ROBERT NOZICK

NEWCOMB'S PROBLEM AND TWO PRINCIPLES OF CHOICE*

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

I

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

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

$$\text{(B1)}\ \{\$1000\} \qquad \text{(B2)} \begin{Bmatrix} \$M \\ \text{or} \\ \$0 \end{Bmatrix}$$

You have a choice between two actions:

(1) taking what is in both boxes

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

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

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

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

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

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

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

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

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

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

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

The Second: The being has already made his prediction, placed the $\$M$ in the second box or not, and then left. This happened one week ago; this happened one year ago. Box (B1) is transparent. You can see the \$1000 sitting there. The $\$M$ is already either in the box (B2) or not (though you cannot see which). Are you going to take only what is in (B2)? To emphasize further, from your side, you cannot see through (B2), but from the other side it is transparent. I have been sitting on the other side of (B2), looking in and seeing what is there. Either I have already been looking at the $\$M$ for a week or I have already been looking at an empty box for a week. If the money is already there, it will stay there whatever you choose. It is not going to disappear. If it is not already there, if I am looking at an empty box, it is not going to suddenly appear if you choose only what is in the second box. Are you going to take only what is in the second box, passing up the additional \$1000 which you can plainly see? Furthermore, I have been sitting there looking at the boxes, hoping that you will perform a particular action. Internally, I am giving you advice. And, of course, you already know which advice I am silently giving to you. In either case (whether or not I see the $\$M$ in the second box) I am hoping that you will take what is in both boxes. You know that the person sitting and watching it all hopes that you will take the contents of both boxes. Are you going to take only what is in the second box, passing up the additional $1000 which you can plainly see, and ignoring my internally given hope that you take both? Of course, my presence makes no difference. You are sitting there alone, but you know that if some friend having your interests at heart *were* observing from the other side, looking into both boxes, he *would* be hoping that you would take both. So will you take only what is in the second box, passing up the additional $1000 which you can plainly see?

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

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

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

II