Brit. J. Phil. Sci. **49** (1998), 67–87

Newcomb's Paradox Realized with Backward Causation

Jan Hendrik Schmidt

ABSTRACT

In order to refute the widely held belief that the game known as 'Newcomb's paradox' is physically nonsensical and impossible to imagine (e.g. because it involves backward causation), I tell a story in which the game is realized in a classical, deterministic universe in a physically plausible way. The predictor is a collection of beings which are by many orders of magnitude smaller than the player and which can, with their exquisite measurement techniques, observe the particles in the player's body so accurately that they can predict his choice (in much the same way as we can predict the motion of celestial bodies). I argue that the player, by choosing whether to take only one box or both boxes, *influences* whether or not, in the past, the predictor put a million pounds into the second box. Yet, I establish that no causal paradox can arise in this set-up.

1 Introduction

The puzzle of backward causation. Can effects sometimes precede their causes in our universe? If not, is it at least imaginable, or logically consistent, that they might? Philosophers have given a wide range of answers to this question. Some have argued that backward causation is in principle impossible. Others have assumed a more 'optimistic' attitude. Dummett ([1964], p. 125), for example, tells the following story:

> Suppose we come across a tribe who have the following custom. Every second year the young men of the tribe are sent, as part of their initiation ritual, on a lion hunt: they have to prove their manhood. They travel for two days, hunt lions for two days, and spend two days on the return journey; observers go with them, and report to the chief upon their return whether the young men acquitted themselves with bravery or not. ... While the young men are away from the village, the chief performs ceremonies—dances, let us say—intended to cause the young men to act bravely. We notice that he continues to perform these dances for the whole six days that

> the party is away, that is to say, for two days during which the events that the dancing is supposed to influence have already taken place.

Until the young men return to the village, the tribal leader doesn't know whether or not they have been brave, so for him it still makes sense to keep dancing until they return. The exact physical mechanisms of causation are not of interest to him. What is important is that he has evidence from his past experience that, if he dances, his men will have been brave, and if he doesn't dance, they won't have been brave; and since he feels that he can choose whether or not to dance, he feels that it is also up to him, at least to some extent, whether or not his men will have been brave.

Although such stories involving 'magical' backward causation have been told, there is a general consensus that physically acceptable examples of backward causation are lacking at present: as yet, no one has provided a description of a scenario which

1. is physically plausible, classically or quantum-mechanically;
2. involves some kind of backward causation; and
3. does not involve the existence of tachyons or closed time-like curves (e.g. wormholes), which arguably pose a threat of causal paradox.

A physical realization of Newcomb's paradox. In the present paper, I describe such a scenario. To illustrate it, I use the idea of Newcomb's paradox, which is thought to be paradoxical exactly because it seems to involve backward causation.

Originally, Newcomb's paradox was presented as a paradox in decision theory (Nozick [1969]): it is one of the rare cases of a game in which the so-called 'dominance principle' and the 'utility principle' seem to suggest different playing strategies. Up to now, however, no physically plausible way has been proposed in which this game could actually be realized. This has led philosophers to reject it as 'incredible' (Cargile [1975], p. 238) or to claim that it 'simply cannot occur' (Mackie [1977], p. 233).

To prevent this beautiful paradox from being classified as physical nonsense, I describe a physically plausible way in which it can be realized in a classical universe (Section **2**). I explain how the player's choice can be infallibly predicted and how, in accordance with this prediction, either £1,000,000 or £0 can be written on a cheque before the player makes his choice. The proposed prediction method does *not* use mechanisms involving closed time-like curves or tachyons, by means of which information could be sent into the past. Such mechanisms are beset by serious philosophical problems, like the 'grandfather paradox', to which there are as yet no generally accepted solutions.

In my proposal, there is no threat of causal paradox (Section **3**). Its underlying idea is the following. The predictor is a collection of beings which are by many

orders of magnitude smaller than the player. If the physical laws describing human behaviour are deterministic, and if these beings can, with their exquisite measurement techniques, measure the state of a person and the person's environment with such a high accuracy that they can predict that person's behaviour in the near future (say, for the next 24 hours), in much the same way as we can predict the motion of celestial bodies, then they are in a position to play Newcomb's game with us.

In the Appendix, I will explain the physical foundations of my proposal: it is based on recent results about the predictability of events (Schmidt [1997] and [1998a]), and about the limitations on the accuracy of the observations which we can make (Schmidt [1998b]), in classical universes.

While the physics of the set-up is understandable, what is going on philosophically is far from obvious. In what sense does the scenario involve backward causation—that is, in what sense can the player's choice be said to 'influence' whether £0 or £1,000,000 was written on the cheque by the predictor *before* the player made his choice? I will argue that the intuition that there is, in a specific sense, backward causation, is natural and reasonable. However, I will not deny that there are other senses of 'causation', according to which there is no backward causation in this scenario (Section **3**).

I will point out that authors who have concluded that the player in Newcomb's game should take both boxes have made an assumption which fails in the present set-up. Finally, I will show that the kind of backward causation arising here cannot lead to causal paradoxes.

2 A strange, but possible, story in a classical world

Many have struggled to formulate theories of singular event causation, and the difficulties met are enormous. Virtually any theory which claims validity over a wide range of cases faces serious objections, and I think many would agree that, at present, no generally acceptable, comprehensive account of event causation is in sight.

I therefore will not embark on the enterprise of constructing a general account of causation of my own. Instead, I will restrict myself to the discussion of a particular case, and argue that under the—admittedly rather extraordinary—circumstances described, we would have the *intuition* that we can, by an action in the present, influence an event in the past: i.e. that there is backward causation in that particular case.

The case I consider is simple and physically unmysterious. Everything is deterministic, and, apart from its potential for backward causation, I do not think any of the elements which often cause concern for theories of causation (pre-emption, prevention, omission, multiple causes, distinctions between sufficient and necessary causes, and the like) play an important role here.

Since my thesis is about human intuitions in a particular situation, I am not concerned with 'metaphysical causation'. In this paper, 'causation' is understood as a concept in our language, as a practical tool for understanding the relations between events in our lives—not as a fundamental metaphysical category. Thus, my rhetoric will not (primarily) take on the form of a philosophical argument: for I cannot establish that we *would* have a certain intuition by trying to convince my readers that, for some philosophical reasons, we *should* have this intuition.

Rather, I am concerned with the question, which causal description of a scenario is most appealing to the people who experience the scenario. To answer this question, we need to *empathize* with these people; and the best way to achieve this is by following the thoughts and emotions they have as a result of their experiences. This is why I have chosen the format of a short story for this section. I will describe the situation in the way it presents itself to the individuals concerned: this will enable us to 'observe' how these individuals form, and maybe change, their beliefs and intuitions about causation.

The subsequent discussion in a more usual philosophical style (Section **3**) aims to separate the different aspects of our intuitions. It reflects some of the thoughts the player may have when he tries to make up his mind. Ultimately, however, it is of course left to everyone's own judgement whether or not their intuition agrees with the one I would favour. Yet, it should be said that if you, in the player's skin, would decide to leave the first box on the table *in order that* a million will have been put into the second box, this strongly indicates, in a very simple, intuitive, and practical sense, that you would *believe*, under the circumstances described, that you can have an effect on the past.

So the story is this.

Newcomb's game. It is your birthday. You get up in the morning, and to your great surprise, you find two open[1] boxes on your kitchen table: in the first box, there is a £1,000 cheque, and the second box is apparently empty. Next to the two boxes, there is a note. It says:

> Hello!
>
> We are pleased to tell you that we are about to play the Newcomb game with you. We've made great efforts to prepare it—please read on.
>
> The first box, there, in front of you, has a cheque for £1,000 in it. We'll tell you in a moment what you can do with it. The second box, you may think, is empty—but not so! In the second box, there is a tiny cheque: it's so tiny that you can't see it, but it is there now. It's a tiny cheque with either £0 or £1,000,000 written on it—which of the two, we won't tell you.

[1] In Nozick's [1969] formulation, the second box is closed. Here, both boxes are open, because I want to make clear that there are physical reasons why cheating, by looking to see which cheque is in the second box, is impossible.

> Whatever is in the second box, whether £0 or £1,000,000, is yours. We recommend that you close the box carefully, so the tiny cheque won't fly away as you carry the box. You can take it to the bank, whenever you want. The bankers will exchange the tiny cheque for the corresponding ordinary cheque for you, which you can then cash: you'll get either £1,000,000 or nothing.
>
> You have the choice of either taking the first box with the cheque for £1,000, or leaving it there. If you decide to take it, we ask you to take it within a minute, and it will be yours. If you decide not to take it, you need to do no more: leave the cheque there in the box, and we will set fire to it.
>
> You have exactly one minute to make up your mind.
>
> Just keep the following in mind: we are infallible predictors. When you read this note, we will already have predicted whether or not you will decide to take the £1,000 cheque. If we predict that you will leave the £1,000 cheque on the table, we write the sum of £1,000,000 on to the second cheque. If, on the other hand, we predict that you will take it, then we write £0 on to the cheque in the second box.
>
> Look at your watch: the countdown starts now.

What will you do?

Is someone making fun of you? Initially, you may be perplexed. Apparently someone wants you to think that he can predict your behaviour. Someone wants you to believe that the sum of £1,000,000 will be written on the second cheque if and only if you leave the cheque with £1,000 on the table. If this someone is right, then you may be better off leaving the £1,000 on the table, because, one may argue, only then is it sure that you will receive the million.

But who could this someone be? You just cannot imagine. As far as you know, human behaviour cannot be predicted. And even if you believe that God or some Laplacian demon-like creature knows your behaviour in advance, you would probably doubt that such a creature is playing Newcomb's game with you this very minute.

You will quickly convince yourself that the most plausible explanation you can find for this morning's surprise is that your younger sister is playing a trick on you for your birthday. She probably thinks that, as a philosopher, you would like this little joke. Greedy as you are, you take everything that there is to be taken: you take the £1,000 from the first box, and in order not to disappoint your little sister, you take the second box all the way to the bank, where, of course, they give you nothing in return.

However, when you meet your sister in the evening, wanting to thank her for the surprise and the £1,000, she denies any connection with the game.

News, discoveries, experiences. During the next couple of months, the newspapers are filled with reports from bankers who have received letters from unknown authors calling themselves 'the dwarfs' and offering the most

extraordinary services in exchange for the establishment of so-called 'detection rooms' in the banks: they ask the bankers, whenever a customer arrives with an apparently empty box and claims that it contains a tiny cheque, to put the box into this 'detection room' and to simply wait for five minutes. During those five minutes, so 'the dwarfs' write, they would replace any tiny cheque found in the box by an ordinary cheque of the same value as the tiny cheque. In spite of the strangeness of this offer and in spite of the fact that no banker ever saw any such 'dwarf' in their 'detection room', more and more bankers accepted the deal and built the detection rooms, and the dwarfs' letters were apparently right: the amazing services promised by the dwarfs were delivered, and no customer ever experienced any difficulties in cashing these human-size £1,000,000 cheques.

Furthermore, thousands of people report that they were confronted with this 'Newcomb game' on their birthdays. Indeed, those who chose to take the cheque for £1,000 were told at the bank that there was a £0 cheque in the other box; whereas all those who left the £1,000 cheque in the box observed it burn to ashes, and then received £1,000,000 from their bank.

In many laboratories, physicists keep finding letters also signed 'the dwarfs', suggesting new experiments and describing the results to be expected. Those physicists who take up the suggestions make revolutionary progress: they keep discovering new particles, new methods for improving the precision of their measurements ... . It is as though they had been given a physics textbook from the end of the next millennium.

Thus it becomes clear that the human species has started communicating with a population of a new kind of creature. They are so tiny that it is impossible to see them even with the best microscopes available, and their knowledge of science is centuries, if not millennia, ahead of ours. For some reason they have decided to take up contact with us and to teach us some new physics.

After six months of rapid scientific progress, all leading newspapers publish the following message from the dwarfs:

> Dear human species,
>
> You have now discovered all the particles we call 'gigantions': they are the particles your bodies consist of, as in fact does most matter on Earth. If you want to discover physics on an even smaller scale, you will have to refine your measurement accuracy by many more orders of magnitude, because the next new particle you can discover is already what we call a 'tinion': one of those particles our own tiny bodies consist of.
>
> But be warned: starting from your current technology, it will take you a long time to reach the measurement accuracy required to detect a tinion, even with our help.
>
> The dwarfs.

The technicalities of the game. The day before your next birthday, the question what this Newcomb game is all about becomes more pressing for you. You leave a letter on your kitchen table, saying:

> Dear dwarfs,
>
> Would you be so kind as to explain to me the technicalities of your Newcomb game? In particular, I would like to know whether you will really put the cheque into the second box before I make my choice.
>
> How do you manage to interact with us human beings, if you are so small that we can't even see you?
>
> And, by the way, would you recommend that I take the £1,000 cheque, or would I do better to leave it on the table? I'm tempted to imagine an omniscient observer,[2] who would see what's written on the tiny cheque and who then, knowing what there is to be taken, would of course advise me to take both cheques—can you give me any reason why I shouldn't follow his imagined advice?
>
> Yours, a philosopher.

And indeed, on the morning of your birthday you find, in addition to the expected two boxes and the note explaining the game, a reply to your letter. It reads:

> Dear philosopher,
>
> From our perspective, the gigantions, that is, the particles you human beings consist of, are huge. As we are so tiny compared to any gigantion, we can measure their positions and velocities very accurately, and by now, we have pretty well figured out the physical laws that describe their motion. We have a central computer which, for decades, has been recording the motion of all gigantions on earth. On the basis of this data and by application of these physical laws, we are able to predict the motion of the gigantions with sufficient accuracy so that we can predict all the nuclear and chemical reactions happening in your atoms and molecules for the next 24 hours. (After more than 24 hours, though, the resolution of our predictions becomes so bad that we can no longer discern events on a nuclear scale.)
>
> Since human bodies consist of atoms, this means that we can also predict human behaviour up to 24 hours in advance. So when we play Newcomb's game (see Figure 1), we already know 24 hours minus one minute before you finish reading the note, whether or not you will leave the £1,000 cheque on the table. Accordingly, we write the sum of either £0 or £1,000,000 on to a dwarf-size cheque, which we place into the second box—*24 hours minus one minute before you finish reading the note.*
>
> So there's the answer to your first question: the tiny cheque *is* there before you make your choice. However, this cheque consists solely of tinions; and therefore, even if you bring the box to a laboratory and ask the physicists what's in it, they will not be able to find the tiny cheque in the box, because they can only detect gigantions—of which there aren't any

[2] This argument has been put forward e.g. by Schlesinger ([1974], p. 211).

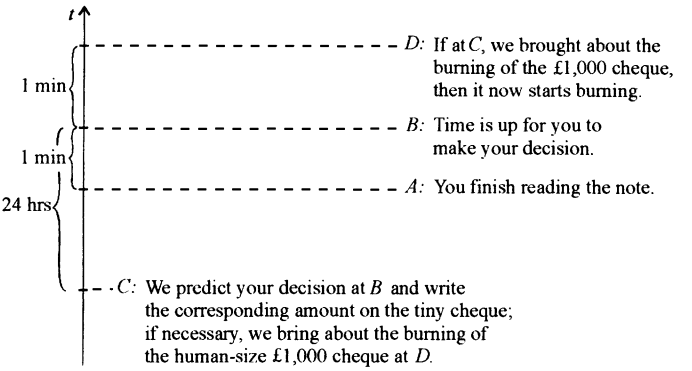

Fig. 1 What happens when

in the cheque. Thus, they will not be able to tell you whether there is a cheque for £0 or £1,000,000 in your box.

However, what we can do on our tiny scale—predicting human behaviour and placing tiny cheques—is only half of the story. We can also produce effects on the human scale: we can put proper big notes (like this one) and boxes with cheques on your kitchen tables, and write letters to your scientists, on A4-size paper. But you should appreciate that fiddling with gigantic things like atoms costs us a lot of effort. Can you imagine what it takes for us tiny dwarfs to deflect even slightly the trajectory of one single gigantion?

But we enjoy this game. Pushing and pulling your gigantions just by tiny amounts, we can bring about events that you can eventually perceive. However, since we can predict all events on a nuclear scale for the next 24 hours, it is clear that our pushing and pulling cannot produce any significant effects on a nuclear scale—nor, *a fortiori*, on any larger scale—in less than 24 hours: for otherwise, we could render our own predictions false. But given 24 hours *plus one minute*, we *can* produce effects on a nuclear scale, and soon afterwards also on even larger scales.

We can cause the most amazing physical processes: for example, produce a fresh paper cheque out of air, by bringing about little nuclear reactions transforming oxygen and nitrogen from the air into carbon and hydrogen in the paper of the cheque. Such events are very *unlikely* under ordinary circumstances, that is, without our subtle pushing and pulling. But there is always a tiny chance that even such extraordinary nuclear (or similarly, chemical) reactions will happen: all the electrons and quarks are there in the air, and all we need to do is to push and pull the gigantions a little, to turn these tiny probabilities into certainties and bring it about that 24 hours plus one minute later, the electrons and quarks will move into a configuration where they no longer form oxygen in air, but carbon in paper instead.

That is how we create cheques, letters, and boxes out of air and how we set fire to cheques, if we want to. It requires a lot of careful measuring and calculating of the gigantion trajectories, and we need to invest a lot of our tiny energy to deflect the trajectories of billions and billions of gigantions in

exactly the way we need to, in order to make the desired object (e.g. an ordinary £1,000 cheque) emerge 24 hours plus one minute later. But we can do it.

As we said, what we decide to do now will influence the events on a nuclear scale 24 hours plus one minute later. However, from our perspective, the events we can still influence now will become inevitable a minute from now: we will then be able to predict them, and nothing we can then still do will be powerful enough to prevent these predictions from coming true. We know the extent to which we are able to deflect the gigantions' trajectories, and this extent is not sufficient to produce a significant effect on a nuclear, or a larger, scale in 24 hours or less.

So now you know how we interact with you human beings.

You also asked for our advice as to whether you should take the £1,000 cheque or leave it. Well, the easiest way of giving you advice would of course be for us to tell you whether we wrote £0 or £1,000,000 on the cheque in the second box: if we could tell you in time, you would know what's there to be taken, and of course you would take it all. The problem is that, even if we wanted to tell you, we would not be able to tell you *in time*. For, although *we* know the answer from the moment we predict your decision, that is, 24 hours minus one minute before you finish reading the note, it will take us at least 24 hours plus one minute to let *you* know the answer—and by that time, it is too late for you, since you only had one minute to make your decision.

In short, we can't let you know in time what's written on the cheque. But we can still give you advice: if you ask us before we predict your choice, we would recommend that you leave the £1,000 cheque on the table: for if that is what you will do, we will predict this and accordingly, we will write £1,000,000 on the tiny cheque. But once we have predicted your choice and written the corresponding amount on the second cheque, it is too late: your decision is already inevitable from our perspective, and it would be futile for us still to be trying to give you any advice.

Therefore, at this very moment, when we still do not know what you will decide and when you still have a chance of following our advice, we recommend that you leave the £1,000 on the table if you want to receive the million.

You also wrote that you are tempted to imagine an omniscient observer,

1. who would see whether £0 or £1,000,000 is written on the cheque and
2. whom you would imagine to tell you in time to take everything there is to be taken, i.e. to take the £1,000 cheque as well.

If you honestly think you can fare better by following the imagined advice of a fictitious omniscient observer, then do—it's your life, you can do whatever you want. All we can do is tell you that, so far as we know, such an observer doesn't actually exist. We do not know of anyone in this universe who can help you any better than we can.

We hope that this answers your questions.

Yours, the dwarfs.

Reflections. Now you are really perplexed. Would you have believed such a story a year ago? If you had found such a letter last year, it might have made an even more amusing birthday surprise, but you would still have thought it a joke. But now, after all these news reports, after all these letters signed 'the dwarfs', containing the most extraordinary information enabling this incredible progress of science, now that anyone seriously questioning the existence of the dwarfs would be recommended to visit a psychiatrist—who would doubt that the situation is as the dwarfs described it in their letter? After all the extraordinary developments in the last year, you are now convinced that there is a £1,000,000 cheque in the second box if and only if you decide to leave the £1,000 cheque on the table.

You read the other note. It's the same text as last year: '... the countdown starts now.' You stare at the second box. As far as you can see, it is empty, but now you know that there is a tiny cheque right there in front of you, with either £0 or £1,000,000 written on it.

'Suppose there is £1,000,000 written on it,' you think. 'Will it change to £0 if I decide to take the £1,000 cheque now?' You know that it won't. The dwarfs explained to you how everything works and that the cheque doesn't change once it is in place. 'Or am I, in that case, forced to leave the £1,000 cheque there? Has my decision already been made for me?' That sounds absurd. You feel that you are as free as ever to decide whatever you want.

You stare and stare in amazement. You lose track of time ... and suddenly the £1,000 cheque bursts into flames. You realize that your time is up. Relieved, you pick up the second box and take it to the bank, where you receive a million pounds in exchange.

Still, you remain puzzled. You understand what's going on physically, yet it seems strange to you that you are apparently able to make a choice now, in order that the dwarfs would have written £1,000,000 on your cheque yesterday. How can you make sense of this?

3 Does the player influence what was put into the second box?

People's initial reactions when they are told of Newcomb's paradox vary greatly. Nozick [1969] writes:

> I have put this problem to a large number of people, both friends and students in class. 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 opposite half is just being silly.

To me, this suggests that it is not a clear-cut matter whether or not Newcomb's paradox involves backward causation: Some people would emphasize that, since the Newtonian universe I described is deterministic, the player's action in

the present scenario is the result of the previous configurations of the particles in the universe, just as much as are the predictions of the dwarfs placing the tiny cheques. Therefore, they would say that the player's action and the placing of the cheques are effects of a common cause, and thus the player's choice cannot influence the sum written on the tiny cheque.

Others would point out that recognizing these facts does not help the player in his decision whether or not to take the £1,000. They would consider it more relevant that there is a strict correlation between the player's choice and the sum written on the cheque: whenever the player takes the £1,000, £0 was written on the cheque, and whenever he leaves it, the second box contains a million. Since the player can decide whether or not to take the £1,000, he can, in virtue of this strict correlation, also bring about the corresponding event in the past.

These two lines of thought represent differing intuitions about whether or not this scenario involves backward causation. Although they agree, *which* aspects they think speak in favour of, and which they think speak against backward causation in this case, they differ in the *extents* to which they consider the different aspects *relevant* and *important*.

Rather than abstractly advocating a certain line of thought and attempting to demonstrate that the other is mistaken, I propose—in accordance with my strategy of Section **1**—to adopt a more subjectivist attitude: First, I will briefly mention different connotations of causation, in order then to compare two causal accounts of the scenario. I will conclude that (i) whether or not there is considered to be backward causation is a matter of personal judgement which depends on the judging person's particular interests and purposes, and (ii) that for our ordinary human purposes, there *is* backward causation.

Finally, I will answer the objection that backward causation is in principle impossible because it may imply causal paradoxes, and explain where authors who have concluded that the player in Newcomb's game should take both boxes have made an assumption which fails in the scenario as I described it.

Connotations of causation. Let me first set aside some common connotations of causation which do not give a clear indication about whether or not there is backward causation in my realization of Newcomb's paradox.

- *Counterfactual dependence.* Actually, the player left the £1,000 on the table, and he received the million. A counterfactual analysis of causation (e.g. Lewis [1973]) may suggest the following statement as a criterion for backward causation in the present case: 'If the player had taken the £1,000, the million wouldn't have been placed into the second box.'

  People's intuitions about the truth value of this counterfactual differ:

Lewis would presumably deny it,[3] while Dummett might think it true. Because of its intuitive ambiguity, I will not discuss this counterfactual, nor its relation to causation here.

- *Entailment.* Causation has also been analysed in terms of entailment (e.g. Mackie's [1965] INUS conditions). Like the counterfactual analysis, these analyses don't provide a unique and generally accepted answer to the question whether or not there is backward causation in this case, and I will not discuss them here.
- *Explanation.* Causation often serves as a tool for explaining relations between events. In the present case, however, it is unclear what the 'best' or 'correct' causal explanation of the scenario is. Either we say, without backward causation, that both the prediction and the player's decision were caused by a configuration of particles shortly before the prediction; or we say that the player's decision is the beginning of a causal chain, partly going backwards in time, which leads to the placing of the tiny cheque (and further, of course).

But not all connotations of causation speak ambiguously under the present circumstances. Some connotations indicate that there is no backward causation:

- *There is no 'traditionally recognized' causal mechanism.* Most, if not all, known causal influences are propagated along a spatiotemporally continuous causal chain with some kind of matter or energy. In the present scenario, however, neither matter nor energy is being propagated into the past, and there is no causal propagation along a continuous path through space-time, leading from cause to effect. Therefore, one may argue that it would be extraordinary if there were backward causation.

  However, there is no consensus that causation must involve matter or energy transfer and continuous propagation, and it is not unusual for philosophers to consider scenarios, involving backward causation, where indeed this is not the case (cf. Dummett's [1964] example of the dancing tribal leader). Yet, this criterion for our intuitions persists.
- *There is a common cause (of the player's choice and the amount in the second box) in the past.* The existence of a past common cause X of both an event E and of an alleged cause C of E in E's future is a widely used criterion for ruling out that C is a cause of E: in such a case, rather than saying that C causes E, it would be said that X causes both C and E.

  In deterministic universes, however, this is an unreasonable criterion, since *any* event F is deterministically caused by the state $X(t)$ of the universe

[3] At least, Lewis's counterfactual analysis of causation [1973] avoids backtracking counterfactuals such as the above.

at any time t before F's occurrence, and so, for *any* two events C and E, this criterion could be used to show that C does not cause E (by appealing to a common cause $X(t)$ occurring before both C and E)—regardless of C and E's actual causal relationship.

Other connotations of causation speak in favour of backward causation in the present scenario, i.e. they indicate that the player's choice indeed *influences* the contents of the second box:

- *The player's choice is a means to bring about a desired end in the past (the placing of the million).* Under ordinary circumstances, if A is a means in order to achieve B, then A is among the causes of B—at least, no cases are known where this is false. In the present case, if the player decides to leave the £1,000 on the table (A), he does so *in order that* the dwarfs will have predicted this and thus written a million pounds on the tiny cheque (B). This suggests that A is a cause of B, backwards in time.
- *The player is held responsible for the consequences of his choice, even though some of them lie in the past.* Under all known circumstances, a person is held responsible (and receives blame or praise) for an event B whose occurrence is correlated with his action A if and only if his action A was a cause of B and the person knew (or at least, could and should have known) that A would cause B.[4]

  To see how this criterion enters the present discussion, imagine a slightly altered 'game': Instead of being nice dwarfs, placing a tiny cheque into the second box, the predictors are horrible tiny robots which are programmed to kill a dwarf if and only if they predict that the player will take the box with the £1,000. Moreover, the player is informed of this unpleasant alteration of the game. If it is well known that these robots are perfectly reliable predictors, and a greedy player decides to take the £1,000, and 'in consequence' a dwarf was killed, then this greedy player will be held responsible for the dwarf's death and probably face life-long imprisonment. In court, it will be judged: 'By taking the £1,000, the player knowingly *caused* the killing of the dwarf, so he is responsible for it—although the effect of his action lies in its past in this peculiar case.' (Rest assured: the programmer of the killing robots will not go free, either.)

Backward causation in court. This alteration adds a dramatic aspect to Newcomb's game. It is not now merely the amount of increase in the player's wealth which is at stake—it is the question, whether or not the player committed a murder: did the greedy player 'cause' the killing of an innocent dwarf in the past?

[4] I take this claim to be plausible enough for the present purposes. Although there are various ways of justifying it, I do not consider it necessary to do so here.

We have seen that there are arguments both for and against backward causation in this scenario, so we might say that we are unable to give an objective answer. But what if we are urged to make up our minds? What if we were involved in a court case about the player's (alleged?) crime? It would seem, intuitively, that the player should be punished if and only if he caused the killing of the dwarf.

In such a court case, the player's lawyer would probably argue as follows.

> The player is not at fault. Rather, it is his genes, his upbringing, the collection of all his experiences, and the exact circumstances he encountered during his entire life, which deterministically led him to take the £1,000, although he of course knew—and regretted—that this involved a dwarf's having been killed. The player's actions are mere symptoms of his and the world's predispositions, and it would be unjust to punish him for showing symptoms of a condition which lies entirely out of his control.

The prosecutor's response could be this:

> Had the player known of particular circumstances which already determined him to take the £1,000 before he made his choice, he should be freed of his charge. For example, if the player had seen someone grab his hand and violently force him to take the £1,000, and if the player had struggled to prevent this in order to save the poor dwarf's life, then he certainly should not have been accused of murder.
>
> But as it was, the player did not, and indeed, *could not* see anything that would force him to take the £1,000. He felt that it was entirely up to him what he would do, and he contemplated both possibilities: either leaving the money and thus ensuring that the dwarf lived, or taking the money and thus ensuring she was killed. Yet he decided for the latter. Such behaviour should be punished, even if it has always been determined. Determinism cannot relieve people of their responsibilities.

As a result of this devastating indictment, the judgement would sound:

> The player knew that, by taking the £1,000, he would cause the killing of the dwarf in the past. Yet, he was so greedy that he took them. He will be sentenced to life-long imprisonment for having killed a dwarf. Let this be a lesson to all those who will in the future be facing choices similar to this player's.

Two legitimate causal descriptions. So much for this court case and its pronounced judgement. But did the player *really* 'cause' the killing of the dwarf? One might still maintain that the player didn't cause the killing, and argue that what mattered in the court case was that the player is *responsible* for the killing, and not whether he 'caused' the killing.

Indeed, if you are a dwarf technician whose job it is to program dwarf computers and robots to predict human behaviour 24 hours in advance, and who is not concerned with the player's moral dilemmas, you may well think this way and thus favour the

> *Mechanically oriented description of the scenario:* The particular configuration (and motion) C of the particles in the player's body, in the boxes, and in the environment shortly before the moment when the prediction was made *caused* the robot to predict that the player would take the £1,000. (In consequence, it killed a dwarf.) Also, the configuration C deterministically led to, i.e. *caused*, the player's taking the £1,000 a day later.

The dwarf technician has a specification of C at his disposal, which is so detailed that he can understand *in principle* exactly *how* it will lead to the player's taking the £1,000. This is why he is satisfied by this mechanically oriented description—although it would probably take him a lifetime to grasp all the physical happenings in the player's body and its surroundings in sufficient detail for him to *actually* understand the mechanical explanation.

However, unless the configuration C is specified in such great detail, the mechanically oriented description won't explain the course of events very well: merely to say that C is 'a configuration leading to the taking of £1,000 a day later' gives us no insight into *how* the player's decision came about. Besides, we human beings don't normally share much interest in such detailed descriptions, and, as explained in Section **2**, it is indeed impossible for us to gain knowledge of the details of such a description before the player makes his choice.

For our everyday human purposes, a much simpler, and—I would contend—equally explanatory, account of the scenario is at hand:

> *The anthropically oriented description:* After thinking about the tiny innocent dwarf for a while, which would be killed if and only if the player were to take the £1,000, the greedy player nevertheless decided to take the £1,000, because he wanted the money to buy a new car. *As a result*, this is what the tiny robot predicted (by some technical method which we understand in principle, but in which we have no particular interest), and so it killed the innocent dwarf.

To the player, the judges, and indeed to most of mankind, this description is more appealing, since it relates concepts and events which are of importance to us in our human understanding of our world, rather than diving into the technical details of which particle configurations lead, by which trajectories, to which other particle configurations (which we are anyway unable to follow in practice). It is much more satisfying to say that 'I greedily took the £1,000; *thus*, this is what was predicted by the horrible robot, which in consequence killed the dwarf', rather than 'I took the £1,000 because the particles in my body were moving in such a way that I would. And, by the way, this was also predicted and so a dwarf was killed, unfortunately.' The latter answer would merely prompt the further question: 'Yes yes, but what were you thinking? What made you decide to take the £1,000?' And in the end, we humans would

hardly be able to avoid giving an anthropically oriented explanation like the one above.

I believe that, in the long run, the language adopted by humans who are not primarily concerned with the technicalities of the dwarfs' predictions will favour the anthropically oriented description. Eventually, the vast majority of people will say *and believe* that the player was recklessly greedy and murdered an innocent dwarf just for £1,000. Experienced and well-informed people will believe that there is backward causation in this scenario; and, being used to this kind of situation, they will find nothing extraordinary about it. Someone who still insists that their belief is false would have to maintain that everyone else except him is wrong, although he has no more information about the physical state of affairs than the others have. Probably such a person would be considered a bit strange.

What implications does all this have for the player in Newcomb's game: should he, or should he not take the £1,000? I, of course, maintain that he shouldn't. But others disagree.

Swain ([1988], pp. 395f.), for example, writes that 'The only way that the [player] could increase [his] chances of becoming a millionaire is by influencing the prediction that has already been made.' Indeed, as I have shown, the player can influence the dwarfs' prediction, by choosing to leave the £1,000 on the table. But Swain does not consider this possibility: right at the beginning of his argument, he adds a veto on backward causation.

Mackie ([1977], p. 217) also acknowledges that 'what is important in the example is surely not probabilistic but causal dependence' and that 'the choice [of leaving the £1,000] could ... be defended with the help of the ... extravagant assumption that there occurs the extreme form of backward causation, the bringing about of the past' (p. 218). But he doesn't produce any argument why there shouldn't be backward causation in Newcomb's extraordinary game. He merely states: 'What the player does cannot affect what is in the [second] box' (p. 217).

Indeed, if one assumes that the predictor is human, or at least that he is some being with the power to place a *human-size* cheque into the second box *within a minute of predicting the player's choice*, then the conclusion that one should take the £1,000 or that the game is altogether impossible to imagine, may well follow. But this assumption fails in the scenario as I described it.

Is this kind of backward causation paradoxical? I hope to have convinced the reader not only that the anthropically oriented causal description, involving backward causation, represents one of many natural and reasonable intuitions for a human being to have under the extraordinary circumstances described; but also that the scenario actually involves backward causation, because in the eyes of people who regularly experience such scenarios and who are well informed

about what is going on, the anthropically oriented causal description is *the* most appropriate causal description of the scenario.

Some philosophers, however, have put forward arguments aiming to establish that backward causation is in principle impossible. If their arguments are valid, there can be no backward causation in the present scenario *in any sense*—including the anthropically oriented sense I suggested—which would undermine my conclusions, in whatever way I attempt to support them.

Of course, I do not have space here to answer all objections to backward causation. Therefore, I will restrict myself to the most common one: the threat of causal loops producing a causal paradox. (The discussion of objections of a more metaphysical character, e.g. based on the idea that the past is fixed, or that causation fixes the direction of time, is more complex: I will take it up elsewhere.)

> *The threat of 'grandfather paradoxes'.* To put it simply: if I can influence the past, why, then, can't I bring it about that my grandfather was killed before my parents were conceived, thus preventing my own existence—i.e. producing a causal loop with a logical contradiction?

It is not difficult to reply to this objection in the context of the present scenario, and in fact I have already done so in Section **2**, when the dwarfs explained why they can't let you know what's in the second box *before* you have to make your choice.

More formally, the reason why it is impossible to obtain any causal loops with the present mechanisms of causation is the following. In a universe like the one described, we know of three kinds of anthropic causal influence (of which (i) and (iii) are also 'mechanical' causal influences, as well as bring 'anthropic'):

(i) 'Orthodox' causal influences: these stay within human scale or within dwarf scale and go strictly forward in time.

(ii) Anthropic causal influences such as the player's choice causing the killing of the innocent dwarf: these influences go from human-scale events to dwarf-scale events, via dwarfs' predictions of human scale events, and they reach *at most* 24 hours into the past.

(iii) Anthropic causal influences such as the dwarfs' initiation of the burning of the human-size cheque, causing its burning: these influences go from dwarf-scale events to human-scale events, via the dwarfs' pushing and pulling of gigantions, bringing about reactions on a nuclear scale which would not have happened without their pushing and pulling; they go *at least* 24 hours plus one minute into the future.

Since none of these causal influences can cross the dotted lines (Figure 2) in the backward direction, it is impossible to set up any causal loops with these three kinds of anthropic causal influence. Of course, this does not rule out the

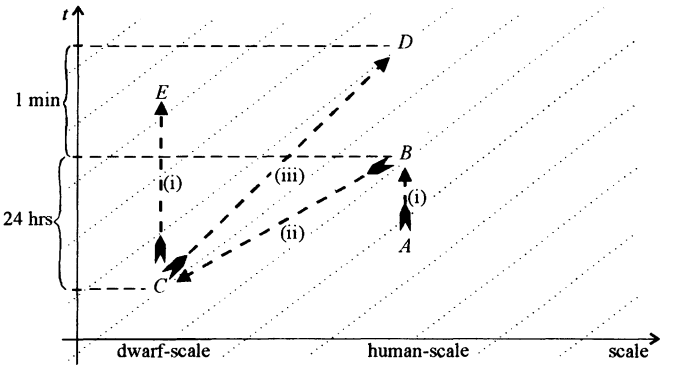

Fig. 2 The three known kinds of anthropic causation cannot give rise to causal loops.

possibility of other (new) kinds of anthropic causal influence, of which we are not yet aware and which, in new ways, could give rise to causal paradoxes. But at least I have shown that the new causal mechanisms I described don't pose a new threat of causal paradox. If we wish to rule out the possibility of causal loops by *any* means of causation, we are led back to the old question, whether there are mechanisms of backward causation which are as yet undiscovered—which of course I cannot answer.

4 Conclusion

I have shown that Newcomb's game is not, as many have claimed, physically nonsensical or impossible to imagine: it is not ruled out that one morning, you might actually be confronted with it.

Admittedly, however, this is not very likely to happen. The prevailing theories of quantum phenomena are usually interpreted as indeterministic, and so it would seem that our actual universe cannot accommodate scenarios like the one described. Yet—as is demonstrated by the recent recognition received by the de Broglie–Bohm interpretation of quantum mechanics (Holland [1993])—physical determinism remains unrefuted, and the story I have told is by no means excluded as a physical possibility.

I have also argued that the player should leave the £1,000 on the table, because, in an anthropically oriented sense, his choice *influences* what the predictor put into the second box: i.e. that backward causation in this sense is possible. I have shown that, nevertheless, no causal loops can arise. Furthermore, this 'anthropic' sense of 'causation' is not just one of many equally relevant senses of the word 'causation': I have argued that it will in fact be the central one for human beings living in a universe like the one I described.

Acknowledgements

I am grateful to my supervisor Jeremy Butterfield for commenting on drafts of this paper, as well as to seminar audiences in Pittsburgh, Budapest, and Cambridge, and in particular to John Earman, John Norton, Miklós Rédei, and László Szabó, for helpful discussions.

Trinity College
Cambridge CB2 1TQ
UK

References

Cargile, J. [1975]: 'Newcomb's Paradox', *British Journal for the Philosophy of Science*, **26**, pp. 234–39.

Dummett, M. [1964]: 'Bringing about the Past', in R. Le Poidevin and M. MacBeath (*eds*) [1993]: *The Philosophy of Time*, Oxford, Oxford University Press.

Holland, P. R. [1993]: *The Quantum Theory of Motion: an Account of the de Broglie–Bohm Causal Interpretation of Quantum Mechanics*, Cambridge, Cambridge University Press.

Lewis, D. [1973]: 'Causation', in E. Sosa and M. Tooley (*eds*) [1993]: *Causation*, Oxford, Oxford University Press.

Lewis, D. [1973]: *Counterfactuals*, Oxford, Blackwell.

Mackie, J. L. [1965] 'Causes and Conditions', in E. Sosa and M. Tooley (*eds*) [1993]: *Causation*, Oxford, Oxford University Press.

Mackie, J. L. [1977]: 'Newcomb's Paradox and the Direction of Causation', *Canadian Journal of Philosophy*, **7**, 2, pp. 213–225.

Mellor, D. H. [1981]: *Real Time*, Cambridge, Cambridge University Press.

Nozick, R. [1969]: 'Newcomb's Problem and Two Principles of Choice', in N. Rescher (*ed.*): *Essays in Honor of Carl G. Hempel*, Dordrecht, Reidel Publishing Company, pp. 114–6.

Schlesinger, G. [1974]: 'The Unpredictability of Free Choices,' *British Journal for the Philosophy of Science*, **25**, pp. 209–21.

Schmidt, J. H. [1997]: 'Classical Universes are Perfectly Predictable', forthcoming in *Studies in the History and Philosophy of Modern Physics.*

Schmidt, J. H. [1998a]: 'Predicting the Trajectories of Particles in Newtonian Mechanics and Special Relativity', forthcoming in *Studies in the History and Philosophy of Modern Physics.*

Schmidt, J. H. [1998b]: 'What We Cannot Know about the Past: Uncertainty Principles in Classical Physics', submitted to *Philosophy of Science.*

Swain, C. G. [1988]: 'Cutting a Gordian Knot: The Solution to Newcomb's Problem', *Philosophical Studies* **53**, Kluwer Academic Publishers.

Appendix: Does the physics work?

In order for Newcomb's game to be possible in the way I described it, it

needs to be the case that dwarfs can exist which satisfy the following requirements:

1. They need to be able to perform observations which are sufficiently accurate and comprehensive that, on their basis, the player's actions can be predicted 24 hours in advance.
2. They need to be able to calculate their predictions on the basis of this observational data in a sufficiently short time—or at least, to look up their predictions in a 'big book',[5] whose entries (one for each set of observational data) they have calculated in advance.

   If this is too strong a requirement, it would be sufficient to require, alternatively, that the dwarfs need only to be able to *recognize* a few particular sets of observational data, for which they have calculated the corresponding prediction in advance. This would enable them to predict the player's action 24 hours in advance at least when one of these pre-calculated situations occurs—which is clearly sufficient to yield the philosophical conclusions of this paper.
3. On the one hand, their influence on the trajectories of gigantions needs to be sufficiently limited not to disturb the validity of their predictions for the next 24 hours.
4. On the other hand, their influence on the trajectories of gigantions needs to be sufficiently strong that they can *eventually* produce the desired effects on a human scale (like the burning of human-size cheques).[6]

Proving these statements for universes satisfying particular physical laws (for example, Newtonian mechanics or special relativity with point particles), is a rather technical matter; and indeed I have not so far completed it. But there is good reason to think the proof will go through: I have shown, in the case of Newtonian point-particle mechanics, and made plausible, in the case of special-relativistic point-particle electromagnetism, that under physically reasonable conditions, *it is possible to predict any event anytime with any required accuracy just on the basis of sufficiently accurate local observational data* (Schmidt [1998a]). Thus, provided the dwarfs have sufficiently sensitive techniques at their disposal for measuring local quantities, requirement 1—which I think is the most critical—holds. However, an explicit construction of a species of dwarfs satisfying all four requirements, e.g. under the laws of Newtonian mechanics, remains to be given.

These technical complications are one reason why I chose to describe the

[5] I have shown (unpublished) that such a book would only need to contain a finite amount of information.

[6] In the story, 'eventually' was in 24 hours plus one minute. However, it would not make a difference to my conclusions if it took the dwarfs longer than that to produce effects on a human scale.

'technicalities' of the scenario in the form of a story: I think what is most important is the scenario's physical *plausibility*, rather than the details of a rigorous proof of its possibility in any particular physical setting. Such a proof would probably be much less intuitively appealing than the story I have told, and thus much less effective in supporting my thesis that the scenario *intuitively* involves backward causation (cf. Section **2**).

Another reason stems from the methodological fact that any such proof would have to start from particular physical laws, such as Newton's law of gravity or special-relativistic electromagnetism. But my thesis is not just about these particular universes: it is about *any* universe in which the above statements hold. And to date, the possibility of their holding even in *our* universe has not been excluded.