# *The Meta-Newcomb Problem*

NICK BOSTROM

[Published in *Analysis*, 2001, Vol. 61, No. 4., pp. 309-310.]

Consider the following twist of the Newcomb problem.

> *Meta-Newcomb*
>
> There are two boxes in front of you and you are asked to choose between taking only box *B* or taking both box *A* and box *B*. Box *A* contains $ 1,000. Box *B* will contain either nothing or $ 1,000,000. What *B* will contain is (or will be) determined by Predictor, who has an excellent track record of predicting your choices. There are two possibilities. Either Predictor has already made his move by predicting your choice and putting a million dollars in *B* iff he predicted that you will take only *B* (like in the standard Newcomb problem); or else Predictor has not yet made his move but will wait and observe what box you choose and then put a million dollars in *B* iff you take only *B*. In cases like this, Predictor makes his move before the subject roughly half of the time. However, there is a Metapredictor, who has an excellent track record of predicting Predictor's choices as well as your own. You know all this. Metapredictor informs you of the following truth functional: Either you choose *A* and *B*, and Predictor will make his move after you make your choice; or else you choose only *B*, and Predictor has already made his choice. Now, what do you choose?

"Piece of cake!" says a naïve non-causal decision theorist. She takes just box *B* and walks off, her pockets bulging with a million dollars.

But if you are a causal decision theorist you seem to be in for a hard time. The additional difficulty you face compared to the standard Newcomb problem is that you

don't know whether your choice will have a causal influence on what box *B* contains. If Predictor made his move before you make your choice, then (let us assume) your choice doesn't affect what's in the box. But if he makes his move after yours, by observing what choice you made, then you certainly do causally determine what *B* contains. A preliminary decision about what to choose seems to undermine itself. If you think you will choose two boxes then you have reason to think that your choice will causally influence what's in the boxes, and hence that you ought to take only one box. But if you think you will take only one box then you should think that your choice will *not* affect the contents, and thus you would be led back to the decision to take both boxes; and so on *ad infinitum*.

Yale University
New Haven, CT 06520-208306, USA
nick@nickbostrom.com
*Web*: www.nickbostrom.com