Newcomb's problem: the causalists get rich

PHYLLIS MCKAY

Analysis 64.2, April 2004, pp. 187–89.

Suppose you are offered two boxes, one red and one black. You have to decide whether to take one box, or both boxes. There is always £1,000 in the black box, but what is in the red box varies. A predictor will have put £1,000,000 in cash in the red box if she judges that you will take only the red box. But if she judges that you will take both boxes, she will put nothing in the red box. We suppose that this predictor is a spookily good judge of character, and certain to be right about you. We could suppose that this is a long-running game show, and the predictor has never yet been wrong. In the past, if the contestant has taken the red box, it turned out to contain £1,000,000. But when contestants in the past have taken both boxes, the red box contained nothing.

It is usually thought that evidential and causal decision-theory give different prescriptions for action for such a choice. Evidentialists advocate taking one box, since this choice is evidence for there *already being* £1,000,000 in the red box and will render the best option of getting £1,000,000 most likely. On the causalist view, you should take both boxes. The deposit in the boxes has *already happened*, and can no longer be affected by you – or by anyone at all. In the best case, you will have £1,001,000, and in the worst case you will at least have £1,000.

To make the right choice, you must decide whether you are in a *genuine*, or merely an *apparent* Newcomb case. Which case you are in is determined by what the shadowy figure of the predictor does or can do. In recent years, the predictor has got more attention, and this attention is along the right lines, but has not yet identified the right question, which is this: can your action now in choosing one or both of the boxes have a *causal* influence on the predictor's decision to put £1,000,000 in the red box?

The answer usually given is 'no', because the action of the predictor is in the past, and backwards-causation is impossible. If this is the answer, you are in a genuine Newcomb case, and you should two-box, as the causalist says you should, since you cannot affect the prior actions of the predictor.

But then intuitions do waver. The 100% past success rate of the predictor is very distracting. There is £1,000,000 at stake and the evidentialist cries, 'How come you ain't rich?' Even those previously convinced by the causal reasoning are unsettled and begin to wonder.

But this is not the whole story, because the answer to the crucial question might be 'yes'. That is, presented with what is apparently a Newcomb case, you might still decide that there is reason to believe your action now can have a causal influence on the apparently past action of the predictor. You might legitimately decide this simply because the 100% past success rate of the predictor is so unusual. It is unusual enough that it undermines

your belief that your choice can have no causal influence on the action of the predictor. A 100% success rate in the absence of any causal relation is so unlikely that it challenges the conviction that the action of the predictor is genuinely in the past. In extreme cases, it might challenge your conviction that backwards-causation is impossible.[1] So faced with a real predictor with a 100% success rate, it is not impossible that you would come to believe that there is some cleverly arranged cheating going on. It looks as if you have a case of backwards causation, but really the relation between your choice and the predictor's action is a case of very cleverly disguised but otherwise normal forwards causation. You may be unable to work out how it could be done, but still think that the 100% success rate is enough evidence for you to believe that it is being done somehow.[2]

The right way to approach the Newcomb problem is to attempt to work out the underlying causal structure, just as the causalists prescribe. You must decide whether or not there is a concealed causal connection. And for either possible answer on the underlying causal structure, the causalist prescribes the right choice. If you still think there *must* be no causal connection since the action of the predictor really is in the past, you should two-box. Alternatively, if you think there probably is some cheating going on undetected by you, then you think there probably is a causal connection, and you should one-box. That is why there is no set answer to the Newcomb problem. There are two possible answers and always will be, because the right choice depends on extra information about the actions of the predictor not given in standard descriptions of the case. And it is the causalist who tells you what information you need, and what to do with it.

Intuitions waver in Newcomb cases because the one-box choice is attractive. Now that I have shown that this is due to implicit *causal* reasoning based on the 100% success rate of the predictor, the Newcomb problem will no longer draw people into evidentialism.

The University of Stirling
Stirling, FK9 4LA, UK
p.k.mckay@stir.ac.uk

References

Carlson, E. 1998. Fischer on backtracking and Newcomb's problem. *Analysis* 58: 229-31.
Fischer, J. M. 2001. Newcomb's problem: a reply to Carlson. *Analysis* 61: 229-35.

[1] I will ignore this since introducing the possibility of genuine precognition brings in a whole host of new problems.

[2] I think that recent concerns about the counterfactual dependence of the predictor's (apparently) past choice on your current choice and whether that counterfactual dependence 'backtracks' would be better directed at wondering whether this counterfactual dependence implies a causal relation. See for example Carlson, and Fischer.