

Princeton Philosophical Society: Problems in Decision Theory

Newcomb's Problem¹:

“There are two closed boxes on the table, Box A and Box B. Box A contains \$1,000. Box B contains either \$1 million or no money at all. You have a choice between two actions: 1) taking what is in both boxes; or 2) taking just what is in Box B.

“Now here comes the interesting part. Imagine a Being that can predict your choices with high accuracy. You can think of this Being as a genie, or a superior intelligence from another planet, or a supercomputer that can scan your mind, or God. He has correctly predicted your choices in the past, and you have enormous confidence in his predictive powers. Yesterday, the Being made a prediction as to which choice you are about to make, and it is this prediction that determines the contents of Box B. If the Being predicted that you will take what is in both boxes, he put nothing in Box B. If he predicted that you will take only what is in Box B, he put \$1 million in Box B. You know these facts, he knows you know them, etc. So, do you take both boxes, or only Box B?

“Well, obviously you should take only Box B, right? For if this is your choice, the Being has almost certainly predicted it and put \$1 million in Box B. If you were to take both boxes, the Being would almost certainly have anticipated this and left Box B empty. Therefore, with very high likelihood, you would get only the \$1,000 in Box A. The wisdom of the one-box choice seems confirmed when you notice that of all your friends who have played this game, the one-boxers among them are overwhelmingly millionaires, and the two-boxers are overwhelmingly not.

“But wait a minute. The Being made his prediction yesterday. He either put \$1 million in Box B, or he didn't. If it's there, it's not going to vanish just because you choose to take both boxes; if it's not there, it's not going to materialize suddenly just because you choose only Box B. Whatever the Being's prediction, you are guaranteed to end up \$1,000 richer if you choose both boxes. Choosing just Box B is like leaving a \$1,000 bill lying on the sidewalk. To make the logic of the two-box choice even more vivid, suppose the backs of the boxes are made of glass and your wife is sitting on the other side of the table. She can plainly see what's in each box. You know which choice she wants you to make: Take both boxes!”

Kavka's Toxin Puzzle²:

“[An eccentric billionaire] places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. (. . .) The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. (. . .) All you have to do is. . . intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin.”

¹ From <http://www.slate.com/?id=2061419>

² From http://en.wikipedia.org/wiki/Kavka's_toxin_puzzle