Newcomb's Problem and an argument for 1-boxing

Part 1

Newcomb's problem has to do with the idea of causation and probabilities. It involves two boxes, A and B and a choice between '1-boxing' and '2-boxing'. Choosing to 1-box entails obtaining the contents of box B while choosing to 2-box entails obtaining the contents of boxes A and B. The contents of box A is $1,000 and the contents of box B depends on a prediction made by a predictor P, assumed to be 99% reliable. Box B would be filled with $1,000,000 or $0 depending on whether P predicts 1-boxing or 2-boxing respectively[1]. The filling of box B is done before the choice of n-boxing[2] and it is assumed that the contents of the boxes are not tampered with once the prediction is made. Here is a table of the payoffs in the different cases:

v(x,y)	**X=E**	**X=F**
Y=2B	$1,000	$1,001,000
Y=1B	$0	$1,000,000

i. X is a variable that characterises the state of box B, E signifies that box B is empty and F signifies that it is filled
ii. Y is a variable that characterises whether one chooses to 1-box or 2-box
iii. v(X,Y) stands for the amount gained as a result of choosing Y when box B is in state X

If not for P, it is straightforward to choose to 2-box as it 'dominates' 1-boxing; in the sense that 2-boxing always ensures the larger amount. However, when the predictor P comes in, an argument based on probabilities seems to support 1-boxing. The idea is that 1-boxing is the choice that maximises expected value and, thus, should be the preferred one. The 'value' of the choice is simply taken to be its $ value and the probabilities are as so:

[1] The specific rewards and accuracy of predictor are of no special consequence but are rather used to allow a clearer outline of the matter.

[2] A generic term used for either 1-boxing or 2-boxing.

$p(Y \rightarrow X)$[3]	X=E	X=F
Y=2B	0.99	0.01
Y=1B	0.01	0.99

$$\begin{aligned} E(2B) &= v(E,2B)\, p(2B \rightarrow E) \;+\; v(F,2B)\, p(F \rightarrow 2B) \\ &= \$1000\ 0.99 \;+\; \$1{,}001{,}000\ 0.01 \;=\; \$11{,}000 \end{aligned}$$

$$\begin{aligned} E(1B) &= v(E,1B)\, p(E \rightarrow 1B) \;+\; v(F,1B)\, p(F \rightarrow 1B) \\ &= \$0\ 0.01 \;+\; \$1{,}000{,}000\ 0.99 \;=\; \$990{,}000 \end{aligned}$$

and thus expected value maximisation argues that one should 1-box.

A causal decision theorist might argue that there is no causal link between n-boxing and the contents of the boxes and thus it always makes sense to choose the option that would guarantee the larger amount, 2-boxing. They might argue that what is relevant here are not the indicative conditional probabilities used in the preceding analysis but rather the subjunctive conditional probabilities. They claim, for example, that what matters is not the probability of the box being full given that 2-boxing is chosen but rather the probability that the box would be full had you to choose to 2-box, in other words, a causal link is sought. In this alternate codification,

$$\begin{aligned} EV(2B) &= v(E,2B)\, p(2B \Box\!\!\rightarrow E) \;+\; v(F,2B)\, p(1B \Box\!\!\rightarrow F)^{4} \\ &= \$,1000\, p(E) \;+\; \$1{,}001{,}000\, p(F) \end{aligned}$$

$$\begin{aligned} EV(1B) &= v(E,1B)\, p(1B \Box\!\!\rightarrow E) \;+\; v(F,1B)\, p(1B \Box\!\!\rightarrow F) \\ &= \$0\, p(E) \;+\; \$1{,}000{,}000\, p(F) \end{aligned}$$

since there is no causal link between n-boxing and the contents of box B, and so $p(2B \Box\!\!\rightarrow E) = p(1B \Box\!\!\rightarrow E) = p(E)$, the prior probability that the box is empty. It is thus clear to a causal decision theorist that 2-boxing is the better alternative, as irrespective of p(F) and p(E), EV(2B)>EV(1B).

The seeming contradiction might be because of an incorrect use of the idea of causation. To begin with, it is assumed that there exists predictor P that has an accuracy of 99% and later on it is assumed that these probabilities ought not to be

[3] $p(X \rightarrow Y)$ is the conditional probability of Y given X

[4] $p(X \Box\!\!\rightarrow Y)$ is the subjunctive probability of x given y, that is, it is the probability that X would lead to Y

used as there is no causal link between the probabilities and the choice to n-box. This is a contradiction and might be at the core of the reason for the inconsistency. In part 2 I shall attempt to show that indicative conditionals are to be used whether or not there is an argument about causality.

Part 2

The crucial point in the discussion is the probabilities and this is where the two approaches diverge. I believe that the correct probabilities to use are the indicative conditionals irrespective of whether or not there is a causal link and this is why using an argument of causal dependence leads to inconsistency.

To see why this is so, the definition of the probability has to be dissected. The indicative probabilities are constructed by considering all the possibilities leading to the different circumstances and then making a statement about how relatively likely they are. So, the statement $p(F \rightarrow 2B) = 0.01$ implies that among all the different possible ways in which the situation may evolve, the box would be full given that you choose to 2-box one hundredth of the times. There is no mention of causality and neither is it required, all that is relevant is that there be a suitable P; that is, the laws of the universe entail that there is a predictor P that achieves an accuracy of 99% in the n-boxing cases. There is the issue of the plausibility of the existence of such a predictor, but it is irrelevant to the current discussion and shall be dealt with later.

This development parallels that of the grandfather paradox and a brief discussion shall show how. Bruno goes back to the past and attempts to kill his grandfather some time before his father is born. In order to maintain consistency, Bruno must fail because otherwise he would not have been born to carry out the said task. The response is that the laws of physics ensure that Bruno would not be able to carry out such a task. It is a mistake, however, to assume that the laws 'actively' intervene to prevent Bruno from carrying out the task; it is rather that there is no consistent outcome of the laws of physics in which Bruno is able to carry out the said task. The situation is observed from 'outside' the universe in a sense. In a similar manner, the laws of this universe entail that there is P which is able to predict n-boxers with 99% accuracy. And thus, even if the decision to n-box cannot 'actively' change the contents of box B, the indicative probabilities still hold true on account of the universe entailing the existence of P.

There is also the issue raised about wet sidewalks and umbrellas, that one's decision to take an umbrella does not causally affect the weather. This case is distinct from the case of n-boxing as there is no predictor involved. There is obviously no causal link between taking out one's umbrella and it raining but there is also no analogue of P here that affects the rain. Imagine that the rain was actually caused by a god who predicts beforehand (with 99% accuracy; though powerful the god isn't omniscient) whether somebody would bring out their umbrella and would let it rain only in case she predicted they would bring it out. In such a case it would make sense to not carry out one's umbrella even if the god's decision is not causally controlled by one's decision to take out an umbrella. A similar argument would work for the case of *mathematosis* and a gene that predicts beforehand whether one would be inclined to do mathematics.

The key issue might be the assumed power of the predictor. It not only involves an analysis of one's first order thought but also their second, third and nth order thoughts in order to arrive at its decision[5]. It is, however, not difficult to see that such predictors are not as impossible as one might initially believe. Probability theory is a fairly well established branch of mathematics and its achievements in multiple fields well documented. The assumptions used in these arguments are not too far off the arguments used in many successful applications of probability theory and it thus can be argued that they are not unreasonable.

Concluding, an argument was made for 1-boxing and it is hoped that the reader is left with some sense of an inclination towards it. Much deliberation would be required before putting this matter to rest but it is hoped that this article might play some small part in dealing with Newcomb's problem.

References

Lectures 4 and 5, MITx: 24.118x 'Paradox and Infinity'
Newcomb's Paradox, wikipedia.org

[5] A second order thought is a thought about a first order thought, for example, the thought about whether thinking about the Newcomb predictor will affect its outcome. Third order thoughts are thoughts about second order thoughts and so on.