Against a Newcomb Dominance Argument*

Daniel J. Singer
singerdj@umich.edu

In the following discussion, I will put forth an intuitively appealing argument for two-boxing in the Newcomb problem[1] which employs dominance reasoning. I will then suggest a potential issue with this argument as formulated. I'd appreciate any feedback here, especially reformulations of the two-box argument that stick to the intention of the original but employ different notions of dominance.

1 Statewise Dominance

One restraint on any reasoner with monotonic preferences is the following:

Statewise Dominance A strategy A *statewise dominates* a strategy B if for every epistemically possible state of affairs, the actor prefers the outcome produced by A at least as much as the actor prefers the outcome of B, and there is at least one epistemically possible state of affairs where the actor prefers the outcome of A strictly more than the outcome of B. If some strategy A' dominates another strategy B', then the actor prefers A' to B'.

Essentially, one strategy dominates another strategy iff the former strategy is at least as good as the latter in every possible state of the world and strictly better than the former in at least one possible state of the world. I take this to be uncontroversial as a constraint on a reasoner with monotonic preferences.

2 Dominance, Newcomb and Problems

One very intuitively plausible argument in the Newcomb case with an infallible predictor goes something like this: The million dollars is there or it's not; either way, taking both boxes is at least as good as taking just the one, so I should two-box. We can flesh out this argument more as follows: The argument is presented in the first person where the speaker is the actor in a Newcomb problem.

*Much of the thinking that went into this was inspired by David Wiens, Stephen Campbell, Shenyi Liao, and Jason Konek.

[1]Here I will assume a standard formulation of the Newcomb problem with an infallible predictor, which can be found here: http://en.wikipedia.org/wiki/Newcomb%27s_paradox

1. I can decide to take one box or two boxes. (Premise of the NP.)
2. My choice will not causally affect the contents of the boxes. (Premise of the NP.)
3. In every epistemically possible state of the world, either there is a million dollars in the opaque box or there is nothing in the opaque box.
4. If there is a million dollars in the opaque box, I strictly prefer to take two boxes.
5. If there is nothing in the opaque box, I strictly prefer to take two boxes.
6. Hence, taking two boxes statewise dominates taking one box. (by 3-5)
7. Hence, I prefer to take two boxes to taking one box.
8. Therefore, I should take two boxes.

This type of argument is intuitively very compelling, and I was briefly taken with its merit. I think I have identified a source of concern here though.

My concern here is that the conclusion that two-boxing statewise dominates one-boxing in line 6 does not follow from lines 3-5 as it should. To show this though, I must rely on the following principle that I take to be a reasonable constraint on rational preferences:

Preference Possibility If an ideal actor prefers A to B, then if B is metaphysically or epistemically possible, then A is metaphysically or epistemically possible.

I feel confident that a stronger constraint is probably true (something like: if an ideal actor has a preference for A, then A is metaphysically or epistemically possible), but I am not willing to defend such a constraint here. Instead, to bring out a problem with the above argument, it is sufficient to show the following two claims:

Claim. *It is not the case that if there is a million dollars in the opaque box, I strictly prefer to take two boxes, i.e. premise 4 is false.*

Proof. Let Σ be a possible world (to the Newcomb-problem world) in which there is a million dollars in the opaque box. Certainly, in Σ, it is metaphysically possible that I one-box; this is due to the condition of the Newcomb problem that requires that if there is a million dollars in the opaque box, then the infallible predictor predicted that I would one-box, and since the predictor is infallible, I must actually one-box. But, in Σ, if I were to two-box, then the predictor would have predicted that I would two-box, so there would not be a million dollars in the opaque box. $\Rightarrow\Leftarrow$ So, it is not metaphysically possible that I two-box. Also, because I am aware of the Newcomb constraints, I know that I cannot two-box. So, in Σ, it is neither metaphysically nor epistemically possible that I two-box. Therefore, by the preference possibility principle, I do not prefer two-boxing to one-boxing in Σ. □

Claim. *The conclusion that two-boxing statewise dominates one-boxing does not follow by the statewise dominance principle from only lines 5 and 3.*

Proof. The conclusion that two-boxing statewise dominates one-boxing would follow by the principle only if for every epistemically possible state of affairs, the actor prefers the outcome produced by two-boxing at least as much as the actor prefers the outcome of one-boxing. By the assumptions of the Newcomb problem, there is an epistemically possible state of affairs (since the actor in the Newcomb problem knows the setup of the problem) in which there is a million dollars in the opaque box.[2] Lines 5 and 3 do not show that in this possible state of affairs, the actor prefers two-boxing to one-boxing, so lines 5 and 3 do not entail that two-boxing statewise dominates one-boxing. □

It seems like the main problem here that makes the argument fail is that in the case where the million dollars is present, the actor is forced to compare his preference for the million dollars to his preference that a contradiction obtain (that is, that he two-box with the million dollars). This is clearly a nonsensical comparison, and does not entitle the arguer of the dominance argument to the conclusion that two-boxing dominates in this case. Hence, the argument above from dominance does not entail the conclusion that the actor in an infallible predictor Newcomb problem should two-box. Also, it seems that an analogous refutation of this type of argument could be run by looking closer at the comparison between two-boxing in the case where the million dollars is not present and the case where a contradiction is instantiated by one-boxing with no money.

Thoughts?

[2]One may contest that no ideal actor would one-box, so it may not be possible that there is a million dollars present. But, in general, the Newcomb problem does not require ideal actors, so we could at least imagine a very stubbern one-boxer.