Newcomb's Problem and Entanglement

Vishal Johnson

August 2019

Newcomb's Problem is a game involving two entities; there is a contestant(C) who takes part in the game and a predictor(P) who predicts the actions of the contestant[4]. Two boxes are brought before C, a small box and a large box, and C is given the choice to either select both boxes(2-box) or to select only the large box(1-box). Whereas the small box definitely contains $1,000, the large box either contains $0(empty, E) or $1,000,000(full, F). P is asked to predict in advance whether C would decide to 2-box or 1-box; and the contents of the large box is filled according to P's prediction. The large box is filled with $0(E) if P predicts that C would 2-box and with $1,000,000(F) if P predicts that C would 1-box[1].

1 Part II

C's decision to 1-box or 2-box strongly depends on P's predictive ability. Start with the case where P simply tosses a fair coin to make the decision. What should C decide to do? Consider the expected value of C's choices.

$$\begin{aligned}\mathcal{E}(2-box) &= v(2-box,E)p(E|2-box)+v(2-box,F)p(F|2-box)\\ \mathcal{E}(1-box) &= v(1-box,E)p(E|1-box)+v(1-box,F)p(F|1-box)\end{aligned} \tag{1}$$

The various values and probabilities are given by:

v(X,Y)	Y=E	Y=F
X=2-box	$1,000	$1,001,000
X=1-box	$0	$1,000,000

(2)

p(Y | X)	Y=E	Y=F
X=2-box,1-box	0.5	0.5

(3)

And thus, the expected values are:

$$\begin{aligned}\mathcal{E}(2-box) &= \$1,000\times 0.5+\$1,001,000\times 0.5=\$501,000\\ \mathcal{E}(1-box) &= \$0\times 0.5+\$1,000,000\times 0.5=\$500,000\end{aligned}$$

[1] it's critical that P makes the prediction well in advance of C playing the game. The events of P's prediction and C's game must be causally non-connected. That is, P's prediction must lie in the past light cone of C's decision.

And it's thus clear that 2-boxing is the preferred alternative as it has the higher expected value. In fact, this remains true even if one tossed a biased coin, as long as the probability of the large box being filled is independent of choosing to 1-box or 2-box. 2-boxing, in this sense, dominates 1-boxing.

How about the case of a perfect predictor? One way to arrange for a perfect predictor is to use an entangled pair of qubits[1].

Remark. *A qubit could be thought of as a quantum mechanical switch. While a classical switch can be in only one of two states, say $|\mathcal{ON}\rangle$ and $|\mathcal{OFF}\rangle$. A qubit, however, could be in a superposition state such as $\sqrt{p}\,|\mathcal{ON}\rangle + \sqrt{1-p}\,|\mathcal{OFF}\rangle$; to be in this superposition state is to measure the state $|\mathcal{ON}\rangle$ and $|\mathcal{OFF}\rangle$ with probabilities p and $1-p$ respectively.*

Another quantum property is that of entanglement[1]. Just as two switches could be 4 states, two qubits could be in a superposition of 4 states. If the superposition is chosen in a special way, there is entanglement. Consider the state $\sqrt{p}\,|\mathcal{ON}\rangle_1 \otimes |\mathcal{ON}\rangle_2 + \sqrt{1-p}\,|\mathcal{OFF}\rangle_S \otimes |\mathcal{OFF}\rangle_2$. This basically means that with probability p both qubits are in a state $|\mathcal{ON}\rangle$ and with probability $1-p$ both in state $|\mathcal{OFF}\rangle$. Thus, the state of one half of the pair confirms the state of the other half. There could be a causality-like effect between the states of two switches without the time travel or retro-causality.

Prepare an entangled state:

$$|\psi\rangle_S \otimes |\psi\rangle_P = \sqrt{p}\,|2-box\rangle_S \otimes |E\rangle_P + \sqrt{1-p}\,|1-box\rangle_S \otimes |F\rangle_P \tag{4}$$

In this way it's ensured that there is complete correlation between P's prediction and S's action.

p(Y | X)	Y=E	Y=F
X=2-box	1	0
1-box	0	1

(5)

$$\begin{aligned}\mathcal{E}(2-box) &= \$1,000\\ \mathcal{E}(1-box) &= \$1,000,000\end{aligned}$$

The overall expected value is:

$$\mathcal{E}(p) = \$1,000 \times p + \$1,000,000 \times (1-p) \tag{6}$$

As can be seen from the expected values of 2-boxing and 1-boxing, it's much better to 2-box. More convincingly, the expected value as a function of p clarifies that it's much better to tweak p so as to bias 2-boxing.

It might seem that S does not act freely in this situation as S's actions are guided by the entangled qubit in S's possession. The argument is that p and S's half of the entangled qubit form the psychological state of S[2]. Thus,

[2] The psychological state of S with respect to Newcomb's game.

p represents the probability that S would 2-box. What about P's half of the entangled qubits? P's half represents the information carried by the environment about S's psychological state. Insofar as there is perfect entanglement between S's psychological state and P's knowledge thereof, it's definitely better to 1-box. And thus, there is no contradiction with free will. It's not the case that S does not act freely in making the decision to 2-box or 1-box, just that the laws of this universe are such that perfect entanglement between the state of S's mind and of P's knowledge ensures that there is perfect correlation between prediction and action.

In order for this situation to accurately model reality, there are two points that need to be clarified. Firstly, there is the issue that S might change their psychological state just before making the decision. This is easily handled by assuming that the psychological state in question is the state just as S makes the decision. There is no apparent violation of causality here, the assumption is that P has enough knowledge about S to accurately predict S's psychological state whilst making the decision. There is a entanglement between S's psychological state while making the decision and P's knowledge thereof.

Secondly, It's not reasonable to assume that there is perfect entanglement between S's psychological state and P's knowledge of it, there would be an error. Assume that there is an error of ϵ in the entanglement between the states.

The state is written as,

$$\begin{aligned} |\psi\rangle_S \otimes |\psi\rangle_P =& \sqrt{p} \times (\sqrt{1-\epsilon}\,|2-box\rangle_S + \sqrt{\epsilon}\,|1-box\rangle_S) \otimes |E\rangle_P \\ &+ \sqrt{1-p} \times (\sqrt{1}\,|2-box\rangle_S + \sqrt{1-\epsilon}\,|1-box\rangle_S) \otimes |F\rangle_P\,, \end{aligned} \tag{7}$$

where ϵ is considered small.

The probabilities and expected values are,

p(Y | X)	Y=E	Y=F
X=2-box	1-ϵ	ϵ
1-box	ϵ	1-ϵ

(8)

$$\begin{aligned} \mathcal{E}(2-box) &= \$1,000 \times (1-\epsilon) + \$1,001,000 \times \epsilon \\ \mathcal{E}(1-box) &= \$0 \times \epsilon + \$1,000,000 \times (1-\epsilon) \end{aligned}$$

And the overall expected value is:

$$\begin{aligned} \mathcal{E}(p) =& (\$1,000 \times (1-\epsilon) + \$1,001,000 \times \epsilon) \times p \\ &+ (\$0 \times \epsilon + \$1,000,000 \times (1-\epsilon)) \times (1-p) \end{aligned}$$

Entanglement is the key to this entire discussion. The critical assumption is that it's possible for there to be a strong entanglement between S's psychological state and P's knowledge of this state. This is a physical claim. There is confirmed evidence for entanglement and there a strong consensus in the physics community regarding the fact that entanglement exists[1]. However, the claims made above are stronger; it's implied that there is entanglement between macroscopic entities and this is not such a common view. There is research in this direction however[2][3], and only time will tell whether this is a viable assumption. Personally, I do believe that there can be entanglement between macroscopic entities and that the nature of this entanglement is such as to recommend 1-boxing[3].

2 Part I

The objective of Newcomb's game is to optimise the dollar value of S's decision. In this respect, 2-boxers fail to achieve their objective. While, 2-boxing does optimise a certain function, the Evidential Expected Value according to Causal Decision Theory[4], it fails to optimise the dollar value S's decision. This points to 1-boxing being the optimal choice in the case of Newcomb's Problem.

It would be illustrative to use the Mathematosis[4] example to clarify this issue. Mathematosis is a deadly disease caused by a certain gene. The presence of the gene(G) increase the probability of contracting the disease(D). The gene is also assumed to increasing the probability of doing mathematics(M). It's clarified that doing mathematics does not cause the disease. Thus, there is no causal link between the disease and doing mathematics.

But, it's important to ask whether contracting the disease and doing mathematics are entangled as in Newcomb's Problem. This is again a physical assumption about the nature of the disease and how mathematics is done. In the specified case, it does not seem that there is an analogy. The contraction of the disease and doing mathematics do not seem to be entangled in any way as there there is no "agent" that contracts the disease, only a probability. On the other hand, what seems to happen in this case is that the increase in the *value* of doing mathematics, given that one possess the gene, indirectly leads to an increased likelihood of finding people with mathematosis who like to do mathematics. The details are laid out below.

The psychological state of doing mathematics is related to possessing the gene and to contracting mathematosis, but there is no entanglement. One example of a joint state is:

$$\begin{aligned}
|\psi\rangle_M \otimes |\psi\rangle_G \otimes |\psi\rangle_D = \sqrt{p} \times |0\rangle_M \otimes &(\sqrt{q} \times |0\rangle_G \otimes (\sqrt{1-\delta}\,|0\rangle_D + \sqrt{\delta}\,|1\rangle_D) \\
&+ \sqrt{1-q} \times |1\rangle_G \otimes (\sqrt{\epsilon}\,|0\rangle_D + \sqrt{1-\epsilon}\,|1\rangle_D)) \\
+\sqrt{1-p} \times |1\rangle_M \otimes &(\sqrt{q} \times |0\rangle_G \otimes (\sqrt{1-\delta}\,|0\rangle_D + \sqrt{\delta}\,|1\rangle_D) \\
&+ \sqrt{1-q} \times |1\rangle_G \otimes (\sqrt{\epsilon}\,|0\rangle_D + \sqrt{1-\epsilon}\,|1\rangle_D)).
\end{aligned} \tag{9}$$

[3] Incidentally, the laws of this universe are assumed to be probabilistic. What about the case of a deterministic universe? That is equivalent to setting p to 1 or 0. The argument remains valid; just that the laws could perfectly predict S's actions and P's prediction. In this case, S would be destined to either choose to 2-box or to 1-box and would not be able to do otherwise.

$|0\rangle$ means that one does not possess the property whereas $|1\rangle$ means that one does possess the property. If δ and ϵ are assumed to be small it's clearly seen above that the cases of doing mathematics and not doing mathematics both have the same probabilities of contracting the disease. This is precisely because it's assumed that there is no entanglement between the states of doing mathematics and possessing the disease.

p(Y | X)	Y=$|0\rangle_D$	Y=$|1\rangle_D$
X=$|1\rangle_M$, $|0\rangle_M$	q(1-δ)+(1-q)ϵ	(1-q)δ+q(1-ϵ)

(10)

However, the value table becomes:

v(X,Y)	Y=$|0\rangle_D$	Y=$|1\rangle_D$
X=$|1\rangle_M$	$\mathcal{V}(|\psi\rangle_G)$	\$-1,000,000 + $\mathcal{V}(|\psi\rangle_G)$
X=$|0\rangle_M$	\$0	\$-1,000,000

(11)

It's assumed that there is a large negative value for contracting the disease and that *not doing mathematics* had no intrinsic value. Additionally, the value of doing mathematics depends on the gene G.

As the probability of contracting the disease is independent of the choice to do mathematics, and as the value of doing mathematics in all cases dominates not doing mathematics, it makes sense to always do mathematics. At the same time, because of the feedback (through the gene) in the value of doing mathematics, there is a larger likelihood of finding instances of mathematosis amongst those who do mathematics even though there is no causal link and is no entanglement between the two[4]. The argument is that there is a larger likelihood of finding mathematosis among people who do mathematics because they tend to enjoy doing mathematics, and thus, associate a larger expected value with it.

For mathematosis it's difficult to predict the behaviour in the absence of a model of the disease and its interaction with the gene and doing mathematics. However, an intuitive model for the disease predicts that doing mathematics is the optimal choice in that situation. Entanglement is the key.

References

[1] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, 2000.

[2] Wojciech H. Zurek. "Decoherence, einselection, and the quantum origins of the classical". In: *arXiv:quant-ph/0105127v3* (2003).

[3] Wojciech H. Zurek. "Quantum Darwinism". In: *arXiv:0903.5082v1* (2009).

[4] Agustín Rayo. "MITx: 24.118x Paradox and Infinity, Newcomb's Problem". In: (2019). URL: `https://courses.edx.org/courses/course-v1:MITx+24.118x+2T2019/course/`.

[4] The case of rain causing wet sidewalks[4] is a clearer example of the same. In this case, the occurrence of rain clearly has no entanglement with carrying umbrellas. A similar analysis lead to the same conclusion as the mathematosis case