Volume 11, Number 2, June 1987, pp. 197-213

Newcomb's Problem: The $1,000,000 Solution

KENT BACH
San Francisco State University
1600 Holloway Ave.
San Francisco, CA 94132
U.S.A.

The more you think about it, the more baffling Newcomb's Problem becomes. To most people, at first it is obvious which solution is correct (not that they agree on which one), but their confidence can be eroded easily. Only a puzzled few are torn between the two right from the start, and for years so was I. But at last, thanks to a certain meta-argument, one solution came to seem obvious to me. And yet, imagining myself actually faced with Newcomb's choice, I started to worry that I might experience just enough last-minute ambivalence to unsettle my confidence in that argument. Fortunately, I have found a strategy to ensure making the right choice when the chips are down.

Not only is Newcomb's Problem puzzling in its own right, it is philosophically significant. The appeal of both solutions reflects a conflict between two plausible conceptions of rational choice. In making a decision, should one consider all of its probabilistic consequences or only its causal consequences? Each conception has its supporters, but some philosophers find them both defensible and see no hope of resolving the conflict. I think the conflict can be resolved, at least in the context of Newcomb's Problem, by properly assessing the relevant counterfactual conditionals.

I The Problem and the Dilemma

There are two boxes, transparent and opaque. Box 1 visibly contains a thousand dollars ($1K), and there may or may not be a million dollars ($1M) in Box 2. You can take the contents either of both boxes (BOTH) or of Box 2 only (ONE). There is a certain super-predictor (PR)

who tries to anticipate your choice, placing $1M in Box 2 iff he predicts that you will take ONE. Knowing that your only motivation is wealth (his is truth), he gathers detailed information about you and plugs it into a high-powered psychological theory. PR is extremely reliable, as shown by his near-perfect record of identifying those who would choose ONE and those who would choose BOTH.[1] You know all this, he knows you know all this, and so on.

There are four possible outcomes: if you choose ONE, you get $1M if PR predicted ONE and $0 if he predicted BOTH; if you choose BOTH, you get $1M + $1K if PR predicted ONE and $1K if he predicted BOTH. You can guarantee yourself $1K by taking BOTH but if that was predicted, you won't become a millionaire. What should you do?

The Dilemma. People don't go into philosophy for the money—they seek truth, not wealth. So it is no surprise that philosophers writing on Newcomb's Problem generally endorse BOTH, even though they reckon this would net them but $1K. But pose Newcomb's Problem to non-philosophers and I bet you'll get the following result.

At first they will probably be impressed with PR's reliability and opt for ONE. Why try to outsmart PR, they'll say, and risk $1M just to grab an extra $1K? If PR has usually been right before, he is not likely to be wrong this time. Whether or not they are relying on an argument from expected utility, as Robert Nozick (1974) suggests in his survey of reader response to Martin Gardner's (1973) *Scientific American* column on Newcomb's Problem, at least we can say that each is relying on some argument based on PR's past performance (PPA).[2]

Then point out that what's done is done: PR has long since made and acted on his prediction, putting put $1M in Box 2 iff he predicted ONE, so that the contents of Box 2 cannot now be affected one way or the other: taking ONE cannot make an empty Box 2 full and taking BOTH cannot make a full Box 2 empty. 'Aha!' they exclaim, figuring that whether or not Box 2 contains $1M, BOTH will yield $1K more than ONE will. Now they are relying either on the dominance argument, as Nozick suggests, or on some other argument based on the fact that what's done is done (WDA).[3]

Now remind them of PR's awesome reliability and their impulse is to revert to the PPA. Then mention that the contents of Box 2 cannot now be affected and they will again be tempted by the WDA—but not convinced, since virtually everyone choosing BOTH netted a mere $1K. They reckon that the WDA must seem, to those who use it, more powerful than it is profitable. Yet the PPA, despite its evident record of success, seems to flout the fact that the past cannot be changed. Not knowing what to think at this point, most people find themselves flip-flopping between ONE and BOTH. Now they feel the full force of the apparent paradox that is Newcomb's Problem.

Of course the paradox is real only if the arguments on both sides are sound. For anyone convinced merely of their validity, the way to escape paradox is to find a false premise in one argument or the other.[4] For example, proponents of the WDA argue that to be sound the PPA requires more than a mere probabilistic connection between one's choice and PR's prior prediction. Causal decision theorists like Gib-

---

1 As is clear from the variants of Newcomb's Problem distinguished by Isaac Levi (1978), Robert Nozick's (1970) original presentation of the problem should have made explicit that PR's predictions of ONE and of BOTH are *both* generally accurate, and that there have been plenty of each. Also, we should make explicit that PR does not cause people's choices and that people's choices do not cause his predictions (so, in particular, PR does not rely on precognition). If backward causality were involved (and the conditions of the problem were still possible, contrary to what George Schlesinger [1980, 75-85] has argued), then ONE would be the obvious choice. Finally, for convenience we will assume that the predictions are made in advance, although this assumption is inessential to Newcomb's Problem (Lewis 1979, 237), provided that PR has no direct knowledge of people's choices.

2 I use this label not just because it captures what is common to all the arguments for ONE that I know of but also because Nozick's label is tendentious. According to Gibbard and Harper, for example, Newcomb's Problem is not a conflict between expected utility and dominance (see next paragraph) but is 'rather a conflict between two kinds of expected utility maximization' (1978, 152), orthodox and causal. Ellery Eells (1982) rejects causal decision theory and claims that the conflict merely concerns how the orthodox principle is to be applied to Newcomb's Problem. He argues that it justifies taking BOTH, not ONE, contrary to what Nozick had led us to believe.

3 I use this generic label because proponents of BOTH have criticized the dominance argument and have offered alternative arguments, all of which (so far as I know) appeal in one way or another to the idea that what's done is done. The dominance argument has been attacked by Doris Olin (1978), and Isaac Levi (1978, 375-6) points out that BOTH is the rational choice even in a version of the problem in which the payoffs for BOTH do not dominate those for ONE (if you take BOTH and PR predicts ONE, you get the $1M + $1K less a fine of $1500).

4 Another way out is to deny that Newcomb's Problem is possible, that its stipulated conditions cannot be fulfilled jointly. This strategy may be effective against those prediction paradoxes (as in Scriven 1964) that assume an impossible symmetry between the agent's and the predictor's information, but this is not assumed in Newcomb's Problem. Nozick (1974, 204-5) mentions several arguments that purport to show that the conditions of the problem are logically or physically impossible, but he rightly regards these as wildly implausible.

bard and Harper (1978) and Lewis (1979) maintain that a causal connection is required but that in Newcomb's Problem there is no such connection (see note 1). What you do cannot affect what PR predicted or what is in Box 2, and so it is silly to worry about losing $1M. If Box 2 does not contain $1M that is too bad, but it cannot be due to your choice of BOTH—you can't cut off your nose to spite your face if your face is already spited.

From the standpoint of the PPA, the trouble with the WDA is that instead of taking PR's reliability seriously enough, it takes two of the four possible outcomes too seriously. It represents the outcomes of each choice disjunctively: BOTH yields either $1M + 1K or $1K and ONE yields either $1M or $0, depending on the contents of Box 2. But PR's reliability makes two of the four cases too unlikely (impossible if PR is infallible) to be taken seriously. Practically speaking, your choice is between a predicted BOTH, with a payoff of $1K, and a predicted ONE, with a payoff of $1M.

It is indisputable both that nothing you do now can change what is in Box 2 and that PR is highly reliable, having shown no signs of decline. So we seem to be at an impasse. Arguments for ONE and for BOTH each seem bad from the opposing point of view.

II A Meta-argument for ONE

Nevertheless, one can only be amused by those advocates of BOTH who, as Nozick reports (1974, 102), realize that takers of BOTH almost always get but $1K whereas takers of ONE almost always get $1M, and proceed to bemoan the fact that rational people do so much worse than irrational ones. Despite their logical scruples, they seem to have a curiously low standard of what constitutes a good argument, at least in the context of Newcomb's Problem. Evidently they would rather be right than rich. One would think that a solution requires not merely a seemingly irrefutable argument but an argument that *works*, one whose use is likely to pay off to the tune of at least $1M.

But is there any argument that works? Since we cannot actually put arguments to the test, we can only make inferences from PR's record. Now surely there must be an explanation for his remarkable success.[5] We know that PR has access to detailed information about each person and makes his predictions by plugging this *psychological profile*, as it might be called, into a powerful psychological theory. He is so highly reliable, it would seem, only because people's choices are almost always *psychologically determined*: their ultimate choices are predictable from their psychological profiles at the time of the prediction. Although PR may occasionally get misinformation or make a miscalculation, his errors, rare as they are, are generally due to nonpsychological factors impinging on the process leading to the person's choice. These could be physiological influences, external physical factors, or results of random methods used to make choices.

Now presumably most people engage in some reasoning before arriving at a decision, and presumably their reasoning has much to do with their ultimate decisions. So it seems that the only plausible explanation for PR's success is that in general he anticipates the arguments people rely on. Only for those whose choices are not based on reasoning could he rest content with indicators like motivational tendencies and character traits. Perhaps there is some other way to account for his success, but I will assume that PR tries to anticipate people's reasoning. I readily admit that my meta-argument for ONE depends on this assumption.[6]

As you are trying to decide whether to take ONE or BOTH, you are aware that those who have opted for ONE have almost always gained $1M, although you do not know what argument (or how good an argument) any of them relied on. Even so, you can be confident that *any* argument for ONE, regardless of its quality from a logical point of view, was very likely to lead to a payoff of $1M. As for those who have opted for BOTH, you know that they have almost always had to settle for $1K regardless of what argument (or how good an argument) they may have relied on.[7] And for all you know, any argument

---

5 Eells (1982, 210-11) makes this observation but goes no further than to suggest that the only *kind* of explanation consistent with the conditions of Newcomb's Problem is one that invokes a common cause of the prediction and the choice.

6 Indeed, I admit that if PR did rely on mere indicators (making his success even more astonishing), BOTH would be the obvious choice. Taking ONE would make no more sense than would giving up smoking to reduce the risk of cancer, if smoking were statistically correlated with but not caused by cancer and both were caused by a common genetic factor. Only if PR were not highly reliable would it be reasonable to suppose that PR did not try to anticipate reasoning and relied merely on motivational indicators. That is why I disagree with Lewis (1979, 238) that it is inessential to Newcomb's Problem that PR be *highly* reliable. Merely moderate reliability would make for a different version of the problem and, since the MA would not go through, make for a different solution. A fortiori, I cannot agree with Lewis that 'it is inessential to Newcomb's Problem [at least as I understand it] that any prediction ... should actually take place. It is enough that some potentially predictive process should go on' (Lewis 1979, 237).

7 As applied to the WDA in particular, this was a common objection raised by cor-

for BOTH you are now considering has been used many times before, and if so, probably without success.

If there is any argument for BOTH that has reliably led to a payoff of \$1M + \$1K, you have no idea what it is. Since PR is not infallible but only highly reliable, someone who chose BOTH might have been lucky and received the maximal payoff, but a good argument for BOTH cannot rely on luck. A good one, giving you excellent chances of getting \$1M + \$1K, must be such that despite your using it, PR will not have anticipated your using it and will have predicted you would choose ONE. Indeed, it must include a lemma to that effect. However, there seems to be no way to establish such a lemma, since you have no reason to think that PR will not have anticipated any argument you use, much less that he will have predicted you to do the opposite of what it dictates. For all you know, such an argument may have been used before, but you have no way of knowing if it led to a payoff of \$1M + \$1K, and not by luck. So you must conclude that there is no good argument for BOTH.

III The Trouble with BOTH

However strong their confidence in their favorite argument, proponents of BOTH ought to ask themselves, 'Why does using an argument for BOTH, no matter how good it may seem, generally not result in the big payoff?' The obvious answer is that when people use such an argument, generally PR predicts BOTH. This does not seem to faze them, perhaps because they believe that this generalization is irrelevant to the concrete situation of any new participant.[8] The tendency of arguments for BOTH to result in one's not getting \$1M has no bearing on your decision now. What's done is done: whatever PR may have predicted in your case is independent of what you do now.

Proponents of BOTH may reason in this way, but they should consider that if a person actually faced with Newcomb's choice so reasons, it is very likely that PR anticipated this reasoning and left Box 2 empty. You are such a person and on the supposition that your choice is psychologically determined, if you so reason you can expect to net but \$1K. It is easy for an outsider, a philosopher or any other perfect judge, to endorse an argument for BOTH, but PR is not concerned to predict arguments endorsed by outsiders. If *you* were to use such an argument, you could count on PR having anticipated you would, in which case relying on the argument would be self-defeating. You cannot coherently (1) endorse the argument (thereby intending to take BOTH), (2) believe that if PR anticipated you would endorse it, you would get more than \$1K, (3) believe that PR anticipated you would endorse it and (4) believe that you will get more than \$1K. Even if the WDA might be valid somehow, as a player you cannot afford to accept it. For you cannot accept an argument for BOTH without believing that you will take BOTH, and if you believe that, you must consider it unlikely, in view of PR's reliability, that he predicted you would take ONE and put \$1M in Box 2.

Ifs and BOTHs. Since advocates of BOTH often concede that the reward of using their favorite argument is likely to be \$1K, how can we explain the appeal of BOTH? I suspect its appeal stems from a certain plausible but flawed pattern of counterfactual reasoning common to the diverse arguments for BOTH. I cannot survey these arguments, but George Schlesinger's (1977, 88ff) 'perfect judge' argument offers a good illustration. A certain rational, well-wishing observer is to judge whether you are better off choosing ONE or BOTH. He is allowed to look into Box 2 but not to tell you what he sees. That doesn't matter, claims Schlesinger, for he would give you the same advice whether or not he sees \$1M in Box 2. Either way, he would advise taking BOTH. Indeed, 'it does not matter that in fact there is no such judge at hand, since it is certain that if there was such a judge he would advise the player, no matter what, to take both boxes. It is necessarily true that it is in the best interest of the player to adopt what would be the best choice in the opinion of a sufficiently well-informed and intelligent judge, if he existed' (Schlesinger 1977, 89). However, what Schlesinger fails to take into account is that if you used his perfect-judge argument, PR almost certainly would have anticipated your using it. Knowing that it would lead you to choose BOTH, PR would have predicted BOTH and left Box 2 empty. What Schlesinger regards as 'a solid deductive argument' is in fact a logical way to avoid affluence. Its only merit is to guarantee a \$1K consolation prize and to illustrate, though not by design, a common error in counterfactual reasoning.

Since the perfect judge argument is really an embellished version of the what's-done-is-done argument and since the error I have in mind occurs in all versions of the WDA that I have seen, it might be best to consider the WDA in its bare bones form. Alan Gibbard and Wil-

---

respondents to *Scientific American*. Nozick reports that 'many pointed out that if you thought of that argument and were convinced by it, the predictor would (almost certainly) have predicted it and you would end up with only \$1,000' (1974, 102).

8 This point is made by Eells (1982, 210-18), who uses it to defend the orthodox conception of expected utility against the counterexamples put forth by causal decision theorists.

liam Harper put it this way:[9] 'Rational choice in Newcomb's situation ... depends on a comparison of what *would* happen if one took both boxes with what *would* happen if one took only the opaque box. What the agent knows for sure is this: if he took both boxes, he *would* get a thousand dollars more than he *would* if he took only the opaque box. That, on our view, makes it rational for someone who wants as much as he can get to take both boxes, and irrational to take only one box' (1978, 155; my emphasis). Gibbard and Harper are unfazed by the question 'If you're so smart, why ain't you rich?'; they respond, 'If someone is very good at predicting behavior and rewards predicted irrationality richly, then irrationality will be richly rewarded' (1978, 153). They claim that choosing ONE is the rational thing to do iff the following counterfactuals[10] (thus my emphasis above on 'would') are both true, or at least probably true:

(i) If you took ONE you would be a millionaire.

(ii) If you took BOTH you would be a non-millionaire.

Here they invoke the method they favor for maximizing expected utility, which uses probabilities of (counterfactual) conditionals rather than the conditional probabilities employed by the orthodox method.[11] They proceed to argue that the two conditionals cannot both be true or even probably true. Indeed, they go so far as to claim that both cannot be true even if PR is assumed to be not just highly reliable but downright infallible, the point at which most proponents of BOTH back off.[12] For even if PR is infallible, in which case you become a millionaire iff you choose ONE, they argue that this truth-functional proposition does not imply both counterfactuals.[13]

In order to show this, Gibbard and Harper appeal to a method of evaluating counterfactuals whereby the possible world to consider is the one which is most like the actual world at the time in question and which thereafter obeys all physical laws. They maintain that one should be concerned only with causal consequences of one's decision (which do not include what PR predicted or what he did with Box 2) and 'consider only worlds in which the past is exactly like the actual past, for since the agent cannot now alter the past, those are the only worlds relevant to his decision' (1978, 161). So if you are the agent and evaluate (i) and (ii) in respect to the time you are trying to make your decision, you are to assume that the contents of Box 2, whatever they are, are fixed. If there is $1M in Box 2, (ii) is false, and if not, (i) is false; either way, both are not true.

The result of applying Gibbard and Harper's method of evaluating counterfactuals to (i) and (ii) may seem plausible, but what if it is applied to the closely related (i') and (ii')?

(i') If you took ONE, PR would have predicted ONE.

(ii') If you took BOTH, PR would have predicted BOTH.

Intuitively it seems that if PR is infallible, both (i') and (ii') are true. Yet Gibbard and Harper's method does not yield this result: just as your choice cannot affect the contents of BOX 2, so it cannot cause PR to have predicted it. However, (i') and (ii') are *backward* counterfactuals, for which it is inappropriate to 'consider only worlds in which the past is exactly like the actual past.'[14] Indeed, this would lead to the dubious result that no backward counterfactual can be true unless its consequent is true in the actual world.[15]

---

9 As Lewis puts it, 'no matter how reliable the predictive process ... [BOTH is better because] one thereby gets a thousand more than he *would* if he declined, since he *would* get his million or not regardless of whether he took his thousand' (1979, 240; my emphasis).

10 Gibbard and Harper allow that despite the label, 'For a proposition to be a counterfactual, we do not require that its antecedent be false' (1978, 127).

11 Theirs is not the only version of causal decision theory, but I believe my objection can be extended to other versions as well.

12 Stephen Leeds, for example, thinks that ONE is rational if PR is perfectly infallible, since the choice is effectively between $1M and $1K. Leeds does advocate BOTH when PR is not perfectly infallible, but he is unable to 'explain, or explain away, the curious discrepancy between our intuitions' in the two cases (1984, 106). Eells avoids this problem by arguing that it is irrational to believe that PR, whatever his record, is perfectly infallible (1982, 208).

13 The situation is more complex if PR is not infallible, for then this truth-functional proposition would not be true. Presumably Gibbard and Harper would claim that choosing ONE is rational iff both (i) and (ii) are highly probable and would deny that both are highly probable.

14 Jonathan Bennett (1984), observing that this has led some to treat backward counterfactuals differently from forward ones, criticizes this bifurcated approach and develops a uniform treatment instead.

15 Here is one sort of backward counterfactual that can be true even though its consequent is false. Suppose that a certain event was the inevitable result of a certain prior event. For example, a bicycle is damaged by being run over by a truck. The backward counterfactual, 'If the bicycle were undamaged, it would not have

(i) and (ii) are both true, then so are the forward counterfactuals (i) and (ii). However, since Gibbard and Harper's method may seem appropriate to the latter, rather than criticize their method I will argue disjunctively: either (1) the method does not apply to (i) and (ii) either, because they should be regarded not as ordinary forward counterfactuals but as back-tracking ones, or (2) it doesn't matter that (ii) is false.

(1) *Given* the conditions of Newcomb's Problem, in particular that PR is infallible (as Gibbard and Harper are assuming), both conditionals are true. But if we evaluate them both as true, we are giving priority to PR's infallibility over the the contents of Box 2 when we determine the relevantly closest possible worlds. And since PR's prediction varies with your choice, in determining those worlds we are not holding the past fixed. We are treating (i) and (ii) as back-tracking conditionals, to which, as noted by Lewis (1981, 22), Gibbard and Harper's method is inapplicable.

(2) Or we can agree with Gibbard and Harper that (i) and (ii) are not true, but deny that the choice of ONE requires the truth of (ii). Suppose you intend to choose ONE and then, in order to evaluate (ii) by means of Gibbard and Harper's method, you consider the possible world whose sole difference from the actual world, the world in which you take ONE and become a millionaire, is that you take BOTH. In that world you would become a millionaire. This makes (ii) false, but that does not affect the rationality of choosing ONE. For part of its rationale is that PR will have predicted one's choice, and this is not the case in the possible world in question.

For our purposes it doesn't matter whether we take route (1) or (2), since either way it is irrelevant to the rationality of ONE that Gibbard and Harper's method of evaluating counterfactuals reckons (ii) as false. By taking route (1) we reject their method when applied to the peculiar situation of Newcomb's Problem, in which (i) and (ii) must be regarded as back-tracking conditionals: PR is assumed to have anticipated one's choice, whatever it may be, so that his prediction, along with the contents of Box 2, depends counterfactually (though not causally) on your choice. In this way we accept Gibbard and Harper's method of maximizing expected utility by considering probabilities of counterfactual conditionals.[16] By taking route (2), however, although we can then accept their method of evaluating counterfactuals, even when applied to the peculiar situation of Newcomb's Problem, in denying that choosing ONE requires the truth of (ii) we are thereby rejecting their method of maximizing expected utility. For it just doesn't matter that in a world where taking ONE makes you a millionaire, taking BOTH would do so as well.

The Fallacy Behind BOTH. We need not tread further on the treacherous topic of counterfactuals to identify the fallacy which, in my view, is common to the various arguments for BOTH. The basic issue is what is and what is not implied by their common assumption:

(WDID) Since what's done is done, your choice cannot affect PR's prediction or what he put in Box 2.

Since your choice does not cause PR's prediction and since it is his prediction that causes him either to put or not to put $1M in Box 2, clearly we cannot quarrel with WDID. But does it really imply that his prediction (and action) is counterfactually independent of your choice, so as to justify taking BOTH? In the normal case, where PR's information and calculations were accurate and where the psychological causal chain he predicted is not broken by nonpsychological factors, your choice and his prediction have a common cause. The common cause is whatever it is about you at the time of PR's prediction that (a) leads you to make your choice and (b) leads him to predict it. Yet advocates of BOTH suppose that because what is in Box 2 is determined at the time of your choice, the way to decide is by considering each of the two possible cases (since you don't know what is in Box 2) and comparing the payoff of each option. If you take ONE, you get either $1M or $0; if you take BOTH, you get either $1M + $1K or $0 + $1K; either way you should choose BOTH.

What is the fallacy in this argument? At the time you are considering the argument you do not yet know which choice you will make. You do not know whether the actual world is the world in which you choose ONE or the world in which you choose BOTH, and which one it is makes a world of difference! For what you can infer from your possibly but not actually making a certain choice is different from what you can infer from your actually making that choice. Suppose that you will actually choose BOTH – the actual world is the world in which you choose BOTH. Then probably there is no $1M in Box 2. In the closest possible world in which you choose ONE instead, you get $0 rather than $1K. But so what? On the other hand, suppose that you actually will choose ONE – the actual world is the world in which you choose ONE. Then probably there is $1M in Box 2. In the closest possible world

---

been run over by a truck,' is true – but not according to Gibbard and Harper's method.

16 I should note that even if Gibbard and Harper's argument assumed that PR is not infallible but merely highly reliable, still it would be highly probable that both (i) and (ii) are true.

in which you choose BOTH instead, you would get $1M + $1K. But again, so what?[17]

I am suggesting, then, that even though your choice cannot affect what PR predicted or what he put in Box 2, you should not regard them as counterfactually independent of your choice. To do so would be to assume that just because you are free to make either choice, your choice was not psychologically determined at the time of the prediction, and there is no reason to bet on that. Since it is unlikely that any nonpsychological factors have intervened, probably your choice and his prediction have a common cause: your total psychological state (as given by his profile of you) at the time of his prediction. Since you have no way of knowing what your total psychological state was or how it provided PR with a basis for his prediction, your only evidence for what he predicted is your actual choice, but you have not yet made that choice. If you make a decision irrevocably, if you don't merely 'try it on for size' as the WDA asks you to do, you will thereby obtain evidence for what PR predicted, but it will be too late for you exploit that evidence.

Another way to appreciate the fallacy behind BOTH is to recall how advocates of BOTH would answer the meta-argument for ONE. Though admitting that your choice is probably predictable, they would insist that it is pointless to consider this fact in making your choice. For PR has long since made his prediction. And they would argue that it matters that your choice was predictable only if this fact implies *both* that if you chose ONE, probably PR will have predicted ONE *and* that if you choose BOTH, probably PR will have predicted BOTH (where these are not construed as back-tracking conditionals). However, the MA does not presuppose this implication. As we have seen, it assumes only that one of these conditionals holds, namely the one involving your actual choice. Since you do not yet know what your choice will be, you do not yet know which one that is. You do not have to know that both hold in order to be confident that PR predicted whatever choice you will in fact make.

17 And it is no fair using retroactive reasoning to justify BOTH. So if a ONEr opens Box 2 and finds $1M in it, a BOTHer might say, 'If only you had taken BOTH, you would have an extra $1K.' And then when it is the BOTHer's turn and he opens Box 2 only to find it empty, he might say, 'See, if I had taken ONE, I would have ended up with $1K less, i.e. nothing.' The BOTHer might argue that these past-tense counterfactuals not only are true but imply that before the choices were made the corresponding future-tense counterfactuals were true: 'If you were to take BOTH, you would get an extra $1K,' and 'If I were to take ONE, I would get $1K less.' To this the ONEr could only laugh—all the way to the bank.

IV One Last Problem: Temptation

No doubt there will be those who object that I am just begging the question against the arguments for choosing BOTH, but I could reverse the charge and demand a refutation of the meta-argument for ONE, and we could trade punches indefinitely.[18] However, I do admit that I am not so certain of the MA as to ignore a different objection which, ironically, exploits the fact that anyone who (like me) accepts the MA might still harbor residual doubts about it.

To enhance the force of the objection I have in mind, let us make the following cosmetic changes in the conditions of Newcomb's Problem. Instead of being either filled or empty, Box 2 is filled with $1M—either in real money or in play money. Before making your final decision, you get to shake Box 2 so as to hear what is inside, but of course no peeking is allowed. Inasmuch as its contents are yours to keep, let us say that your choice is whether to TAKE or to LEAVE Box 1. You don't know whether you hear real money or play money in Box 2, but you do know that nothing will happen to whatever is in there after you decide whether or not to TAKE Box 1.

The objection is based on what might be called 'the argument from temptation.' Imagine yourself faced with Newcomb's Problem and deliberating about which choice to make when the moment of truth arrives. Presently you are persuaded by the MA but fear that later, as you are shaking Box 2, you'll think, 'whether it contains real money or play money, I will get $1K more if I TAKE Box 1 than if I LEAVE it. What's done is done. So I might as well TAKE it.' You are not harboring the hope that PR will have predicted you would LEAVE it even if you should choose to TAKE it, but you fear that PR will have suspected you might yield to a last-minute temptation. In this way the mere possibility, once you recognize it, of yielding to a last-minute temptation to TAKE Box 1 seems to justify doing so. For if *you* can't be sure that you won't yield to this temptation, you can't be sure that PR was sure when he made his prediction. So he may have put play

18 For example, I can anticipate the following rebuttal to the MA. They may say that whatever argument leads you to TAKE Box 1 will have been anticipated by PR and thereby invalidated (no $1M in Box 2) but, since PR did not put $1M in Box 2, you would have done worse to LEAVE Box 1 and therefore that any argument leading you to *LEAVE* it would thereby have been invalidated. Of course my reply, however futile at this point, would be that it is only on the supposition that you TAKE Box 1 that we infer no $1M in Box 2. Once again, obviously, the issue is how to assess the relevant counterfactuals. The issue can be disputed indefinitely, but meanwhile you have to make the choice.

money in Box 2.[19] Thus you are afraid that as persuasive as the MA seems now, it will ring hollow later when, as you shake Box 2, you feel the urge to TAKE Box 1. You just cannot convince yourself now that later you will remain convinced by it in the face of the thought that 'what's done is done,' and not be overwhelmed by that urge.

Why BOTH is Still Tempting. There is good reason for this fear. The decision to LEAVE Box 1 will be difficult to carry out because you realize, in effect, that it is not *ratifiable*, in Richard Jeffrey's (1981) sense. A decision is ratifiable only if it is one you can 'live with,' i.e. see as rational once you have made it and up to the time of action, and it does seem, as you look forward to when you hear the shuffling of paper inside Box 2 and must take action, that the decision to LEAVE Box 1 is not ratifiable at the last minute. Now Jeffrey claims that to be rational a decision must be ratifiable,[20] in which case you should TAKE Box 1. However, Stephen Leeds (1984, 99-102) has argued against this claim.[21] His counterexamples are rather out of the ordinary, but there are many down-to-earth instances of rational but unratifiable decisions, which also illustrate why an unratifiable decision can be difficult to carry out. Suppose, for example, that you want to cut down on your smoking but not to give it up. However, whenever the thought of smoking occurs to you, you reason that having one cigarette has a negligible impact on how much you smoke: so you smoke it. Unfortunately, this reasoning is as valid after each cigarette as before. The decision not to have one more cigarette is not ratifiable and hard to adhere to (also, it illustrates why slippery slopes are slippery), but still it is rational.

Similarly, the decision to LEAVE Box 1 is rational without being ratifiable, at least not at the time of action.[22] The argument from temptation seeks to exploit the fact that one who realizes this while deliberating will suspect that the decision will be difficult to carry out. To that extent he will wonder if PR really did predict that he would decide to LEAVE Box 1. Thus this very suspicion is enough to undermine his decision. That is true, I admit, but only insofar as the suspicion persists. Fortunately, there is something you can do to get rid of it.

Commit Yourself. If you believe that the decision to LEAVE Box 1 is rational but fear that you won't be able to ratify it and will TAKE Box 1, what can you do? You can take measures to remove the possibility of succumbing to a last-minute temptation to TAKE Box 1 and to make sure that you will LEAVE it. And you should do this because you can count on PR to have anticipated it. So instead of worrying about what you might do later, now you can either cause yourself later to LEAVE it involuntarily or, if that is offensive, give yourself an incentive to LEAVE it voluntarily. Undergo hypnosis or make an appropriate side bet (bet someone $2000 to $1 that you will LEAVE Box 1), but do whatever it takes to make sure your choice will have been predictably to LEAVE Box 1. If you have no doubt of that, PR will have had no doubt that you would have no doubt and will have counted on you to LEAVE it.

One might object that it cannot be rational to limit one's subsequent freedom of choice. Yet this can be rational, as in taking on such a position as judge or doctor, which constrains one's personal discretion (Rawls 1955). Another example of rationally limiting one's freedom is the commitment required to make credible a threat that is very costly to carry out. For instance, if your son has stolen something and refuses to return it, you might threaten to call the police even though that would cost both legal fees and your son's affection. Recognizing this, he might think it would be crazy of you to call the police (no wonder that certain threats are more credible when posed by crazy people), and so if you are to make the threat credible enough to induce compliance, you must demonstrate conviction that you will carry it out if necessary. If you know you cannot bluff while still demonstrating conviction, you must do what it takes to convince *yourself*. You must make sure you will not be inhibited by the thought that 'this will hurt me more than it will hurt you.'

---

19 Formulations of Newcomb's Problem are not explicit about what PR does when he is unsure. Since Nozick (1974, 102) does stipulate that PR does not put $1M in Box 2 if he thinks the choice will be made on the basis of a random event, we could extend that stipulation to include momentary impulses and other sources of uncertainty. However, the problem of temptation would remain even on the stipulation that PR puts $1M in Box 2 if he thinks that LEAVE it is at least more likely than TAKE it, since you couldn't be sure that he thought this.

20 Jeffrey incorporates this requirement into orthodox decision theory, and like Eells he argues that the various counterexamples put forth by causal decision theorists are only apparent.

21 Curiously enough Leeds still believes, 'for reasons I am entirely unable to explain' (1984, 106) that one should TAKE Box 1.

22 Notice that *prior* to the time of action the decision to *TAKE* Box 1 is not ratifiable. For the more certain you are that you will adhere to it and not change your mind, the more certain you are that there is $0 in Box 2. The decision to TAKE Box 1 *is* ratifiable at the time of action. For then you know that there is probably $0 in Box 2, hence that if you were now to LEAVE Box 1, you would get $0 rather than $1K.

Similarly, in the case of Newcomb's Problem the strategy is to commit yourself to the decision to LEAVE Box 1 and thereby render yourself incapable of being tempted to TAKE it. Then you will not be seduced by the argument, however compelling, that later, when you hear the shuffling of paper in Box 2 and know you can TAKE Box 1, you will be unable freely and rationally to LEAVE it. You can freely and rationally make sure now that later you will LEAVE it. In denying yourself any further freedom of choice, you have every reason to believe that PR predicted you would.

*****

The rational decision in Newcomb's Problem is to choose ONE, as shown by our meta-argument. Any argument you use for BOTH can be good only if despite your using it, PR predicted you would choose ONE; but there is no way to establish the required lemma to that effect. So there is no good argument for BOTH. Arguments for ONE require no such lemma and, indeed, on the presumption that PR anticipated you would use it, you can use *any* argument for ONE. Nevertheless, wondering how PR anticipates people's choices, you cannot forget that what's done is done and not worry about yielding to a last-minute temptation to take BOTH. To combat that worry you should commit yourself to taking ONE. And instead of wondering how PR does it, you should recall what Muhammad Ali said prior to facing the seemingly invincible heavyweight champion Sonny Liston: 'If Cassius Clay says a rooster can lay an egg, don't ask how—grease that skillet!'

Received July, 1985

References

Jonathan Bennett, 'Counterfactuals and Temporal Direction,' *Philosophical Review* **93** (1984), 57-91.

Ellery Eells, *Rational Decision and Causality* (Cambridge, England: Cambridge University Press 1982).

Martin Gardner, 'Mathematical Games,' *Scientific American* **30** (1973), 104-8.

Alan Gibbard and William Harper, 'Counterfactuals and Two Kinds of Expected Utility,' in C.A. Hooker, J.J. Leach, and E.F. McClennen, eds., *Foundations and Applications of Decision Theory* (Dordrecht, Holland: Reidel 1978), 125-62.

Richard Jeffrey, 'The Logic of Decision Defended,' *Synthese* **48** (1981), 473-92.

Stephen Leeds, 'Eells and Jeffrey on Newcomb's Problem,' *Philosophical Studies* **46** (1984), 97-107.

Issac Levi, 'Newcomb's Many Problems,' in C.A. Hooker, J.J. Leach, and E.F. McClennen, eds., *Foundations and Applications of Decision Theory* (Dordrecht, Holland: Reidel 1978), 369-83.

David Lewis, 'Prisoner's Dilemma Is a Newcomb Problem,' *Philosophy and Public Affairs* **8** (1979), 235-40.

David Lewis, 'Causal Decision Theory,' *Australasian Journal of Philosophy* **59** (1981), 5-30.

Robert Nozick, 'Newcomb's Problem and Two Principles of Choice,' in Nicholas Rescher et al., eds., *Essays in Honor of Carl G. Hempel* (Dordrecht, Holland: Reidel 1970), 114-46.

Robert Nozick, 'Reflections on Newcomb's Problem,' *Scientific American* **31** (1974), 102-8.

Doris Olin, 'Newcomb's Problem, Dominance and Expected Utility,' in C.A. Hooker, J.J. Leach, and E.F. McClennen, eds., *Foundations and Applications of Decision Theory* (Dordrecht, Holland: Reidel 1978), 385-98.

John Rawls, 'Two Concepts of Rules,' *Philosophical Review* **64** (1955), 3-32.

George N. Schlesinger, *Religion and Scientific Method* (Dordrecht, Holland: Reidel 1977).

George N. Schlesinger, *Aspects of Time* (Indianapolis: Hackett 1980).

Michael Scriven, 'An Essential Unpredictability in Human Behavior,' in B.B. Wolman and E. Nagel, eds., *Scientific Psychology: Principles and Approaches* (New York: Basic Books 1964), 411-25.