1	S_2	S_3	S_4
	$(A_1 > H_1)$ & $(A_2 > H_2)$	$(A_1 > H_1)$ & $(A_2 > H_1)$	$(A_1 > H_2)$ & $(A_2 > H_2)$	$(A_1 > H_2)$ & $(A_2 > H_1)$
A_1	M	M	O	O
A_2	1000	M + 1000	1000	M + 1000

A_1: You take the one box only.
A_2: You take both boxes.
H_1: A_1 (your taking only one box) is predicted.
H_2: A_2 (your taking both boxes) is predicted.

It is not difficult to show that the states are pairwise exclusive. Let i, j ϵ {1,2,3,4}. Suppose S_i & S_j. If $|j - i| > 1$, then $A_1 > H_1$ and $A_1 > H_2$. Since $H_1 \equiv \sim H_2$, $A_1 > (H_1 \ \& \sim H_1)$. Thus, $\sim \Diamond A_1$. Since $\Diamond A_1$, $\sim \Diamond (S_i \ \& \ S_j)$. If $|j - i| = 1$, then $A_2 > H_1$ and $A_2 > H_2$. Thus, $\sim \Diamond A_2$, and since $\Diamond A_2$, $\sim \Diamond (S_i \ \& \ S_j)$. To ensure exhaustivity, we could include '$\sim S_1$ & $\sim S_2$ & $\sim S_3$ & $\sim S_4$' (= S_5), which is obviously inconsistent with any other state.

Suppose that the predictor is infallible. Then $\sim(A_1 \ \& \ H_2)$ and $\sim(A_2 \ \& \ H_1)$. Equivalently, $(A_1 \supset H_1)$ and $(A_2 \supset H_2)$. Since $\prod_{i=1}^{5} \Box[(A_1 \ \& \ S_i \ \& \ (A_1 \supset H_1)) \supset H_1]$ and $\prod_{i=1}^{5} \Box[(A_2 \ \& \ S_i \ \& \ (A_2 \supset H_2)) \supset H_2]$, $\prod_{i=1}^{5} [(A_1 \ \& \ S_i \ \& \ (A_1 \supset H_1)) > H_1]$ and $\prod_{i=1}^{5} [(A_2 \ \& \ S_i \ \& \ (A_2 \supset H_2)) > H_2]$. ('$\prod_{i=1}^{n}$' symbolizes generalized conjuction and '$\sum_{i=1}^{n}$' expresses generalized disjunction.) On the imported view, it follows that $\prod_{i=1}^{5} [(A_1 \ \& \ S_i) > H_1]$ and $\prod_{i=1}^{5} [(A_2 \ \& \ S_i) > H_2]$. By employing a principle of dilemma, '((P > R) & (Q > R)) $\supset$ ((P v Q) > R)', it follows that $(\sum_{i=1}^{5} (A_1 \ \& \ S_i)) > H_1$ and $(\sum_{i=1}^{5} (A_2 \ \& \ S_i)) > H_2$. Substitution of equivalent antecedents yields '$(A_1 \ \& \sum_{i=1}^{5} S_i) > H_1$' and '$(A_2 \ \& \sum_{i=1}^{5} S_i) > H_2$'. Given that $\Box \sum_{i=1}^{5} S_i$, substituting equivalent antecedents produces '$A_1 > H_1$' and '$A_2 > H_2$', which is S_1. Given

pairwise exclusivity, we have S_1 as the only possible state. But if S_1 is the only possible state, A_1 clearly dominates. So the argument for taking just one box is sound.

There is an equally convincing argument that you should take both boxes. Recall that your action does not affect the contents of the boxes. Nozick (1970, pp. 131-132) in the original presentation of the problem assumes that '...the actions or decisions to do the actions do not affect, help bring about, influence, etc., *which* state obtains...' If you know that nothing you can do will affect what is in the boxes, then whatever the contents of the boxes, you are better off taking both. The assumption that the predictor is infallible seems to have done nothing to undermine the cogency of the argument for taking both boxes.

This appearance is correct. Nozick's stipulation amounts to a constraint on the puzzle that the contents of the boxes be independent of your actions.[3] The independence of state and action entails that either $(A_1 > H_1)$ & $(A_2 > H_1)$ or $(A_1 > H_2)$ & $(A_2 > H_2)$. Let us call this the fixity constraint. This constraint is equivalent to the disjunction of S_2 and S_3. Given pairwise exclusivity, S_1, S_4, and S_5 are inconsistent with this puzzle condition. Thus, S_2 and S_3 are the only possible states and A_2 dominates. So the argument for taking both boxes is sound.

The resolution of the problem is straightforward. The puzzle is overconstrained. Suppose that we have the infallibility constraint, '$(A_1 \supset H_1)$ & $(A_2 \supset H_2)$'; the fixity constraint, '$((A_1 > H_1)$ & $(A_2 > H_1))$ v $((A_1 > H_2)$ & $(A_2 > H_2))$'; and the generic constraint, '$\Diamond A_1$ & $\Diamond A_2$'. We have shown above that, on the imported view, the infallibility constraint entails that $(A_1 > H_1)$ & $(A_2 > H_2)$. Suppose that $A_1 > H_2$. It follows that $A_1 > (H_1$ & $H_2)$, or equivalently, that $A_1 > (H_1$ & $\sim H_1)$. But then, $\sim \Diamond A_1$. Therefore, $\sim (A_1 > H_2)$. Suppose that $A_2 > H_1$. Then $A_2 > (H_1$ & $H_2)$, or equivalently, $A_2 > (H_1$ & $\sim H_1)$. Thus $\sim \Diamond A_2$, which contradicts the above. Therefore, $\sim (A_2 > H_1)$. Consequently, $\sim (A_1 > H_2)$ and $\sim (A_2 > H_1)$, contradicting the fixity constraint.

If it is given that the predictor is infallible but not that the contents of the boxes are fixed independent of your decision, your choice is clear. You should take one box. If the puzzle stipulates that your choice will not affect the contents of the boxes and the predictor is fallible, your choice is again clear. Take both boxes. But in the present case we are asked to assume that both constraints hold. It follows from these constraints that at least one act is impossible. If the decision problem is to be genuine, then one of the constraints must be given up.

Thus, neither a one-box nor a two-box solution to the perfect predictor puzzle is satisfactory. In solving a hypothetical practical problem, the stipulated puzzle conditions must be assumed to hold no matter what act is performed. Once this role of puzzle conditions

is understood, the force of both the one-box and the two-box arguments can be appreciated. The solution lies not in choosing between the two arguments but in choosing between different coherent formulations of Newcomb's problem.[4]

References

Eells, Ellery. "Causality, Utility and Decision," *Synthese* 48 (1981): 295—329.

Gibbard, Allan and Harper, William L. "Counterfactuals and Two Kinds of Expected Utility," in W. L. Harper, R. Stalnaker, and G. Pearce, eds., *Ifs* (Dordrecht: D. Reidel Publishing Co., 1981): 153—190.

Jeffrey, Richard C., *The Logic of Decision* (New York: McGraw-Hill, 1965).

———, "The Logic of Decision Defended," *Synthese* 48 (1981): 473—492.

Lewis, David. "Counterfactual Dependence and Time's Arrow," *Nous* 13 (1979): 455—466.

———, "Causal Decision Theory," *Australasian Journal of Philosophy* 59 (1981): 5—30.

Nozick, Robert. "Newcomb's Problem and Two Principles of Choice," in Nicholas Rescher, ed., *Essays in Honor of Carl G. Hempel* (Dordrecht: D. Reidel Publishing Co., 1970): 114—146.

Skyrms, Brian. *Causal Necessity* (New Haven: Yale University Press, 1980).

———, "Causal Decision Theory," *Journal of Philosophy* 79 (1982): 695—711.

Sobel, Jordan Howard. "Newcomb's Problem and the Theory of Rational Agency," unpublished manuscript, 1977.

Notes

[1] An anonymous referee for this journal has pointed out that at least on Skyrms' (1980) version of causal decision theory, the recommendation is highly indeterminate. The indeterminacy arises from the fact that the K-expectation of an act involves conditional probabilities in which the condition has zero probability, (e.g., the probability that you get $0 conditional on the predictors predicting that you take both boxes and your taking just one). Obviously, these conditional probabilities cannot be defined in the standard way.

[2] We often say that a person decided to do something even though he had no decision problem (i.e., when only one alternative was possible). This is partly because decision is viewed as a mental act which requires that the agent *believe* he has a genuine choice but not that this belief be true. It is also partly because of the connection between one's deciding to do an act and that act being voluntary. Clearly, an act can be voluntary even if one has no choice. But it does not follow from the fact that someone made a decision and acted voluntarily that he had a decision problem.

[3] It is possible to interpret Nozick's statement of Newcomb's problem in various ways. Nozick admits (1970, p. 146) that his use of 'influence' and 'affect' is imprecise. If we do not interpret this requirement as ruling out S_1, then, given that the predictor is infallible the solution to the puzzle is clear: take only one box. It is difficult, however, to understand how the fixity constraint could fail without one's choice influencing the contents of the boxes.

[4] We are indebted to Keith Lehrer, Alvin Plantinga, and Terry Horgan for their many helpful comments on an earlier version of this paper. We are also grateful to Franklin and Marshall College for a research grant in support of this project.