Jordan Howard Sobel

Predicted Choices

*First Day**

It is sometimes said that there cannot be a predicted choice, that the very phrase, truly predicted choice, is a contradiction in terms. An argument for this *extreme* thesis might go like this:

> Suppose—we will show that this is impossible—that one of my choices has been truly predicted. Then I have to make it, for if I do not make it, then it was not truly predicted. For the same reason, given the supposition that a choice of mine has been truly predicted, it is not possible for me to make any other choice in its place, and I have in connection with it no choice. But a choice that I *have* to make, a choice in connection with which I have no choice, is no *choice* at all. So contrary to our supposition, it is not the case that one of my choices has been truly predicted.

I will discuss neither the extreme thesis that there cannot be a predicted choice nor this argument for it. I have stated them solely in order to set them aside so that in what follows their intrusions may be resisted.

The problem I take up is circumscribed in several ways. First I am interested only in *rational* choices. And second, in the cases with which I am concerned, I do not assume that choices made have been truly predicted: I do not *insist* on this. I require only that the agent should be nearly certain that his choice has been truly predicted. In the cases that interest me the agent has what are in his view the best of reasons for thinking that his choice has been truly predicted. He is not sure what choice has been predicted, and he will not be sure of this until he makes his choice, but he is already sure, or nearly sure, that whatever choice he will make, it has already been predicted that he will make it. It can seem that in at leat some such cases there is no possibility of a rational choice. I will consider two cases in which rational choice at least *seems* to be impossible.

NEWCOMB'S PROBLEM

The situation: There are two boxes on a table. A thousand dollars are in Box 1, and either a million or nothing in Box 2.
Possible Choices: take both boxes; take only Box 2.
The ringer: There is either a million dollars or nothing in Box 2, depending on what the organizer of the problem has predicted that you will choose to do. He has put $M in Box 2 if he has predicted that you would take only Box 2, and $0 if he has predicted that you would take both boxes. And you have the best of reasons for thinking that he has truly predicted the choice you will make. You are nearly *certain* that he has truly predicted the choice you will make. Indeed, whatever your choice were to turn out to be you would be nearly sure that he had predicted it—learning of either choice would be, for you, nearly tantamount to learning that that was the choice he had predicted. And you are certain that whatever prediction he made he acted on, and that he is now quite out of the picture. You are certain that your choice can have no effect on the contents of Box 2.

The structure of the problem

		Possible contents of Box 2	
		$M	$0
Possible choices	take both	$M + 1000	$1000
	take only Box 2	$M	$0

What do you do? What is your *rational* choice? Evidently, or *apparently*, your rational choice is *not* to take *both boxes*.

> For if you made this choice you would be nearly certain that the organizer had predicted it and put nothing in Box 2; you would be nearly certain of getting a mere $1000. Whereas if you were to take only Box 2, you would be nearly certain of getting $M. Your choice is between a near certainty of $1000, and a near certainty of $M. There appears to be no contest here. It would it seems, be *irrational* to take *both* boxes when by taking only the second you can be sure of $M.

But it seems that your rational choice is not to take only Box 2, *either*!

> For you are sure that the prediction has been *made* and the boxes set up with money according to the prediction that *has* been made. You are sure that your choice will not affect any of that.

> How could it? All of that is over and done with. So it seems that you should *not* take only Box 2, for whether the $M is in this box or not, you come out ahead by taking both boxes. It would be irrational to leave the $1000 behind on the table when it is there for the taking. It would, it seems, be irrational to take only Box 2.

Only two choices are possible, and it has been argued that each is irrational. From this it follows that in this case there is no possibility of a rational choice. The structure of the case—including most prominently the agent's confidence in the correctness of the organizer's *prediction* of his choice—seems to make *rational* choice in the case *impossible.* (For a life-sized variant of Newcomb's Problem, begin by supposing that a life of misery and self-denial is a sign of pre-ordained compensations and infinitely better things to come.)

APPOINTMENT IN SAMARRA
Death Speaks

> There was a merchant in Bagdad who sent his servant to market to buy provisions and in a little while the servant came back, white and trembling, and said, Master, just now when I was in the market-place I was jostled by a woman in the crowd and when I turned I saw it was Death that jostled me. She looked at me and made a threatening gesture; now, lend me your horse, and I will ride away from this city and avoid my fate. I will go to Samarra and there Death will not find me. The merchant lent him his horse, and the sevant mounted it, and he dug his spurs in its flanks and as fast as the horse could gallop he went. Then the merchant went down to the market-place and he saw me standing in the crowd and he came to me and said, why did you make a threatening gesture to my servant when you saw him this morning? That was not a threatening gesture, I said, it was only a start of surprise. I was astonished to see him in Bagdad, for I had an appointment with him tonight in Samarra.
>
> W. Somerset Maugham

This story appears as front matter to John O'Hara's *Appointment in Samarra.*

Here is a related, but somewhat different story. My enemy will kill me if he and I are in the same city tomorrow. He has predicted where I shall be and is on his way there now. Nothing I do will alter his course. I am sure of all of this. My choices are limited; I can be in either Bagdad or Samarra. And as long as I avoid my enemy it doesn't matter to me where I am tomorrow. The *problem* is that I have the best of reasons for thinking that my enemy has correctly predicted where I will be. I know that he is a highly reliable predictor, an especially good one in these situations. So I am nearly certain that he has correctly predicted where I shall be. As in Newcomb's Problem, learning of my decision

would be tantamount to learning of his prediction—*whichever* place I learned I was going, I would be nearly certain that *that* was where he had *predicted* I was going. I have that much confidence in his perspicacity. Here is the choices-circumstances-consequences structure of my problem.

		Possible destinations of my enemy	
		Bagdad	Samarra
Possible choices	go to Bagdad	death	life
	go to Samarra	life	death

Where shall I go? Which choice would be rational?

It can seem that it does not *matter* where I go, or *which* choice I make, and that therefore either choice would be rational, or at least not irrational. An argument for this might be:

> It is nearly certain that if I were to go to Bagdad my enemy would have predicted that I was going there, and that he would meet me there. Similarly, supposing I were to go to Samarra. So it doesn't matter from the life/death standpoint where I go. And since we have said that it doesn't matter from any other standpoint, it doesn't matter at all. But when it doesn't matter what I do, then whatever I do is reasonable, or at least not unreasonable.

But consider: *Could* it be rational for me to choose Bagdad?

> No—it can be argued—because if I were to decide for Bagdad, then I would, while secure in this decision, be nearly sure that my enemy was going there. But then (look at the first column) it would be irrational for me to go there, and I ought to change my mind, and to go to Samarra instead. So choosing to go to Bagdad would be irrational.

And it is obvious that it can be argued similarly that it would be irrational to choose to go to Samarra. The general principle at work in *these* arguments is that a choice or decision is rational only if realizing that one had made it would not give one sufficient reason for a *change* of mind and decision. On this principle there is no possibility of a rational choice in the present case *because* there is no possibility of a fully reflected upon *stable* choice. So once again we have a case whose structure—including most prominently the agent's confidence in another's prediction of his choice—*seems* on *final* analysis to make a rational choice impossible.

Second Day

What should we make of all of this? I think the *final* analysis given of the Samarra Case is *correct*. I think that where no fully reflected upon *stable* choice is possible, no *rational* choice is possible. And given that in the Samarra Case, whatever his choice, the agent would be nearly certain that his enemy had predicted that choice, I think that no fully reflective choice *would* be stable. (Discussion of defects of the *initial* analysis of this case is omitted.)

So I think the final analysis of the Samarra Case is correct. But I think that the analysis given of Newcomb's Problem is defective. I think that there *is* a uniquely rational choice in that problem, namely, to take both boxes and not leave the $1000 behind. The argument against *this* choice sounds right, I think, only because it confuses two different kinds of claims about the case. The claims

> (1a) If I were to take both boxes, I would be nearly certain that I was getting only $1000.

and

> (1b) I am nearly certain that if I were to take both boxes, I would get only $1000.

are confused, as are the claims

> (2a) If I were to take only Box 2, I would be nearly certain that I was getting $M.

and

> (2b) I am nearly certain that if I were to take only Box 2, I would get $M.

These confusions are abetted by such words as 'if I were to take only Box 2, I would be nearly certain *of getting* $M' which, while specific to (2a) are so similar to 'if I were to take only Box 2, I would be nearly certain *to get* $M', *which* words are much closer to (2b). To make the intended senses of these four displayed claims *quite* clear, we could insert 'then, either as a consequence of or independently of that,' in each, before the words 'I would'.

The *second* members of the displayed pairs of claims would be supremely relevant to what I ought to do since they concern probable consequences—probable 'material' consequences—of my actions, namely money. In contrast, the *first* claims in these pairs concern relatively insignificant consequences and are thus hardly relevant at

all. They concern (possibly passing) states of mind—admittedly, in the case of (2a) a very pleasant (possibly passing) state of mind, though not one that would be, just as such, worth *much* money to me, and *certainly* not one for the experience of which I would pay, or give up, $1000. What is *awkward* is that *while* only the *second* claims in these pairs are really *relevant* to choice, only the *first* claims of these pairs are have both been stipulated to be *true* in Newcomb's Problem. Indeed, the *second* claims of the pairs are presumably *not* both true in the case: they *could* not *both* be true unless I (the agent) do not understand the case or am very confused. In the case, I should be certain that

(1*) If I were to take both boxes, I would get $1000.

is *true* if and only if

(2*) If I were to take only Box 2, I would get $M.

is *false*. In the Problem, since my choice cannot affect what is in the boxes, either whatever I do I get at least $M, or whatever I do I get at most $1000: see columns one and two of the Problem's matrix. So unless I (the agent in the case) am obtuse or very confused, I *can't be certain of both* (1*) and (2*). Thus (1b) and (2b), though they would be supremely relevant, are not both available for purposes of determining what I ought to do. In contrast, (1a) and (2a) *are* both available—they are both true in the case. But *they* are not very relevant. The argument against both boxes *seems* compelling only because it does not make these distinctions and proceeds as if *relevant* premises, the b-claims, were identical with similar sounding *true* premises, a-claims.

There are dilemmas in which no rational choice is possible. Some cases in which the agent is nearly certain that the choice he will make has been predicted are like this. The Samarra Case is like this, even if, as I think, Newcomb's Problem is not. But what is it that allows the Samarra Case to 'work'? And what is it that makes Newcomb's Problem 'work' to the extent that it does?

What makes these cases 'work' to the extent that they do has nothing *essentially* to do with predictions of choices. What these cases have in common, and what makes them 'work,' is that in each, *choices* would be for the agent *signs* of choice-relevant factors of which he would not consider them to be *causes*. For example, that I had chosen only Box 2 would be a *sign* that there was $M in this box, though certainly not a *cause* of $M's being in this box. Similarly, that I had chosen to go to Samarra would be a sign that my enemy was going there too, but not a cause of *that*. Now *one* way in which choices can be signs but certainly not causes of choice-relevant factors is by these factors being caused by reliable *predictions* of these choices. That is the way of our two cases. But there are *other* ways. For example, such choices can be caused by such factors.

To illustrate this last possibility, here is a prediction-free case that is otherwise like the Samarra Case. I am sure that I have either disease A or disease B, and I am trying to decide whether or not to drink water. Now I am nearly sure that if I had disease A it would strongly dispose me to decide, possibly for what I took to be good reasons, *to* drink; whereas if I had disease B, it would strongly dispose me to decide *not* to drink. And I am sure that if I have disease A I will recover if and only if I do *not* drink, and if I have disease B I will recover if and only if I *do* drink. Finally, my interest here is in my health, and not at all in drinking or not drinking as such.

		Possible diseases	
		A	B
Possible choices	drink	death	life
	abstain	life	death

Neither choice could, if fully reflective, be stable in the case. Choosing to drink would be a sign that I had disease A and thus on reflection a reason for choosing to abstain, while choosing to abstain would be a sign that I had disease B and so a reason for choosing to drink. The case is like the Samarra Case. Though no views regarding predictions are involved, choices would be signs but certainly not causes of choice-relevant factors, and given the structure of these cases this peculiar relation would undo any choice I made, unless I simply made a choice without thinking about it, without continuing to think about it.

In the Disease Case, choices would be signs of choice-relevant factors because choices are viewed as *caused* by these factors. And it would be natural to include similar features in fuller statements of Newcomb's Problem and the Samarra Case. In Newcomb's Problem, for example, my choices are signs but certainly not causes of the state of Box 2, presumably because I believe the predictor could see what deliberative processes I am disposed to employ and what choices they figure to *produce* (causal notion). My confidence in the predictor, and in my enemy in the Samarra Case, might well have such grounds, and include a view on my part of my choices as things that are caused, a view of choices I am trying to make as choices I will be *caused* to make. That view of my choices is an explicit feature of the Disease Case.

Reflection on this feature of at least some choice-problems of kinds we have considered leads to a final question. Is it ever *reasonable* for a

deliberating agent to view in this way choices he is trying to make—to view them as things that will be *caused*, rather than as *uncaused causes* of things? If not, then although there *are* cases in which rational choices are not possible, *perhaps* there are no cases possible for *fully rational agents* in which rational choices are not possible. It would, I suppose, be nice if at least *that* were true. I leave open the question whether or not it is.

NOTE

* This essay has its origin in two lectures given to members of a philosophy in literature class who were reading Sophocles *Oedipus Rex*. Numerical utilities and probabilities were avoided. I offer scripts of these lectures in the belief that issues raised by Newcomb's Problem and the like can be taken some distance without introducing complications of mathematical decision theory, and in the hope of bringing these issues to the attention of a wider audience than their technical treatments are likely to attract.

References

Richmond Campbell and Lanning Sowden, eds., *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, University of British Columbia Press, forthcoming in 1985. (An anthology with a very useful introductory essay by the editors.)

Allan Gibbard and William L. Harper, "Counterfactuals and Two Kinds of Expected Utility," *Ifs*, eds. W.L. Harper and others, Reidel 1981. (An analysis of Newcomb's Problem is included as well as a discussion of a 'Samarra Problem'.)

Robert Nozick, "Newcomb's Problem and Two Principles of Choices," *Essays in Honor of Carl G. Hempel*, ed. N. Rescher, Reidel 1969. (The first published discussion of Newcomb's Problem.)

Robert Nozick, "Reflections on Newcomb's Problem," *Scientific American*, March 1974. (A report in the Mathematical Games department on correspondence received concerning a column by Martin Gardiner on Newcomb's Problem.)

Jordan Howard Sobel, "Rational Actions and Choices, and Expected Utilities," *Theoria*, 1983, "Circumstances and Dominance in a Causal Decision Theory," *Synthese*, forthcoming. (Formal complications are courted in these papers.)