Newcomb's Paradox Revisited

by MAYA BAR-HILLEL and AVISHAI MARGALIT

This paper attempts to provide a solution to the Newcomb Problem, which was first presented in Nozick [1969]. The author suggested there a solution of his own, with which he admitted to being dissatisfied, and invited further comments that might 'enable [Nozick] to stop returning periodically to [the paradox]' (*op. cit.* p. 143). We found the paradox every bit as intriguing as Nozick did, and hope that our solution can restore his peace of mind.

Suppose you are playing a game with a Being whom you believe to possess extraordinary predictive powers. The game proceeds as follows: Before you are two boxes. In one you can plainly see $1,000. The other is covered, so you cannot see what it contains. But you know that the Being has put into it either a million dollars or nothing, depending on what he had predicted that you will do come your turn to play. You have a choice between two actions: taking what is in both boxes, or taking what is in the covered box only. You know, however, that the Being played his move as follows: if he predicted that you will take both boxes, he has left the covered box empty; if he predicted that you will take the covered box only, he has put a million dollars in it. You are not allowed to use a chance device to determine your choice. You have enormous confidence in the Being's ability to predict your actions, and you know that he has correctly predicted all choices of all players who have played this game with him to date (in fact, he has predicted correctly all choices that *you* have made in this game in some previous 'warming up' trials, played for points, say, rather than money). The Being has just now (or an hour ago, or a year ago . . .) made his prediction and played his move. It is now your turn. What will you do?

Both actions can be argued for very persuasively and intuitively.

An argument for taking only the covered box might run as follows: You firmly believe that whatever you ultimately decide to do, the Being has probably foreseen. In other words, you are, for some reason, almost sure that if you will take both boxes you will end up with $1,000, whereas if you will take just the covered box, you will end up with a million dollars. It seems a shame to sacrifice a million dollars for a thousand.

An argument for taking both boxes might run as follows: The Being has already played his move. Whatever you now do will not affect the amount of

Received 12 *April* 1972

money in the covered box. Regardless of whether the Being did or did not put a million dollars in the covered box, you stand to receive \$1,000 more by taking both boxes than by taking the covered one alone. It seems a shame not to take advantage of the fact that the Being played before you, and you play second.

Further elaborations of these arguments can be found in Nozick [1969].

Our solution is based on a game theoretical approach, following that of Nozick. Below is a representation of the game in matrix form.

		Being	
		P_1	P_2
you	A_1	\$1,000,000	\$0
	A_2	\$1,001,000	\$1,000

where: P_1—Being predicts you will take only covered box
P_2—Being predicts you will take both boxes
A_1—You take only the covered box
A_2—You take both boxes,
and the cell entries are your payoffs.

The first argument above is based on the principle of maximising subjective expected utility (*SEU*). This principle states that one should select that action which leads to the highest expected utility. The (subjective) expected utility of an action is the sum of the products of the utility of the (mutually exclusive and exhaustive) outcomes it may lead to, by their probability given this action. Thus, in the present case, the *SEU* of A_1 is $p.u(\$1{,}000{,}000)+(1-p)\cdot u(\$0)$, and of A_2 is $q\cdot u(\$1{,}001{,}000)+(1-q)\cdot u(\$1{,}000)$, where $0<p,q<1$, p is close to 1 and q is close to 0. No doubt the reader's utility function yields a higher *SEU* for A_1 than for A_2.

The second argument above is based on the dominance principle. This principle states that if for every possible state of the world (or, alternatively move of your opponent) you are at least as well off by doing one act as by doing another, and even better off for some state (or move), then you should select the former act in preference to the latter. Thus, in the present case, the possible states are that there are a million dollars in the covered box, or that there are not, and in either case you are better off by doing A_2 than by doing A_1.

This apparent discrepancy between the recommendations of the *SEU* principle and the dominance principle results from an injudicious application of the dominance principle. Consider, for instance, the following example: Israel must decide whether to withdraw from its occupied territories or not, and Egypt must decide whether to declare war on Israel or not. Suppose the following matrix represents the possible payoffs to Israel on some ordinal utility scale.

		Egypt	
		war	peace
Israel	withdraw	0	2
	remain	1	3

Clearly, remaining in the occupied territories is the dominant strategy, since 1>0 and 3>2. Suppose, however, that you believe that with a high probability withdrawal will be conducive to peace while remaining in the territories will eventually lead to war. Then you might prefer to withdraw and end up in the 2 cell than to remain and end up in the 1 cell.

This example makes clear that the dominance principle loses its appeal when applied to situations where the states of the world (or the opponent's moves) are affected by the decision maker's actions, and its logic is overriding only when the states are independent of the actions, *i.e.* when the probability distribution over states of the world (matrix columns) is the same for all actions (matrix rows). Let us refer to this case as the unconditional case. In the unconditional case a dominant strategy, if one exists, coincides with the maximun *SEU* strategy, and in fact with the strategy recommended by any of the other principles of 'rational' choice. This constraint on the applicability of the dominance principle, though obvious and undisputed, has nevertheless rarely been explicitly stated in the literature (see, however, Jeffrey [1965], pp. 8–10). Since in our problem the states of the world are not probabilistically independent of actions, there is no case for the dominance principle. Furthermore, it can be proven that for every finite partition of the world into states that are not probabilistically independent of actions, there is a refinement of this partition into an unconditional case. We show this for a simple example. The general case is discussed and proved in Krantz, Luce, Suppes and Tversky [1971].

Let the following be a payoff matrix:

		states	
		S_1	S_2
actions	A_1	1/4 0	3/4 4
	A_2	2/3 2	1/3 6

where A_1 and A_2 are the two possible actions; S_1 and S_2 is a partition of the world; 0, 2, 4 and 6 are the possible outcomes on some ordinal utility scale; and the fractions give the probability of the respective cells for a given action. Obviously, we do not have probabilistic independence in this case. We now show how to refine this partition to obtain another one, under which states *are* probabilistically independent of actions.

Suppose that which state of the world obtains is determined by randomly drawing chips from a poker bag. The poker bag contains twelve chips: three are red, five are green, and four are yellow. Drawing a red chip always entails S_1; drawing a yellow chip always entails S_2; but drawing a green chips entails S_2 if the action is A_1, and S_1 if the action is A_2. Let us now define S'_1 as the state defined by drawing a red chip; S'_3 as the state defined by drawing a yellow chip; and S'_2 as the state defined by drawing a green chip. This yields the following matrix:

states

	S'_1	S'_2	S'_3
A_1	1/4 0	5/12 4	1/3 4
A_2	1/4 2	5/12 2	1/3 6

The new states are probabilistically independent of your actions, yet obviously every action now leads to its possible outcomes with the same probability as before. Note, however, that A_2 is no longer the dominant strategy, since, for S'_2, A_1 is the better action. Although [S_1,S_2] may be a more natural partition of the world in some sense, it does not enjoy any formal privilege over [S'_1, S'_2, S'_3]. Sometimes, in fact, a new partition, arrived at formally, may be interpretable in some meaningful way. Suppose, for instance, for simplicity's sake, that in our problem $p=1-q$, *i.e.* the probability for getting a million dollars when taking one box is equal to the probability of getting a thousand dollars when taking both boxes.

We can now take that probability to be the Being's likelihood of predicting correctly, and repartition of the world into states not merely independent of your actions, but also with a meaningful interpretation, as follows:

		Being	
		D_1	D_2
you	A_1	p $1,000,000	$1-p$ $0
	A_2	p $1,000	$1-p$ $1,001,000

where A_1 and A_2 are as above; D_1 and D_2 are the possibilities that the Being predicts right or wrong, respectively; and p is the probability of the Being predicting correctly.

The problem, when thus presented, is seen to be not really one of deciding between two principles of choice. However, doing away with the case for the dominance principle does not seem to do away with the paradoxical nature of the problem. One is left with the uneasy feeling that choosing A_1, though defensible on game-theoretical grounds, is somehow 'wrong' in a very fundamental way. That it is, in fact, tantamount to subscribing to backwards causality. In other words, choosing A_1 rather than A_2 seems essentially to be justified by the fact that thereby a very high, rather than very low, probability can be assigned to a certain desired event, namely that the Being put a million dollars into the covered box. But to go about assigning probabilities to past events, unaffected by present events, in what appears to be a completely arbitrary, *ad hoc*, and wilful fashion is, to say the least, highly unorthodox and more than a little unsettling. Before answering this objection, we would like to criticise Nozick's way of dealing with it.

Nozick persists in presenting the dilemma embodied in this situation as one of choice between two decision principles. He therefore proceeds to distinguish between probabilistic independence and logical independence, where states are logically independent of actions if the actions 'do not affect, help bring about, influence, *etc.*' (*op. cit.* p. 132) which state obtains, and then recommends as a general policy that where states are logically independent of actions, even if they are not probabilistically independent of them, 'one should perform the dominant action' (*op. cit.* p. 132).

This solution, though relieving one of any suspicion of adherence to backwards causality, poses its own threats to one's image as a rational decision maker. (*i*) It puts you in the extremely uncomfortable position of acting 'against what you would rationally want to bet on' (*op. cit.* p. 116).

(*ii*) It implies that even if you believed that choosing A_2 would *surely* mean loss of the million dollars, *i.e.* even when you believe that the Being has perfect predictive powers, you should still choose A_2 whenever the Being played his move before you, for reasons of 'logical independence'. This is a conclusion which Nozick himself is unwilling to draw, which leads him both into an inconsistency with his own maxim, and causes him to recommend a different strategy for the case when the Being's probability of correct prediction is 1 and for the case when it is $n/n+1$, for an arbitrarily large n. The distinction between logical and probabilistic independence here is at best shaky.

The question now naturally arising is whether the paradox does not lie in the very assumption of such a Being itself. In others word, the question may seem to be not what strategy a rational decision maker should employ under the proposed circumstances, but whether a rational man could ever find himself in such circumstances to begin with. What kind of evidence would lead to such overwhelming faith in the predictive powers of any Being? Can any situation lead a rational man to simultaneously believe that the Being plays his move, irrevocably, prior to your move and yet that the probability of there being a million dollars in the covered box is different if you play one strategy than if you play another? Let us examine this question first psychologically, and then logically.

Since neither the funds nor the Being for a real-life tryout of this game are available to us, we urge the reader to follow us through the following thought experiment: Suppose you have volunteered to participate in a psychological experiment at the local university. Sitting behind a one-way screen, you watch many subjects play this game against an experimenter. Time and again, you see the experimenter put into a covered box a check made out either for a million dollars (play money, of course . . .) or zero dollars; you then see a subject entering the room, receiving instructions on the nature of the game (essentially by way of its payoff matrix, and possibly with statistics on how former players have fared), and then playing his move. You note down for each subject the amount he wins in the game; to your extreme surprise you soon realise that all those who come their turn had taken only the covered box had found the check in it to be made out for a million dollars, and all those who had taken both boxes had found their check to be for zero dollars. You are now summoned to play the game yourself. The experimenter had played his move, and it is your turn. What would you do?

First and foremost, we predict, what you would do is to discredit the evidence of your senses. You would suspect foul play. You would suspect that you are being taken. You would either tell yourself that the checks were, by some clever sleight of hand, tampered with or exchanged *after* the subjects had played their move (and you need not feel any more called upon to explain how this was done than if you had been watching rabbits being pulled out of a hat); or you might believe that the subjects were actually collaborators of the experimenter, preinstructed as to how to play, *etc.*; or you would just simply think you were imagining things. Anyway, you would play your move depending on the way you interpreted the goings on (if, indeed, you were still motivated to play it rationally . . .). In all probability you would *not*, however, be holding simultaneously to both beliefs, *i.e.* that the check was written out irrevocably prior to the subjects deciding what to play, and that the probability for subjects playing the two strategies in good faith is different (for the event of having the check written out to the sum of a million dollars). Indeed, to go back to our Being, each of the two premisses we are given would seem to be counterevidence to the other. That people who play A_1 find a different sum of money in the covered box than people who play A_2 would normally be taken as proof that the sums were put there after they made their choice; and that the money was put in the box before you choose would normally be taken as a guarantee that the probability for finding a million dollars in the box is the same whether you will now take both boxes or just one. That this is not so in our case, we claim, is maybe counterintuitive, but not logically contradictory.

Let us check our intuitions. The two seemingly incompatible premisses above are essentially equivalent to the single premiss that the Being has near perfect foreknowledge of which of two acts you will perform. Foreknowledge in itself is not a logical impossibility. If, for instance, the Being were playing not against yourself but against some robot, say, it would be perfectly plausible that he could predict how the robot would play. It does however, seem virtually impossible that *your* moves and choices are so predictable! After all, you, unlike the robot, have free will! But do you? Maybe you are only under an illusion of free will, like a robot may be programmed to be (for example, program him to make a different response than he would have otherwise every time he is told what his response will be . . .). In fact, our Being, if he existed, would contribute just the kind of evidence that would disprove your illusion that you can choose arbitrarily in the game, if you want to.

The paradox may now have been brought into clearer focus. To the extent that you refuse to take seriously the possibility of any conceivable circumstances that would lead you to believe in the existence of such a Being, the paradox for you becomes void. Prior evidence becomes immaterial to your present decision, and you would act on the 'sure thing'—

namely, take both boxes. Furthermore, even granted that such a Being could, logically speaking, exist, you might feel that its existence is such a major departure from a world in which you can trust your (tried and true) concepts of rationality, that you would not really want to test them as they are in such a world, and this again would render the paradox void. One may, however, go along and think this 'as if' situation to the limits which *are* afforded by our existing concepts of rationality. This is what we shall now proceed to do, and naturally it bears on the age old philosophical questions of what implications, if any, a deterministic point of view has to discussions of rationality, morality, personal responsibility, *etc.* In particular, in what sense can one recommend a certain course of action in a situation in which the course of action to be taken has actually been predetermined.

Suppose you are playing the following game: Someone is tossing a fair die, and after each toss you guess whether it came up on 6 or not. You receive a penny for each correct guess. Clearly, you will be maximising your expected gains by guessing not–6 on every toss. Now suppose that the die was tossed the previous day, the outcomes were noted down on a list, and you are now guessing, entry by entry, whether the number is 6 or not. Had you been able to obtain a copy of the list you could change your previous strategy to one that would ensure that you get a penny on each trial. But without such a list, you can *still* do no better than to guess not–6 on each trial! The fact that the order of 6's and not-6's is predetermined, given that you do not know what it is, does not affect your strategy. What this example serves to point out is that the mere knowledge that things are not what they seem to be does not necessarily supply you with an alternative strategy for dealing with them.

Thus, in our case, although the facts really imply that there is no free choice, the illusion of free choice persists, and you can do no better than to behave as if you do have free choice, *i.e.* 'deliberately' pick that strategy that seems to serve your interests best. The fatalistic claim that in a deterministic world it makes no difference which mode of behaviour one chooses because '*que sera sera*' is obviously false on statistical and phenomenological grounds, if no other. Furthermore, to follow the same argument, though you know that the Being plays before you, you nevertheless cannot do better than to play as if he plays after you. For you cannot outwit the Being except by knowing what he predicted, but you cannot know, or even meaningfully guess, at what he predicted before actually making your final choice. This feature, strange though it may seem, is built into the story by virtue of investing the Being with such astounding predictive powers, and of *behaviour*, no less, rather than just inclinations. (As a side remark, let us say that though we do not wish here to embark on speculations as to how such a Being might go about arriving at his predictions, it seems safe to say that it is *not* by a process that you could ever hope to master yourself. Since if you did, like the Being, know what you were going to do before doing it, you could intervene in the course of events, so to speak, and disprove that prediction by in fact behaving otherwise.)

Finally, note that the uneasiness created by the persistent feeling that choosing A_1 over A_2 amounts to arbitrarily choosing to assign a high rather than a low probability to a past event is shown by this analysis to be no more than an illusion itself; for if indeed your feeling that you are a free agent and can choose arbitrarily between the strategies is illusory, so, of course, is the feeling that probabilities are here being assigned in an arbitrary and *ad hoc* manner.

To sum up, we have tried to show that the Newcomb problem allows for just one, rather than two, 'rational' strategies (given that what you are trying to do is get as much money as you reasonably can. One could, of course, advocate the taking of both boxes on maximin grounds; *i.e.* assuring yourself of at least a thousand dollars.) This strategy is to take only the covered box. It is not justified by arguing that it *makes* the million dollars more likely to be in that box, although that is the way it appears to be, but because it is inductively known to correlate remarkably with the existence of this sum in the box, and though we do not assume a causal relationship, there is no better alternative strategy than to behave as if the relationship was, in fact, causal. The pragmatic nature of this argument is supplemented by other considerations which serve to show that any logical inconsistencies which seem to be attached to choosing A_1 rather than A_2 are due to the unintuitive implications of the existence of such a Being, but are not real contradictions, merely illusory ones.

Thus, though we began by analysing the problems along the lines set out by Nozick, we ended up with the opposite recommendation to his. Our solution manages to avoid the inconsistencies attached to Nozick's solution, but nevertheless, we hope, succeeds in meeting the objections that drove Nozick away from offering it himself. By proposing a rationalisation for a strategy that seemed to be a mere act of faith, we hope to convince the reader to take just the one covered box, and join the millionaires's club!

The Hebrew University of
Jerusalem

REFERENCES

JEFFREY, R. C. [1965]: *The Logic of Decision.*

KRANTZ, D. T., LUCE, R. D., SUPPES, P., TVERSKY, A. [1971]: *Foundations of Measurement, Volume 1.*

NOZICK, R. [1969]: 'Newcomb's Problem and Two Principles of Choice,' in N. Rescher *(ed.): Essays in Honour of Carl G. Hempel,* pp. 114–46.