## Newcomb's Paradox[1]

Rabern
PUZZLES AND PARADOXES
*University of Edinburgh*

PUZZLE. The Predictor is a being who is able to predict your choices with great accuracy. The Predictor has accurately predicted your choices in the past, and has accurately predicted the choices of others in the past. The Predictor has also had great success at making predictions about the choices people make in the following situation: There are two boxes, a transparent box $b_1$ and an opaque box $b_2$. You can see that $b_1$ contains $1,000; $b_2$ contains either $1,000,000 or nothing. You have two choices: either you take what is in both boxes, or you take what is in the opaque box $b_2$ alone. If the Predictor predicted that you will take what is in both boxes, he does not put $1,000,000 in $b_2$; but if the Predictor predicted that you will take what is in $b_2$ alone, he puts $1,000,000 in $b_2$. You value more money to less money. What should you do?[2]

When Newcomb's Problem was presented in Scientific American in 1974, of the first 148 letters, 89 people said they would take one box, while only 37 people thought two boxing was the correct option.

> "To almost everyone, it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly." (Robert Nozick, 1969)

**One-boxing.** One-boxing—taking $b_2$ alone—is a good way of ensuring that you get $1,000,000. Most people who have taken $b_2$ alone have been correctly predicted to do so, and thus most people who have taken $b_2$ alone have walked away millionaires. Whereas, on the other hand most people who have taken both boxes have also correctly been predicted to do so, and thus have walked away with a mere $1,000. That is: If one selects both boxes, $b_2$ will almost certainly be empty. If one selects just $b_2$, it will almost certainly contain $1,000,000. Since we want to get the most money it seems we ought to take $b_2$ alone.

In fact, this reasoning is supported by an influential view on how to calculate the *expected utility* of an action. An agent is concerned with certain outcomes whose importance can be measured by a number; this is the outcome's **utility**. And the **expected utility** of an action is the sum of those utilities each multiplied by the agent's degree of belief that the action will lead to that outcome. Let $A_1$ be the proposition that the agent takes $b_2$ alone; let $A_2$ be the proposition that the agent takes both boxes; and let $M$ be the proposition that there is $1,000,000 in $b_2$. Then, the evidential expected utility of taking $b_2$ alone is (where $V(X)$ is the evidential expected utility of $X$):

$$\begin{aligned} V(A_1) &= P(M \mid A_1) \times u(M \wedge A_1) + P(\neg M \mid A_1) \times u(\neg M \wedge A_1) \\ &= (0.99 \times \$1,000,000) + (0.01 \times \$0) \\ &= \$990,000 \end{aligned}$$

[1] The case is due to the physicist William Newcomb, but was introduced into the literature by Robert Nozick (1969) "Newcomb's Problem and Two Principles of Choice", in N. Rescher, *Essays in Honor of Carl G Hempel*. My description of the problem and the rationale for each choice follows the discussion in Dilip Ninan (2006) "Illusions of Influence in Newcomb's Problem" (http://www.dilipninan.org/papers/newcomb.pdf).

[2] We can be more specific about the Predictor's success. The Predictor has made two thousand predictions prior to your choice. Of those, one thousand people took one box and the Predictor predicted this 99% of the time; the other thousand took two boxes and the Predictor predicted this 99% of the time.

$$\begin{aligned} V(A_2) &= P(M \mid A_2) \times u(M \wedge A_2) + P(\neg M \mid A_2) \times u(\neg M \wedge A_2) \\ &= (0.01 \times \$1,001,000) + (0.99 \times \$1000) \\ &= \$11,000 \end{aligned}$$

**Two-boxing.** Either the Predictor has predicted that you'll take one box or he predicted you'll take two boxes. Suppose he has predicted that you'll take one box. Then there's $1,000,000 in $b_2$ and $1,000 in $b_1$. If you take both, you'll get $1,001,000; if you take $b_2$ alone, you'll get $1,000,000. So you should take both boxes. Suppose now that the Predictor has predicted that you'll take both boxes. That means there is nothing in $b_2$. If you take both, you'll get $1,000; if you take $b_2$ alone, you'll get nothing. So you should take both boxes. So no matter what the Predictor has predicted, you'll be better off if you take two boxes rather than one.

The justification for taking both boxes is a game theory principle called the **dominance principle**. A strategy is said to be dominant when it is always the more beneficial strategy regardless of what your opponent does.

| | prediction one-box | prediction two-box |
|---|---|---|
| take one-box | $1,000,000 | $ 0 |
| take two-box | $1, 001, 000 | $1000 |

It is your turn to choose. **What should you do?**