A reflection on Free Will and Determinism

Marc Rodà Llordés

August 30, 2022

Abstract

Here I discuss some ideas concerning the relationship between the free will of human individuals and the determinism imposed by natural laws. We take Newcomb's paradox as a starting point, a paradoxical game based on predicting the behaviour of human beings. We then review what does it tell us about whether we actually have free will or whether we are fated to act in the way we act.

The discussion is based on topics found in chapters 3, 4, and 5 from the book *On the Brink of Paradox* [Rayo(2019)], such as the one-person Demon's game, casual loop time travel, and the Newcomb's paradox itself.

1 Introduction: Newcomb's Paradox

1.1 The Game

Consider the following game. You enter a room with two closed boxes, a large one and a small one. You are then given two options:

1. Take both boxes.
2. Take the large box only and leave the small box behind.

Now, you are told that the small box contains 1.000 €, while the large box can either be empty or contain 1.000.000 €. The boxes are already set when you enter the room, so your choice will not alter the contexts of the boxes.

At this point, the game might sound dull, but here comes the interesting twist. The contents of the large box are decided by a perfect predictor agent in the following way. If she predicts that you will take both boxes, the large box will be empty, whereas if she predicts that you will only take the large box, she will put the million euros inside.

1.2 The Paradox

Intuitively speaking the paradox arises as follows. On the one hand, when we enter the room, we cannot affect the contents of the boxes anymore, so clearly we must take everything. On the other hand, the agent will predict our move, so if we want the box to contain the million euros, we should take only the large box, and leave those 1000 € behind. So two seemingly valid rational ways of reasoning lead to opposite strategies.

In [Rayo(2019)] this paradox is then discussed in more detail as a conflict between two different principles in *decision theory*:

- **Expected Value Maximization**: Choose an option whose expected value is at least as high as that of any rival option.
- **Dominance**: If option A is always better than option B no matter what happens with things you don't control, you should choose A over B.

These principles are in apparent contradiction in this game, as we just discussed.

1.3 Do we need a perfect predictor?

As discussed in [Rayo(2019)], the existence in practice of such a predictor does not matter, it only matters that it is logically possible. Furthermore, it is not needed that the predictor is 100% accurate for the paradox to arise. By adjusting the money in the large box, we just need it to be better than 50% as we show next.

Consider a predictor whose probability of a correct prediction is $0.5 + \epsilon$, with $\epsilon > 0$. And take the money in the large box to be either zero or x €, according to the predictor's prediction, as in the original game. For us, the players, we have the following table of possibilities

	Predictor is Correct	Predictor is Wrong
We take both boxes	1000 € (0.5 + ϵ)	1000 + x € (0.5 - ϵ)
We take the large box only	x € (0.5 + ϵ)	0 € (0.5 - ϵ)

where we listed the payoff (as well as the probability) of each situation. The expected value for our choices is then given by

$$\begin{aligned} E(\text{both-boxes}) &= 1000(0.5+\epsilon) + (1000+x)(0.5-\epsilon)\text{€} = 1000 + x(0.5-\epsilon)\text{€} \\ E(\text{large-box}) &= x(0.5+\epsilon)\text{€}. \end{aligned} \tag{1}$$

For a paradox to arise, we need the Expected Value Maximization to tell us to take the large box only. Therefore, we just need

$$E(\text{large-box}) > E(\text{both-boxes}) \implies x > 1000/\epsilon\,\text{€}. \tag{2}$$

This implies that for any predictor better than 50% chance (i.e.$\epsilon > 0$) we can obtain a paradox by making the amount of money in the large box big enough. Note also that this is true no matter how much money there is in the small box, as shown in the previous equation, since 1000 could be any number.

2 Perfect predictor

We just saw how a perfect predictor is not needed for Newcomb's paradox to arise. Nevertheless, I think it is most interesting to think about what would happen precisely in the case where the predictor is 100% accurate. If we forget about practical issues, there are several logically possible ways of obtaining a perfect predictor. Here I use some of them to discuss different aspects.

2.1 Time Travel

The first option is to let the predictor be a time traveler. In this case, she only needs to observe what we do, then go back in time, and adjust the boxes accordingly. We can assume she travels back in time just the moment we open the box so that when we do see the contents of the box, they correspond to the correct prediction.

This creates a casual time loop, but since we will make the same choice every time, the loop is consistent. But, this seems to indicate that we are not free when we make the choice, as our choice is conditioned on our "previous" choice. In this case, "previous" refers to previous in super time, as in the super time discussed in [Rayo(2019)]. It seems like we were only free when we made the "first" choice. But was there ever a first choice?

What's even more, are we sure that we would make the same choice? Determinism would tell us that it is the case since we have the same initial conditions. I would argue though, that in principle, according to quantum mechanics, there is inherent randomness in our world, so even with the same conditions, one could observe different outcomes every time one repeats the experiment. One could think, however, that since we are in a time loop, this randomness is again decided on the "first" iteration of the loop and then stays always the same.

2.2 Simulation

The last point in the previous discussion brings us to a second option for the perfect predictor. It could be that, instead of traveling in time, she owns a supercomputer able to simulate the consciousness of the human being together with all the relevant parts of its environment, be it the room with the boxes, the building, the whole town, or even the whole planet or solar system.

If the simulation is aware of all the laws of nature, it should be able to reproduce humans' behaviour exactly, at least according to determinism. In this case, it might be easier to see how quantum theory implies that the simulation can only output probabilities. But again, as long as we don't get at 50% from the simulation all the time, this will lead to a paradox too.

2.3 Yourself

An interesting option I would like to consider is taking yourself as the predictor. Maybe not as the person putting the money in the boxes, but maybe as the player trying to predict itself, to see what will the perfect predictor predict.

It seems like we should be able to predict with 100% accuracy what will we do. Wouldn't that be the point of free will altogether? You decide what to do, and then you do it. However, here is where the Demon's game (seen in chapter 3 of [Rayo(2019)]) tells us that this might not always be easy to achieve. This game explains how there might be changes in the circumstances that will make a totally rational being change their plan. In that same chapter, he discusses the story of Homer, who orders his men to bind him to the mast of his ship, to make sure that he sticks to his original plan of listening to the siren's song without drowning.

One could argue that changing plans according to new situations is the rational thing to do. However, what is paradoxical about these two examples, is that the change of the situation was known from the beginning. Homer knew he would be tempted by the Siren's song, and the player of the Demon game knows that he will be offered infinitely many coins. To further discuss this point, I would like to introduce a situation similar to Newcomb's paradox, Kavka's toxin puzzle.

2.3.1 Kavka's Toxin Puzzle

For this one, suppose that we have a vial of poison that makes you suffer a lot for a certain amount of time, but has no long-lasting consequences nor threatens your life. We just need that drinking this vial is undesirable for anyone. The game can then be explained in different ways, I will go with this one:

You will be paid a huge amount of money today at 10:00 if and only if, at that moment of time, you are willing to drink the vial of poison one hour later at 11:00.

Note that there is no contract involved. In particular, this means that once you have the money, there is no one forcing you to drink the poison. And why would you? At this point you already won the game, so why would you suffer unnecessarily? But here comes the paradoxical question for this game:

Can you intend to drink the toxin if you also intend to change your mind at a later time?

This situation is basically equivalent to Newcomb's paradox. Can you intend to take only one box, even if you also intend to take both boxes in the end? It seems to me like this is impossible. You might trick the predictor, or whichever opponent you are playing against. However, you do know you are not going to drink the poison, and therefore, so should a perfect predictor, or any omniscient being. This seems to suggest though, that any rational being will never be able to win the toxin puzzle, nor the Newcomb's game for that matter. I think this is similar to how the paradoxical trajectories in a world with wormholes are solved in [Rayo(2019)]. Only trajectories that lead to situations consistent with natural laws are allowed to exist.

3 Retro-causality

It's hard to say what is the answer to all the dilemmas in the previous section. My take is that free will might not be properly described. In particular, let's take a closer look at its relation with the passing of time.

These two games seem to suggest that our decisions in the past ought to bind our actions in the future. Imagine that we have faith in the perfect predictor. Then, in the past before going to the room with the boxes, we know that the predictor will predict our move, and we definitely want the money, so we blindly commit to taking only one box, whatever happens. But later we are in the room, and we see both boxes sealed, and whatever we do now will not change the contents, so what in the world will stop us from getting the extra 1000 €? Well, it should be our firm decision from the past, because in the past we decided to believe in the predictor, and from this belief, taking only one box gives us an advantage.

This might sound like a discussion about power of will. Like a person committing to go to the gym the next morning and then oversleeping, because it seems more desirable once the alarm actually rings. However, I do think there is something deeper hidden here. If you do believe in the predictor, you know that if you take both boxes, one of them will be empty, so you will take only the large one because then it will be full.

The previous sentence though can be read in another way. If we take that sentence as true, because we believe in our omniscient predictor, it could also mean that our action in the present affects events in the past. That is, us taking both boxes is the cause of the second box being empty. This leads us to retro-causality, a situation where an act in the future has a consequence in the past.

Retro-causality is something we consider to be wrong, or illogical, therefore, there must be some flaw in our understanding. At this point, I think we have two clear candidates, free will and determinism. Either we don't act freely, and our actions are fully determined from the beginning and there is no problem with someone predicting those actions. Or hard determinism is not true and therefore a perfect predictor is out of the question.

4 Fatalism

In this section, I briefly explore the philosophical consequences of hard determinism. As already discussed thousands of years ago, if we take for granted that everything in the future is already determined (and at this point, it does not even matter whether this determinism comes from initial conditions and natural law or from some sort of divine fate) then we are left with the *idle argument.* An example of this argument would go like this:

Suppose you are ill. Whether you recover from that illness is already fated. If you are fated to recover, then calling a doctor will not change that. If you are fated not to recover, then calling a doctor will not change that. Therefore, it is futile to call a doctor.

Personally, I find this way of reasoning very unappealing. Because it can be taken to a boring extreme by iterating further, as follows. Whether you call a doctor or not is already fated too, so considering whether to call it or not is futile. Even more, the time you spend considering whether you call a doctor or not is also fated, and the anxiety you feel while thinking about the decision is also fated. According to fatalism, whether you believe in fatalism, or whether someone will one day convince you about it is also fated. Whether any of this or anything at all matters to you should also be fated.

I don't think is possible to rule out Fatalism with what humanity knows now, or maybe is not possible at all. However, since it leads to a sort of dead end quite quickly, I think it is more constructive to develop alternatives.

5 Free Will

I think keeping free will leads to more interesting developments, even if this entails abandoning hard determinism. Attempts have been done on keeping both free will and determinism, for instance by changing our definition or understanding of free will (see Compatibilism [McKenna and Coates(2021)]). Here, I want to suggest that determinism might not be quite correct.

Our current basis to defy determinism is some sort of fundamental randomness. Determinism takes as input deterministic initial conditions and uses laws of nature to exactly predict the future. If one adds randomness to the mixture, our ability to exactly predict the future downgrades from exact predictions to probabilistic predictions. This might not look more appealing from a humanistic point of view than fatalism. Some might even say it's worse, as we are no longer destined to follow an already written fate, but instead, we are just subject to random events outside our control. So we gained access to many different futures, but we still do not have the ability to choose. However, I do think we did a step toward our freedom.

One of the currently used physical theories, namely quantum mechanics, includes fundamental randomness. However, the theory is not perfect, it contains flaws and concepts upon which no agreement has been reached. This is to say, that quantum mechanics is most likely not the answer to everything, but it did give us a new element that threatened determinism. In the same way, we might find new evidence and theories in the future that make us question things that we now consider settled. I think it is not logically impossible to consider a way of understanding the world that includes natural laws, randomness, and free will, together with probably many more elements.

It could well be that humanity, and all other extraordinary things happening on our planet is just a deterministic result coming from some random fluctuations in the original state of the universe, together with some random fluctuations on top. A situation where all our actions, even our considerations about free will and the meaning of this are all part of our huge physical system naturally moving towards equilibrium in a way so complex that we will never even understand. However, precisely because of all this complexity, I think it's way too early to surrender our free will to natural laws, so let us explore the world some more and strive for a more satisfactory theory of everything.

References

[Rayo(2019)] A. Rayo, *On the Brink of Paradox: Highlights from the Intersection of Philosophy and Mathematics* (MIT Press, Cambridge, MA, USA, 2019).

[McKenna and Coates(2021)] M. McKenna and D. J. Coates, in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta (Metaphysics Research Lab, Stanford University, 2021) Fall 2021 ed.