WHY YOU SHOULD ONE-BOX IN NEWCOMB'S PROBLEM

by

HOWARD J. SIMMONS

July, 2015

Here is the problem. You have in front of you a transparent box, in which you can see $10. There is also another box, but you do not know what is in it, as it is opaque. You are told that you have a choice between opening both boxes ('two-boxing') or opening the opaque box only ('one-boxing') and you get to keep whatever is in the box or boxes that you open. The twist is this: if you choose to open the opaque box only, then an individual called the Predictor will have predicted that you are going to do this and will have put $100 into the opaque box, but if you choose to open both boxes, she will *not* have put $100 in the opaque box. Knowing all this, what should you do? We assume that maximising your wealth is your only goal.

Many people think that you should one-box. Anybody who one-boxes will find that there is a hundred dollars in the opaqe box, as the Predictor will have known that she was going to one-box and will consequently have put the hundred in there. Anybody who two-boxes, on the other hand, will get the $10 in the transparent box, but nothing in the opaque box, as in that case, the Predictor will have left the opaque box empty. One-boxers end up much better off than two-boxers. Assuming that being as well off as possible is all that matters, why would a rational agent want to do anything else?

But it has to be admitted that there is something odd about one-boxing. If I one-box because I am convinced by the above argument, then in effect I act in order to bring it about that the Predictor has predicted that I will one-box and so put $100 into the opaque box. This seems peculiar, paradoxical even. Causality surely only ever runs forward. We can't reach back into the past and determine the way things *have* been.

Suppose we agree for the moment that this line of thinking is correct. Can we say what we ought to do in Newcomb's problem? Here is a popular argument for resolving the quandry. Either there is $100 in the opaqe box or there isn't. Suppose the $100 is there. Then if I one-box, I will get that $100. But if I two-box, I will get the contents of the transparent box *as well*, and since there is $10 in that box, I will finish up with a total of $110. In other words, I do better by two-boxing than I do by one-boxing. But suppose instead that the opaque box is empty. Then if I one-box I get nothing, whereas if I two-box I get $10. So again I do better by two-boxing. Hence, *whatever* the contents of the opaque box, I do better by two-boxing than I do by one-boxing. Therefore, it might be claimed, two-boxing, not one-boxing, is the rational thing to do.

When one option gives the agent a higher gain than another, whatever the situation, the former option is said to *dominate* the latter. The argument just given for the rationality of two-boxing in Newcomb's problem appeals to dominance. But is the appeal to dominance always a sound one?

In this argument, each possibility—that there is $100 in the opaque box and that there isn't $100 in there—is considered in turn and it is shown that in either event, the agent does better by two-boxing. It may help to evaluate the argument if we consider how we should react if it is applied to a different situation. This situation (call it the 'future oriented Newcomb case') is like the Newcomb situation, but the Predictor is not involved. I can push one of two buttons, A and B. Button A will open the opaque box, while button B opens both boxes. As in the standard Newcomb case, the transparent box visibly contains $10, while the contents of the opaque box depends on my decision. But the mechanism of this dependence is different. Quite simply, when I push button A, $100 is automatically inserted into the opaque box, just before that box is opened to enable me to retrieve the money. (Button B, in contrast, *merely* opens both boxes, leaving the contents of the opaque box empty.)

This case is a no-brainer. Nobody would seriously contend that it would be rational to two-box. In

this situation $100 is there for the taking by one-boxing (or rather *will* be there for the taking if I choose to take it), whereas two-boxing yields only $10. But what is interesting is that the dominance argument in favour of two-boxing can still be constructed. Either the opaque box *will* contain $100 or it won't. In the former case, one-boxing yields $100, but two-boxing yields $110. In the latter, one-boxing yields nothing, but two-boxing yields $10. In both cases, then (so the argument goes), the agent does better by two-boxing, just as in the original Newcomb situation. But that is a ridiculous conclusion. Clearly, dominance alone cannot be relied on to select the rational course of action.

An interesting aspect of the dominance argument is that it does not conform to the normal Baysian procedure for determining rational action (i.e., the action which it is rational to perform). That procedure is as follows:

For each possible act:

> for each possible outcome that might occur if the act were performed:
>
> > multiply the utility of that outcome by the probability that the outcome will occur if the act is performed.
>
> Find the sum of all these products (known as the *expected utility*).

The act(s) which it is/are rational to perform is/are the one(s) with the highest expected utility.

The implicit basis of the dominance argument is different, although it does incorporate the Baysian procedure. It starts by considering a *partition*—a mutually exhaustive and exclusive set of alternative states of affairs (there being $100 in the opaque box and there not being $100 in the opaque box) and then detemines that, for each such state of affairs, the expected utility (calculated using the Baysian procedure described above) of one of the acts (two-boxing) must be higher than the alternative act (one-boxing). Since the set of alternative states of affairs was exhaustive (there either is or there is not $100 in the opaque box), it seems to follow that two-boxing is the rational action.

This procedure (call it the *disjunctive* procedure) is sometimes valid. Suppose that in the future oriented Newcomb's problem, in addition to the money that you receive from opening one or both of the boxes, you will receive either $50 or $40, but you have no idea which. This is all you know. Then it would be legitimate to use the disjuctive procedure, considering in turn the case where you get $50 and the case where you get $40 and calculating, for each possiblity what you could be expected to get in total, depending on whether you one-box or two-box. But this is only possible because, for all that we have been told, the probability of my getting the $50 and the probability of my getting the $40 is unaffected by anything that I do. In contrast, the probability of there being $100 in the opaque box *is* affected by what I do and it is absurd to leave this out of account. Likewise in the original Newcomb problem, the probability of there being $100 in the opaque box is affected by what I do (albeit for a reason that is somewhat bizarre) There is as much reason for taking this into account in the latter case as in the former. So an application of the disjunctive procedure, in which we repeat the Baysian calculation for each possible amount in the opaque box, is indefensible. We should just use the normal Baysian procedure. This gives us an expected utility of $100 for one-boxing and $10 for two-boxing, which tells us that one-boxing is the rational course of action.

What about the counter-intuitiveness of one-boxing noted earlier, the fact that it seems to involve acting so as to try to cause a past event? Well perhaps we don't have to describe it that way. Certainly we one-box *so that* the Predictor will have put the hundred in the opaque box, but it is not clear that we have to refer to this as *causing* the Predictor to do this.[1] If this is right, then endorsing one-boxing is not necessarily to recognise backwards causation. On the other hand, if we were to conclude that one-boxers *are* committed to backwards causation, I personally would rather do this than not be a one-boxer, as the latter would seem to me to be more counter-intuitive, not less.

APPENDIX: PRIEST'S ARGUMENT FOR TWO-BOXING

Graham Priest has produced an argument for two-boxing that is distinct from the dominance argument (Priest (2002), 13), though I think it too is mistaken. Let *c* mean 'whatever is now in the opaque box', where this expression is used as a rigid designator. Now if I one-box, I get *c*, whereas if I two-box, I get *c* plus $10. *c* plus $10 must be greater than *c* and so this again implies that two-boxing is rationally required.[2]

This argument, however, fails. It is indeed perfectly possible to use the phrase 'whatever is now in the opaque box' as a rigid designator. If so, the phrase refers to a specific amount of money, whatever amount is actually in that box, and it does so in respect of all the possible situations that may be created by making one or other decision—the decision taken doesn't affect what amount *c* is. But if so, then a tacit assumption of the argument, which one would normally accept as incontrovertible, is undermined. This is the principle that if the agent chooses to open a particular box, she will necessarily get (at least) whatever is in that box: for example, if she opens the opaque box, she will get at least whatever is in the opaque box. (We can add of course that if she opens the transparent box *as well*, she will also get an extra $10.) This normally quite innocent principle has to be rejected if we understand the designating phrase 'whatever is in the opaque box' rigidly. For understood in that way, it will no longer be true that we will get something different depending on whether we open just the opaque box or both boxes. Suppose it happens to be true now that the opaque box contains $100. Then 'whatever is in the opaque box' refers to $100 even in possible situations in which I two-box. But it is agreed all round that if I two-box, I get only $10. So I do *not* get at least whatever is in the opaque box, where the latter phrase is used as a rigid designator. There is nothing paradoxical about this. It is just a consequence of the way rigid designators work. The air of paradox is due only to the fact that we would *not* normally use the phrase rigidly in that context.

Once it is accepted that with the phrase 'whatever is in the opaque box' understood rigidly, we cannot assume that an agent choosing to two-box gets at least whatever is in the opaque box, we can no longer suppose that such an agent must get whatever is in the opaque box plus $10 and therefore cannot conclude that she does better by two-boxing than she would have done by one-boxing. Priest's argument then collapses.

1 John Leslie calls it 'quasi-causing' (Leslie 1991).

2 However, Priest also accepts that one-boxing is rationally required, for familiar reasons. He thus thinks that Newcomb's problem is a *rational dilemma*, in which there is no uniquely rational decision.

BIBLIOGRAPHY

Leslie, John: 1991. “Ensuring Two Bird Deaths with One Throw”, *Mind*, 100(1): 73-86.

Priest, Graham: 2002. “Rational dilemmas”, *Analysis*, 62(1): 11-16.