Causal Decision Theory's Predetermination Problem

T.C.P. Solomon

December 19, 2018

Abstract

It has often been noted that there is some tension between engaging in decision-making and believing that one's choices might be predetermined. Causal decision theorists need to pay attention to this tension: the possibility that our choices are predetermined forces us to consider act-state pairs in causal decision theory which are inconsistent, but which have, at least *prima facie*, non-zero causal probability, leaving us with undefined expected utilities. There are three ways to solve the problem, but all of them suffer serious costs: give up the reasoning that justifies two boxing, for any credences, in Newcomb's problem; define the outcomes of act-state pairs subjunctively, leading to the wrong results in a series of cases introduced by Arif Ahmed; or assume—contrary to our evidence—that it is impossible that our choices are predetermined, at least for the purposes of decision-making. However they choose to respond, causal decision theorists cannot remain silent: the intuitive tension between decision-making and the possibility of predetermination can be made precise and is not easily removed. Causal decision theorists have a predetermination problem.

Key Words: Causal Decision Theory, Decision Theory, Determinism, Predetermination, Free Will

1 Introduction

It is always a possibility that which option I will choose in a particular decision is, at the time I am making that very decision, already determined. This possibility makes trouble for causal decision theory (CDT) because it gives rise to decision situations in which CDT, at least naïvely understood, renders no verdict. In this paper I will explore this *predetermination problem* and what proponents of CDT might do about it. I will begin with an informal discussion of how the problem arises to give a flavour of what is going on.

Usually we are introduced to CDT via Newcomb's problem.[1] The problem goes roughly like this:

> **Newcomb's Problem:** You are on a game show and there are two boxes in front of you. One of them is transparent and contains $1,000. The other is opaque, but you are told that it contains either nothing or $1,000,000. You must choose whether to take just the opaque box (one boxing), or the opaque and transparent boxes (two boxing). The opaque box contains $1,000,000 if and only if the show's producers have predicted that you will take only the opaque box. The prediction was made 10 minutes ago during the ad break and the money has already been placed in the box or returned to the bank. You believe that the producers of the show are extremely likely to be right. They take detailed questionnaires before the show, scan your brain, run psychological tests, and observe your behaviour during the show to determine whether you will take one or both boxes—and they have never gotten it wrong before.

Should you take one or two boxes? The causal decision theorist tells you to take both, and they typically reason like this:

> The producers have already made their prediction and either put the money in

[1] See (Nozick 1969) for the canonical statement of Newcomb's problem.

> the box or returned it to the bank—nothing you do now can change the past or make money magically appear in the box (violating the laws of nature). Now, if there is already $1,000,000 in the opaque box then taking both boxes will get you $1,000 more than taking just one box. And if there is nothing in the opaque box then taking both boxes will again get you $1,000 more than taking just one box. So, either way, you should take both boxes, because doing so is guaranteed to get you an extra $1,000 and does not affect whether you will get $1,000,000 as well—even though you think that taking both boxes is good evidence that you will walk away with only $1,000.

This reasoning is, I will assume, sound—you should take both boxes.[2] But imagine that during your deliberation the following thought came to mind:

> What if my decision, along with whether the money is in the box, is already determined? Perhaps the conditions of the universe 10 minutes ago, during the ad break, together with the laws of nature, have already settled whether I will take one or two boxes. What should I do? I can't change what the conditions of the universe 10 minutes ago were, nor the laws of nature, nor whether together they already settle which decision I will make; not any more than I can change the fact that the prediction has been made and that this, together with the laws of nature, either entails that there is $1,000,000 in the opaque box or entails that there is not.

This reasoning certainly looks cogent: in Newcomb's problem causal decision theorists rely on the fact that the past and laws of nature are not under our control to argue that whether or not there is money in the opaque box is not under our (present) control. But if that is not under our present control, then neither can it be under our present control whether or

[2] I will also assume throughout that we can use monetary values as a directly proportional substitute for utilities. This is obviously unrealistic, but it will not harm the case to be made below.

not the conditions of the universe when the prediction was made, and the laws of nature, are such as to determine whether we will now decide to take one box or two; these are just facts about the past and the laws of nature. Indeed, they might be *the very same facts* that determine what the prediction was, or the prediction itself might be one of these facts. From the perspective of CDT all facts about the past and laws of nature should be on a par—if the prediction in Newcomb's problem is causally independent of our actions then so is what, if anything, we are predetermined to choose.[3] Let's keep going with the story. As a good causal decision theorist you now start to construct a decision matrix:

> There are six epistemically possible states:[4]
>
> S_1: I am now determined to one box, and the producer's have predicted one boxing.
>
> S_2: I am now determined to one box, and the producer's have predicted two boxing.
>
> S_3: I am now determined to two box, and the producer's have predicted one boxing.
>
> S_4: I am now determined to two box, and the producer's have predicted two boxing.
>
> S_5: My choice is not yet determined, and the producer's have predicted one boxing.
>
> S_6: My choice is not yet determined, and the producer's have predicted two boxing.

[3] So long, of course, as there is no backwards causation involved. I will for the remainder of this paper ignore backwards causation and time travel because all we need to get the problem is that the problematic cases are possible. It is irrelevant if there are other, different, cases involving time travel or backwards causation in which the problem does not arise.

[4] Epistemically possible should in this context be understood to mean *having non-zero probability*, or whatever the relevant sense of possibility is to include a state in our expected utility calculations.

	S_1	S_2	S_3	S_4	S_5	S_6
1B	\$1,000,000	\$0			\$1,000,000	\$0
2B			\$1,001,000	\$1,000	\$1,001,000	\$1,000

Table 1: Decision Matrix for Newcomb's Problem. Throughout I will use the notation 1B as short hand for the proposition that the decision-maker takes only one box and 2B for the proposition that they take two boxes.

> Now, if I one box and the prediction is one boxing I will get \$1,000,000. If I one box and the prediction is two boxing I will get \$0. If I two box and the prediction is one boxing I will get \$1,001,000. And If I two box and the prediction is two boxing I will get \$1,000. I can now fill in some of the utilities in the decision matrix, as in Table 1.
>
> But what about the blank squares? What will happen if I am determined to take take two boxes, but I take only one box? Or what if I am determined to take one box, but I take both boxes? Hmm.

Your puzzlement is well founded. You need to know what the outcome of two boxing would be in a world *in which you one box* (and vice versa). But knowing this is, *prima facie* at least, equivalent to knowing 'what will happen if I do not take both boxes and I do take both boxes?' There is no sensible answer to this question.

And without knowing what utilities to put in the blank squares we cannot determine whether it is better to take one box or two. The probabilities of the corresponding act-state pairs are non-zero—since as argued above what, if anything, you are already determined to choose is causally independent of your choice—and so our expected utility calculations will require us to determine the value of a non-zero value multiplied by an undefined value.

For example:

$$\begin{aligned} EU(1\text{B}) &= P(S_1) \times \$1,000,000 + P(S_2) \times \$0 + P(S_3) \times \text{undefined} + P(S_4) \times \text{undefined} \\ &\quad + P(S_5) \times \$1,000,000 + P(S_6) \times \$0 \\ &= \text{undefined} \end{aligned}$$

Hence, CDT falls silent when we consider the possibility that the outcome of our current choice is predetermined. This is *the predetermination problem*.[5]

Of course, many have noted that there is some worry about believing that the outcome of one's decision-making might be predetermined. They have, however, seldom been very specific about what this worry is. Usually they make only vague gestures towards the idea that one must believe that all of one's options are possible, in some sense or other, in order to have a real decision problem.[6] The most important upshot of this paper is that we can make precise at least one of the worries that predetermination raises for CDT, viz. that it leads to undefined expected utilities.

The lack of a precise statement of the predetermination problem also means there is a lack of solutions to it. Below I have attempted to cover the most obvious solutions to the problem in their most plausible forms (though not all of them are very plausible), but are possibilities that I cannot discuss here. I hope, nonetheless, that what is said below will achieve the primary goal of this paper: to show that the tension between the possibility of

[5] In fact this may only be *a* predetermination problem—I do not wish to rule out the possibility that predetermination (or global determinism) might cause other problems for decision-making or CDT specifically (perhaps to do with ought-implies-can principles or similar)—but it is *the* predetermination problem this paper is about.

[6] See for example (Nozick 1969, 141) or (Fernandes 2016) and those referenced therein. There is also a debate on the relationship between belief in determinism and decision-making (usually called deliberation) in the free will literature, see (McKenna and Pereboom 2016, §12.3) for a short introduction to this debate and further references. Finally, there is the debate about whether it is possible to have credences about one's own actions while engaged in decision-making. The negative answer to this question seems to be at least partly motivated by the worry that believing that one's choices are predetermined might conflict with decision-making. See (Hájek 2016) for a, critical, introduction to that debate.

predetermination and decision-making can be precisely stated and that it is a real problem for causal decision theorists—causal decision theorists have a predetermination problem.

The plan for the rest of the paper is this: next I will give a more formal definition of CDT that will act as our guide for the rest of the paper. In §3 I use this definition to precisely state the problem and discuss its ubiquity for causal decision theory. I then discuss, in §4, the three possible ways of solving the problem—modifying our understanding of causal probabilities, using subjunctive conditionals to assess utilities, and assuming that our choices are not predetermined—and why none of them are attractive. Thus setting up a trilemma for causal decision theorists.

2 Naïve Causal Decision Theory

Causal decision theory is not a single theory but a family of theories which agree on at least one thing: you should take both boxes, unconditionally (no matter what your credences are), in Newcomb's problem. Unfortunately it is not easy to find more agreement than this. In this section I will give a formal definition of CDT that will allow us to precisely state the predetermination problem while remaining neutral, as far as possible, on any contentious issues. I believe that the predetermination problem so stated can be translated into the language of all other major formulations of CDT, but unfortunately I do not have space to show this here and will have to leave it as an exercise for the reader to see how it applies to their favored formalization.

Most generally, CDT is the theory that one should perform an action, A, which maximises the quantity $EU(A)$ when compared to all other available actions. For any partition[7] $\mathbb{S}$ of the set of worlds the decision-maker assigns non-zero probability (credence) to $EU(A)$

[7] A subset whose members are mutually exclusive and jointly exhaustive.

is defined as:

$$EU(A) = \sum_{S_i \in \mathbb{S}} P(S_i || A) U(A \wedge S_i)$$

What makes this formula distinctive of *causal* decision theory—rather than expected utility maximising theories generally—is that we understand the probabilities $P(S_i||A)$ causally. Exactly what such causal probabilities amount to is a matter of some controversy. What matters to us here are three, uncontroversial, properties: First, that $P(S_i||A)$ is not always equal to $P(S_i|A)$—that is, the causal probability of S$_i$ obtaining if I do A is *not* just the conditional (evidential) probability of S$_i$ given A. Second, $P(S_i||A) = P(S_i)$ whenever S$_i$ is *causally independent* of A. Third, the past and the laws of nature are both causally independent of any act that we might perform (at least in normal circumstances with no backwards causation or time travel) and therefore the causal probability of the past and the laws of nature being a particular way on the condition that any action, ϕ, is performed is just the unconditional probability of the past and the laws of nature being that way: $P(\text{Past} + \text{Laws}|\phi) = P(\text{Pass} + \text{Laws})$. Causal decision theorists cannot give up these assumptions without giving up their *raison d'etre*: unconditional two boxing in Newcomb's problem (see §4.1). I therefore take meeting these requirements to be definitive of a decision theory being a *causal* decision theory.

I will here leave the nature of the actions involved in CDT unanalysed, taking A (unitalicised) to be an act and *A* (italicised) the proposition that A is performed. Nothing will depend on any details about what acts are. The states are taken to be sets of possible worlds.[8] The state S$_i$ is the set of possible worlds consistent with the *state description*, S_i, which is a set of propositions. We will also stipulate that for the purposes of this sum zero multiplied by an undefined term is zero; the importance of this requirement will become

[8] If you prefer to think of states as sets of propositions (or long conjunctions) this will make only a terminological difference since sets of worlds and sets of propositions are interdefinable.

apparent in the next section.

For our purposes the most important definition is of the utility, $U(A \wedge S_i)$, attached to each act-state pair. Intuitively $U(A \wedge S_i)$ is just the utility of the outcome of doing A if the world is in state S$_i$. But we need something a little more definite than this. Particularly because it will, most often, be the case that there is more than one way the world might be—more than one possible world—which might be the outcome of performing A in a state S$_i$ (the main exception being where S$_i$ is itself a singleton set containing only one possible world).

In the first instance utilities attach to possible worlds; the most basic appraisal of goodness is the appraisal of how much goodness would obtain if the whole history of the world were some fully specific way. This fact is often overlooked in setting out decision theories and it is instead presumed that we can attach utilities directly to propositions. It is however, I think, a mistake to do so. What would be the utility of $2+2$ being equal to 4? Or the utility of dogs having four legs? Or the utility of Neil Armstrong never having been to the moon? The only way to sensibly answer these questions is to appeal to some appropriate *average of the utilities of the worlds consistent with those propositions*. The question then is what this averaging function should be and, in particular, what this averaging function should be when what we are interested in is assessing the utility of propositions about performing certain actions so as to make decisions about them. It is fairly obvious that evidential decision theorists will say that this should be the evidential expected utility of performing the act in question in each of the possible worlds that are consistent with the state description. And similarly it is intuitive to think that causal decision theorists will say that it is the causal expected utility of performing the action in each of the worlds consistent with that state descriptions. We will, however, find below that we need an intermediary concept—outcomes—to be able to state the second solution to the predetermination problem—that we should understand utilities in CDT *subjunctively*—and since that solution is quite intuitive it would be unacceptable to rule it out by definition. As such I will now provide a definition of $U(A \wedge S_i)$ for causal decision theory in terms only of causal probabilities and the utility of individual possible worlds which goes via the intermediary concept of an *outcome*:

First assume that we have a primitive utility function over worlds; this is just the function telling us how good it would be to be in each possible world.[9] Call this function $u(W)$, where W ranges over all logically possible worlds. Now, let '⇒' be the appropriate conditional, which might be indicative or subjunctive, for assessing what the consequences of performing A in a state S$_i$ will be (we will return to the proper interpretation of this conditional below). We will then define the outcome of performing A in state S$_i$, $\mathbb{O}(A \wedge S_i)$, as the set of worlds given by:

$$\mathbb{O}(A \wedge S_i) = \{W| \text{ If } (A \wedge S_i \Rightarrow p) \text{ then } p \text{ is true at } W\}$$

And, we will use $U(A \wedge S_i)$ as shorthand for $U(\mathbb{O}(A \wedge S_i)$. Now, what is the utility of an outcome $\mathbb{O}(A \wedge S_i)$? First, we will stipulate that when $\mathbb{O}(A \wedge S_i)$ is a singleton set containing only one world, W, then $U(\mathbb{O}(A \wedge S_i)) = u(W)$. That is to say, if the outcome of performing an act in a state is a single possible world then the utility of performing that act in that state is just the utility of being in that world.

But what about when $\mathbb{O}(A \wedge S_i)$ is a set of worlds? A very important desideratum on decision theories—*partition invariance*—will help us work this out. A decision theory is partition invariant if and only if the expected utility of an act is the same no matter which partition of states we calculate it with respect to. Most importantly for our purposes a decision theory is partition invariant if and only if the expected utility of an act is the same calculated with respect to any arbitrarily coarse-grained partition $\mathbb{S}$ and calculated with respect to the maximally fine-grained partition $\mathbb{W}$ whose members are just the epistemically

[9] We might define this by some representation theorem, or by some other method. All that matters here is that it makes sense to associate primitive utilities with fully specified, atemporal ways the world might be.

possible worlds. I will assume here that any acceptable decision theory must be partition invariant. See (Joyce 1999, §5.5) for a detailed discussion of the need for partition invariance in CDT.[10]

Now, we can find a definition of the utility $U(\mathbb{O}(A \wedge S_i))$ when the outcome is a set of worlds by comparing the expected utility calculated with respect to an arbitrarily coarse-grained partition $\mathbb{S}$, denoted $EU_{\mathbb{S}}(A)$, and the expected utility calculated with respect to the maximally fine-grained partition $\mathbb{W}$, denoted $EU_{\mathbb{W}}(A)$. For an arbitrary decision and act we have:

$$EU_{\mathbb{S}}(A) = \sum_{S_i \in \mathbb{S}} P(S_i || A) U(\mathbb{O}(A \wedge S_i)) = \sum_{W_n \in \mathbb{W}} P(W_n || A) U(\mathbb{O}(A \wedge W_n)) = EU_{\mathbb{W}}(A)$$

Now, let $\mathbb{S}$ be a partition with n states, for arbitrary n. Then:

$$\begin{aligned} EU_{\mathbb{S}}(A) &= \sum_{S_i \in \mathbb{S}} P(S_i || A) U(\mathbb{O}(A \wedge S_i)) \\ &= P(S_1 || A) U(\mathbb{O}(A \wedge S_1)) + P(S_2 || A) U(\mathbb{O}(A \wedge S_2)) + ... \\ &\quad + P(S_n || A) U(\mathbb{O}(A \wedge S_n)) \end{aligned}$$

And, by splitting the (singleton sets of) worlds in $\mathbb{W}$ according to which $\mathrm{S}_i \in \mathbb{S}$ they are

[10] There are of course important formulations of CDT that are *not* partition invariant. Most notably the dependency hypothesis formulation due to Lewis (1981). A few comments can be made about this: First, I take it that such formulations were primarily attractive before Joyce (1999) showed that a partition invariant formulation of CDT was possible—it may be that we would accept a partition dependent decision theory if there is no partition invariant version possible, but since there are partition invariant formulations of CDT available we have good reason to accept it. Second, any plausible partition dependent CDT must have a privileged partition, but in order to avoid the predetermination problem not just *any* privileged partition will do. In particular the partition of dependency hypotheses will only avoid the problem by committing us to taking horn two of the trilemma introduced below. I discuss below in f.n. 14, after the predetermination problem is introduced more formally, what kind of privileged partition would be required to avoid the problem entirely, why this is implausible, and why Lewis's partition is just a variation on the subjunctive utilities response to the predetermination problem. Finally note that one might take this paper to constitute a reductio of partition invariant CDT, and that is, as far as I can see, a consistent response. But it is not a plausible one and will, therefore, be ignored here along with several other very implausible but apparently consistent responses.

contained in (remember that the states S_i are just sets of worlds) we get:

$$\begin{aligned} EU_{\mathbb{W}}(A) &= \sum_{W_j \in \mathbb{W}} P(W_j || A) U(\mathbb{O}(A \wedge W_j)) \\ &= \sum_{W_j \in S_1} P(W_j || A) U(\mathbb{O}(A \wedge W_j)) + \sum_{W_k \in S_2} P(W_k || A) U(\mathbb{O}(A \wedge W_k)) + ... \\ &\quad + \sum_{W_p \in S_n} P(W_p || A) U(\mathbb{O}(A \wedge W_p)) \end{aligned}$$

Then, by matching terms we find that for all i:

$$\begin{aligned} P(S_i || A) U(\mathbb{O}(A \wedge S_i)) &= \sum_{W_q \in S_i} P(W_q || A) U(\mathbb{O}(A \wedge W_q)) \\ U(\mathbb{O}(A \wedge S_i)) &= \frac{1}{P(S_i || A)} \sum_{W_q \in S_i} P(W_q || A) U(\mathbb{O}(A \wedge W_q)) \end{aligned}$$

This gives us a definition of $U(\mathbb{O}(A \wedge S_i))$ for arbitrarily coarse partitions, in terms of the outcomes of acts given the maximally fine-grained partition.[11] And for the maximally fine-grained partition the outcome of an act in any state, $\mathbb{O}(A \wedge W_q)$, is—whenever it is defined—a singleton set containing a single possible world. Recall that we have defined $U(\mathbb{O}(A \wedge W_q) = u(W)$ when the outcome is a singleton set. We have then defined the utility of $\mathbb{O}(A \wedge S_i)$ in terms of our primitive utility function $u(W)$, as long as $\mathbb{O}(A \wedge W_q)$ is defined for all W_q:

$$\begin{aligned} U(A \wedge S_i) &= U(\mathbb{O}(A \wedge S_i) \\ &= \frac{1}{P(S_i || A)} \sum_{W_q \in S_i} P(W_q || A) u(W_q) \end{aligned}$$

We are now in a position to see how the predetermination problem arises more formally.

[11] Of course $\frac{1}{P(S_i || A)}$ is undefined whenever $P(S_i || A) = 0$. But we are defining 0 multiplied by an undefined term as zero for the purposes of expected utility calculations. This means that the whole value $P(S_i || A) U(A \wedge S_i)$ will be zero, as expected, whenever $P(S_i || A) = 0$.

3 The Predetermination Problem

Returning to our Newcomb problem from §1, recall that we have six states in our decision matrix:

S_1: I am now determined to one box, and the producers have predicted one boxing.

S_2: I am now determined to one box, and the producers have predicted two boxing.

S_3: I am now determined to two box, and the producers have predicted one boxing.

S_4: I am now determined to two box, and the producers have predicted two boxing.

S_5: My choice is not yet determined, and the producers have predicted one boxing.

S_6: My choice is not yet determined, and the producers have predicted two boxing.

All of these states can be completely described by facts about the past and laws of nature, both of which are out of our causal control. Therefore, focusing on one boxing for a moment, we have that $P(S_i||\text{1B}) = P(S_i)$ for all i. As such the utility of all the states will contributed to our expected utility calculations.[12] Now, for S$_3$ and S$_4$ we get a problem. We need to find the utilities $U(\text{1B} \wedge S_3)$ and $U(\text{1B} \wedge S_4)$. Recall our definition:

$$U(\mathbb{O}(A \wedge S_i)) = \frac{1}{P(S_i||A)} \sum_{W_q \in S_i} P(W_q||A)U(\mathbb{O}(A \wedge W_q))$$

Now, both S$_3$ and S$_4$ are states in which we two box, because they are states in which the past and the laws *entail* that we two box. As such, every world contained in these states is a world in which we two box, and if we take two boxes we cannot also take only one box.

[12] Except where $P(S_i) = 0$. But it will not generally be the case that $P(S_i) = 0$ for all of the states which lead to inconsistency unless we endorse the third solution to the problem discussed in §4.3.

Then we will have for all W_q in S_3 or S_4:

$$\begin{aligned} \mathbb{O}(1\mathrm{B} \wedge W_q) &= \{W \mid \text{If } (1\mathrm{B} \wedge W_q \Rightarrow p) \text{ then } p \text{ is true at } W\} \\ &= \{W \mid \text{If } (1\mathrm{B} \wedge 2\mathrm{B} \wedge W_q \Rightarrow p) \text{ then } p \text{ is true at } W\} \\ &= \{W \mid \text{If } (1\mathrm{B} \wedge \neg 1\mathrm{B} \wedge W_q \Rightarrow p) \text{ then } p \text{ is true at } W\} \end{aligned}$$

So we need to know which conditionals of the form "$1\mathrm{B} \wedge \neg 1\mathrm{B} \wedge W_q \Rightarrow p$" are true. Intuitively decision-making is about deciding what to do in the actual world. We want to know what the consequences of performing each our options would be in each of the ways we take it that the actual world might be. It seems obvious then to understand '$\Rightarrow$' as an indicative conditional—the consequences of performing an action in a state are just the things that would happen if we performed that action in each of the possible worlds that are consistent with the state description. But, if so, there are three possibilities with regards to the conditionals of the form "$1\mathrm{B} \wedge \neg 1\mathrm{B} \wedge W_q \Rightarrow p$" :

1. All such conditionals are true, because a contradiction entails everything.
2. All such conditionals are false, because there is no fact of the matter about what would follow from a contradiction.
3. All such conditionals are without truth value, because there is no fact of the matter about what would follow from a contradiction.

Whichever of these we choose, the set $\{W \mid \text{If } (1\mathrm{B} \wedge \neg 1\mathrm{B} \wedge W_q \Rightarrow p) \text{ then } p \text{ is true at } W\}$ will be the empty set: If we choose option 1, every p and its negation $\neg p$ will be the consequent of some true conditional of the relevant form, but there are no *logically possible* worlds where every p and its negation is true.[13] If we choose option 2 or 3, there are

[13] Here it is worth noting that an assumption I am making throughout this paper is that we cannot sensibly assign primitive utilities to inconsistent worlds. We might resist this by appeal to a paraconsistent logic that allows us to distinguish between worlds where contradictions are true. However, I take it that giving up on

no true conditionals of the relevant form and hence no worlds where the consequents of those conditionals are true. Finally, $u(\varnothing)$—the primitive utility of the empty set—cannot be defined in any sensible way. Hence:

$$\begin{aligned} U(1\text{B} \wedge S_3) &= \frac{1}{P(S_3 \| 1\text{B})} \sum_{W_q \in S_3} P(W_q \| 1\text{B}) U(\mathbb{O}(1\text{B} \wedge W_q)) \\ &= \frac{1}{P(S_3 \| 1\text{B})} \sum_{W_q \in S_3} P(W_q \| 1\text{B}) u(\varnothing) \\ &= \text{undefined} \end{aligned}$$

Clearly the same result holds for S_4. Now, let's try to calculate the expected utility of one boxing (keeping in mind the causal independence claims):

$$\begin{aligned} EU(1\text{B}) &= \sum_{i=1}^{6} P(S_i \| 1\text{B}) U(1\text{B} \wedge S_i) \\ &= \sum_{i=1}^{6} P(S_i) U(1\text{B} \wedge S_i) \\ &= P(S_1) \times 1,000,000 + P(S_2) \times 0 + P(S_3) \times \text{undefined} + P(S_4) \times \text{undefined} \\ &+ P(S_5) \times 1,000,000 + P(S_6) \times 0 \\ &= \text{undefined} \end{aligned}$$

And mutatis mutandis for $EU(2\text{B})$. Causal decision theory's advice to maximise $EU(A)$ gives us no guidance here.

At the heart of the predetermination problem is the question: what will the consequences of doing A be, if I am predetermined not to do A? The problem is that intuitively there is no good answer to this question—it is equivalent to asking what will be true if a contradiction

classical logic would be far more of a cost to solve the predetermination problem than accepting any of the three horns of the trilemma below. As such we can ignore this possibility for the purposes here.

is true. But CDT requires that we have an answer to this question, because it assigns non-zero probability to our being predetermined not to do A—a fact about the past and laws of nature—*even if we do A*.

Notice that the problem would not arise if the probability of the inconsistent act-state pairs were zero. If so the undefined utilities would not make trouble for our expected utility calculations because we are assuming, for the purposes of this calculation, that an undefined term multiplied by zero is equal to zero. We need this requirement to ensure that decision theory gives us answers in *any case at all* because there are inconsistent act-state pairs in every decision if we are allowed to include even those states with zero probability. For example, states which are themselves inconsistent would give rise to undefined utilities by introducing a contradiction to the antecedent of the conditionals used to find outcomes. But this is not a problem for CDT because such states have zero probability, and we have stipulated that an undefined term multiplied by zero is equal to zero.

Similarly, you might think that the issue would arise with state descriptions such as 'I will two box' (making the problem merely about considering *what I will do*, rather than about predetermination specifically). But the *causal* probability that I will two box if I one box is zero, when the fact that I will two box is a fact about the future and not entailed by the past and laws of nature—if I can take either one or two boxes then taking one box will cause me not to take two boxes. This is not the case, however, with predetermination because in such cases the fact that I will perform some particular action is entailed by facts—about the past and the laws—which are out of my causal control. We can see now also why evidential decision theory does not face the predetermination problem: the evidential probability that I am predetermined to do A, given that I do not do A, is zero ($P(A \wedge S_i | \neg A) = 0$), so all of the problematic act-state pairs have zero probability.

Finally, while the predetermination problem only became salient in Newcomb's problem when we explicitly considered the possibility that our choice was predetermined, it is in fact ubiquitous—the predetermination problem affects every possible decision, whether or not we explicitly consider the possibility of predetermination. This is guaranteed by partition invariance. (So long, that is, as it is in fact a possibility at all that our acts are predetermined). Partition invariance requires that the expected utility of an act is the same whatever partition of states we choose, and in particular it must always be the same as the expected utility calculated with respect to the maximally fine-grained partition. But the maximally fine-grained partition—made up of the fully specific worlds—necessarily includes states, for every option A, where we are predetermined not to do A. The expected utility calculated with respect to the maximally fine-grained partition, therefore, *always* includes utility terms for inconsistent act-state pairs whose causal probability is non-zero. Unless we can solve the predetermination problem, partition invariance guarantees that the expected utility of every act in every situation is undefined. The predetermination problem represents a fundamental tension between causal decision theory and the possibility that our choices are predetermined.

4 The Trilemma

We have now stated the predetermination problem quite precisely. How might we solve it? There are three options that will avoid the combination of undefined utility and non-zero probability, allowing us to give a definition of expected utility that avoids the predetermination problem:

Horn One: (Re)Define $P(S_i||A) = 0$ whenever S_i entails that we will not do A—that is, accept that performing an act can *cause* the past and laws not to be a particular way.

Horn Two: (Re)Define $\mathbb{O}(A \wedge S_i)$ so that it is non-empty even when S_i entails that we will not do A—that is, use some account of counter*possible* conditionals to define the consequences of our acts.

Horn Three: (Re)Define $P(S_i) = 0$ whenever S_i entails that we will not do ϕ for *any act* ϕ under consideration in the present decision—that is, assume, for the purposes of CDT, that our present choice is not predetermined.

Horns one and three solve the predetermination problem by ensuring that the inconsistent act-state pairs all have zero probability, while horn two solves the predetermination problem by ensuring that utilities are defined even for inconsistent act-state pairs. I call these options *horns* because each has serious costs, and, as such, they constitute a trilemma for causal decision theorists. I will now discuss the most obvious and plausible ways of implementing each solution and the costs of doing so. Unfortunately there are more possible responses than I can cover here. However, I trust that this discussion will demonstrate that the predetermination problem is not trivially solved: causal decision theorists must take on significant costs to avoid it.[14]

[14] As promised I can now discuss why a privileged partition formulation of CDT is not a plausible response to the predetermination problem. First, Lewis's (1981) formulation of CDT privileges the partition of dependency hypotheses: "maximally specific proposition[s] about how the things [the decision-maker] cares about do and do not depend causally on [their] present actions" (Lewis 1981, 11). Now, if we assume that such dependency hypotheses use indicative conditionals it will be the case that they *cannot* always specify what would happen if the agent performed some action, because the possibility of predetermination ensures that there is not always a fact about this. In other words, if dependency hypotheses use indicative conditionals then some of the dependency hypotheses will have to be *partial*, but if they are then Lewis's formulation of CDT will break down because it assumes that the dependency hypotheses are complete. On the other hand, if we assume (as Lewis does) that the dependency hypotheses should use *subjunctive* conditionals then Lewis's formulation is (implicitly) equivalent to taking horn two of the trilemma. Using subjunctive conditionals in dependency hypotheses ensures that every act-state pair is mapped to some outcome, but only at the expense of ensuring that that outcome is not always consistent with every possible state.

Second, any partition dependent formulation of CDT that did avoid the predetermination problem, without taking one of the horns of the trilemma, would have to have a privileged partition such that it never distinguishes between states in which our choice is predetermined and those in which it is not. But if it does so then it will not be able to model extremely simple decision problems. For example, imagine that you are trying to decide whether to bet $1,000 on the truth of Bohmian mechanics. Intuitively the relevant states are those described by "Bohmian mechanics is true" and "Bohmian mechanics is false". But if the privileged partition does not allow us to distinguish states on the basis of whether or choice is predetermined or not then we cannot use these states—because Bohmian mechanics is a deterministic theory which entails that all our choices are predetermined. But a decision theory which says it is impossible to even consider whether or not Bohmian mechanics is true is a very implausible decision theory indeed.

4.1 Horn One: Causal Probabilities

The past and the laws of nature are causally independent of anything we can do now. Yet surely the probability that I will (would) be predetermined to do *A*—a fact about the past and the laws of nature—if I do (did) *not* do *A* is zero. There is a tension between these two intuitions. The first possible solution to the predetermination problem is to side with the second and define causal probabilities so that the probability of being predetermined to do A is zero on the assumption that I do not do A. That is, define causal probabilities so that $P(S_i || A) = 0$ whenever S$_i$ entails ¬A, even if S$_i$ can be specified with reference only to facts about the past and laws of nature. This will ensure that the inconsistent act-state pairs involved in the predetermination problem always have zero probability, thus solving the problem.

The problem with this solution is that it forces us to give up the first intuition and, therefore, the standard argument for two boxing in Newcomb's problem based on the causal independence of the past from our choices. This will force us to give up either two boxing, no matter what your credences are, in Newcomb's problem or two boxing, no matter what your credences are, in very similar cases I call *Bonus Money Newcomb Problems*. Here I will introduce Bonus Money Newcomb Problems and show that this solution forces us to endorse one boxing in some of them, contrary to the reasoning which led us to embrace causal decision theory in the first place. Consider the following case:

> **$1,000,000 Bonus Newcomb Problem:** You are on a game show and there are two boxes in front of you. One of them is transparent and contains $1,000. The other is opaque, but you are told that it contains either nothing, $1,000,000, or $1,000,000 plus a bonus of $1,000,000. You must choose whether to take just the opaque box (one boxing), or the opaque and transparent boxes (two boxing). The opaque box contains $1,000,000 if the show's producers have

> predicted that you will take only the opaque box. The prediction was made 10 minutes ago during the ad break and the money has already been placed in the box or returned to the bank. You believe that the producers of the show are extremely likely to be right. They take detailed questionnaires before the show, scan your brain, run psychological tests, and observe your behaviour during the show to determine whether you will take one or both boxes—and they have never gotten it wrong before. The opaque box contains the bonus $1,000,000 if and only if the past and laws of nature already determine that you will choose to take only the opaque box.

This case is just the standard Newcomb's problem except that there will be a bonus of $1,000,000 in the opaque box if the past and the laws of physics entail that you will choose to take just one box. Now, should you take one or two boxes? Clearly the causal decision theorist should endorse taking both boxes, no matter what your credences are. Nothing you can do now can change whether or not the bonus money is in the box, no more than it can change what the prediction was and whether the first $1,000,000 is in the opaque box. Of course you will receive the most money if it turns out that you are predetermined to one box, and there is a sense in which that is the best outcome. But you should take two boxes now, because nothing you can do now will change how much money is in the opaque box. No matter what you do now taking both boxes will get you $1,000 more than taking only a single box. Whether or not you get the bonus money is causally independent of what you do now, so causal decision theorists should treat this case as no different to the standard Newcomb problem—if you should two box, no matter what your credences are, in one then you should do so in the other as well.

The amount of the bonus involved in the $1,000,000 Bonus Newcomb Problem is arbitrary: if causal decision theorists should endorse unconditional two boxing here they should do so *for any possible amount of bonus money*. I will call cases that are the same as the above with the \$1,000,000 bonus replaced by an arbitrary bonus of \$$a$, Bonus Money Newcomb Problems.

The only way for a causal decision theorist to argue that you *should* one box in a Bonus Money Newcomb Problem is to admit that the past and laws of nature can depend *causally* on our choices. But if they do that then their argument in the standard Newcomb problem is undermined: why should we take both boxes if it is possible that taking only one box will *cause* there to be more money in the opaque box? Even if causal probabilities can be defined such that CDT endorses two boxing for every credence function in the standard Newcomb problem, this will not be enough if CDT using those same causal probabilities endorses one boxing in a Bonus Money Newcomb Problem because doing so undermines the reason for being a causal decision theorist in the first place.

Now, the suggestion is that we define causal probabilities so that the following requirements are satisfied:

1. $P(S_i||A) = 0$ when $\mathrm{S}_i \supset \neg A$—that is, the causal probability of being predetermined not to do A is zero on the assumption that you do A. This is the requirement that solves the predetermination problem.

2. $P(S_i||A) \neq 0$ when $\mathrm{S}_i \supset A$—that is, the causal probability of being predetermined to do A on the assumption that you do A is *not*, in general, zero. This is the requirement that differentiates this solution from the third horn of the trilemma.

3. $P(S_i||A)$ obeys the axioms of the probability calculus. And in particular it obeys *normalisation*: $\sum_{S_i \in \mathbb{S}} P(S_i||A) = 1$—that is, it is certain that if I do A then the world will be in some state or other among those in our decision problem (we can use $S_i \in \mathbb{S}$ here because $\mathbb{S}$ is assumed to be a partition). This requirement is not explicitly part of the original suggestion, but it is required for any plausible definition of causal probabilities. I take it that normalisation is intuitively obvious enough not to need defence

here, so I will simply assume that our causal probabilities must be so normalised. If a defence is needed it should suffice to note that any causal probability function that is not so normalised will be open to Dutch books.

However, given an acceptable credence function and a mapping from credences to causal probabilities (a definition of causal probabilities) that obeys these three requirements we can always find a value of a for which CDT will endorse one boxing in the Bonus Money Newcomb Problem whose bonus is \$$a$. But, as we have seen, CDT should never endorse one boxing in any Bonus Money Newcomb Problem.

The proof of this is quite simple. First, define the states S_1 through S_6 as for the standard Newcomb problem above. Then the decision matrix for an arbitrary Bonus Money Newcomb Problem, with bonus \$$a$, is as in Table 2.

	S_1	S_2	S_3	S_4	S_5	S_6
1B	\$1,000,000 + \$a	\$0			\$1,000,000	\$0
2B			\$1,001,000	\$1,000	\$1,001,000	\$1,000

Table 2: Decision Matrix for an arbitrary Bonus Money Newcomb Problem.

To meet requirement 1—and avoid the predetermination problem—we must have $P(S_3 \| 1\text{B}) = 0$, $P(S_4 \| 1\text{B}) = 0$, $P(S_1 \| 2\text{B}) = 0$, and $P(S_2 \| 2\text{B}) = 0$. Substituting these into the causal expected utility formula we have:

$$EU(1\text{B}) = P(S_1 \| 1\text{B}) \times (\$1,000,000 + \$a) + P(S_5 \| 1\text{B}) \times \$1,000,000$$

$$EU(2\text{B}) = (P(S_3 \| 2\text{B}) + P(S_5 \| 2\text{B})) \times \$1,001,000 + (P(S_4 \| 2\text{B}) + P(S_6 \| 2\text{B}) \times \$1,000$$

We also know that $EU(2\text{B}) < \$1,001,000$, because $EU(2\text{B})$ is maximised when $P(S_3 \| 2\text{B}) + P(S_5 \| 2\text{B})$ is maximised, but normalisation guarantees $P(S_3 \| 2\text{B}) + P(S_5 \| 2\text{B}) \leq 1$. Hence, the maximum value that $EU(2\text{B})$ can have for any credence function is \$1,001,000.

We can now derive a formula for a (the value of the bonus) which will allow us to construct, given a credence function and definition of causal probabilities, a Bonus Money Newcomb Problem for which CDT endorses one boxing. First we will find the function for a that guarantees $EU(1\text{B}) > EU(2\text{B})$:

$$
\begin{aligned}
&EU(1\text{B}) > EU(2\text{B})\\
&P(S_1||1\text{B}) \times (\$1,000,000 + \$a) + P(S_5||1\text{B}) \times (\$1,000,000) > EU(2\text{B})\\
&P(S_1||1\text{B}) \times \$a + (P(S_1||1\text{B}) + P(S_5||1\text{B})) \times (\$1,000,000) > EU(2\text{B})\\
&P(S_1||1\text{B}) \times \$a > EU(2\text{B}) - (P(S_1||1\text{B}) + P(S_5||1\text{B})) \times (\$1,000,000)\\
&\$a > \frac{EU(2\text{B}) - (P(S_1||1\text{B}) + P(S_5||1\text{B})) \times (\$1,000,000)}{P(S_1||1\text{B})}
\end{aligned}
$$

So, we know that when this inequality is satisfied CDT will endorse one boxing. But there is *always* a value of a which satisfies this inequality for any causal probability function which satisfies the three requirements above. This is guaranteed by the fact that $0 \leq EU(2\text{B}) \leq \$1,001,000$, and that all the probabilities must be values between zero and one. The the numerator of the right hand side is, therefore, a real number between $-\$1,000,000$ and $\$1,001,000$, while the denominator is a real number strictly greater than zero and less than one. The function $\frac{x}{y}$ is well defined for every value of x and y within these bounds, and hence there will always be some well defined minimum value of a satisfying the inequality. *This is true even if the causal probabilities depend on the value of a.*

Now we have a method for producing a Bonus Money Newcomb Problem where CDT endorses one boxing given a credence function and definition of causal probabilities (mapping from credences to causal probabilities) which meets the three requirements: simply take any probability distribution which satisfies the proposed definition of causal probabilities, enter it into the right hand side of the above inequality, calculate the resulting value, and finally pick a value of a that is larger than this value. CDT using the proposed definition of causal probabilities, and the specified crdences, will endorse one boxing for a Bonus

Money Newcomb Problem using this value of a.

Notably, the relevant value of a need not be greater than zero—there are many possible definitions of causal probability meeting the three requirements that will get the answer wrong in the standard Newcomb problem. But even if a plausible definition of causal probabilities can be found which succeeds in guaranteeing two boxing when $a = 0$ (the standard Newcomb problem), there will still be some value of the bonus \$$a$ for which CDT using the given probabilities will endorse one boxing. And we have seen that endorsing one boxing for any value of \$$a$ requires us to give up the original reasoning in favour of two boxing in Newcomb's problem and thus CDT's *raison d'etre*.

It is worth stressing the strength of this result. It is not merely that we can find some credences for which a proposed definition of causal probabilities endorses one boxing in a Bonus Money Newcomb Problem, but rather that we can find a Bonus Money Newcomb Problem for which any given credence function that satisfies the requirements above must endorse one boxing. That is, we can allow the defender this solution to specify not only how credences are mapped to causal probabilities but *which credences in particular we should use* and still come up with an example in which they endorse the wrong result. If causal probabilities obey the requirements above then for *every rational agent* there is some Bonus Money Newcomb Problem in which they should one box. This points again to the fundamental nature of the predetermination problem.

Perhaps an argument can be made that we should sometimes one box in Bonus Money Newcomb Problems, and that the original argument for CDT was mistaken to rely on the causal independence of the past and laws of nature from our choices.[15] But such a theory will not be causal decision theory as we know it.

[15] There are at least two authors, Cantwell (2010, 2013) and Edgington (2011), who argue that we should understand CDT in this way.

4.2 Horn Two: Subjunctive Outcomes

Causal decision theory is often thought to have a strong connection with subjunctive conditionals (see, for example, (Joyce 1999, 172)). The second way to solve the predetermination problem takes up this idea and suggests that while the inconsistent act-state pairs have positive probabilities, we can assign them sensible utilities—and thus avoid the predetermination problem—by appeal to subjunctive conditionals when assessing their outcomes. Recall that the outcome of doing A in state S$_i$ is:

$$\mathbb{O}(A \wedge S_i) = \{W| \text{ If } (A \wedge S_i \Rightarrow p) \text{ then } p \text{ is true at } W\}$$

Above I assumed that '⇒' should be read as an indicative conditional and that, therefore, the set of true claims $(A \wedge S_i \Rightarrow p)$ is either trivial or empty when A and S_i are inconsistent (as they are in the predetermination problem). But given an appropriate subjunctive account of '⇒' the set of true claims $(A \wedge S_i \Rightarrow p)$ will be neither trivial nor empty; we will then have a well defined set of possible worlds $\{W| \text{ If } (A \wedge S_i \Rightarrow p) \text{ then } p \text{ is true at } W\}$ even when A and S$_i$ are inconsistent and the predetermination problem will be avoided.

The first challenge for this response is to provide such an account of subjunctives—that is, to provide an appropriate account of counter*possible* conditionals. Giving such an account is not necessarily easy. Unfortunately it is difficult to assess the plausibility of this subjunctive response to the predetermination problem without knowing the details of the relevant account of counterpossibles. We can, however, make some broad comments about any such account that will show that using counterpossibles to define outcomes is not as unproblematic as it might at first seem.[16]

First note that in order to solve the predetermination problem the worlds picked out by

[16] A minor problem for using counterpossibles in this way: in order to define utilities for non-maximally fine-grained states we needed to assume that the outcome of performing A in any maximally fine-grained state (a singleton set of one possible world) was also a singleton set of one possible world. The relevant account of counterpossibles will have to ensure that this condition is met.

the condition ‘ If $(A \wedge S_i \Rightarrow p)$ then p is true at W’ must be logically possible worlds—otherwise the outcome of A and S$_i$, which ranges over only possible worlds, is still empty. But if A and S$_i$ are inconsistent, then these worlds are logically possible only if at least one of the following is true for each such world:

1. A is not performed.
2. The past or laws of nature are different to the past or laws of nature of S_i.
3. There are miracles (violations of the laws of nature).

If all three of these claims are false then the resulting world is one in which A is performed and it is not, because the past and laws of nature of S$_i$ entail that A is not performed. We can assume that the relevant account of counterpossibles will not tell us that the outcome of doing A is a world in which we do not do A (that it will not make 1 true). Doing so would, extremely implausibly, imply that we should weigh in favour of, or against, doing A what would happen if we did not do A. We are left with two options:

Epistemically Possible Subjunctive Outcomes: We are weighing in favour of, or against, doing A what would happen in some state S_2 ($\neq S_1$) with non-zero subjective probability, which is therefore already included in our expected utility calculation. This will, again, lead to the wrong result in some Bonus Money Newcomb Problems.

Epistemically Impossible Subjunctive Outcomes: We are weighing in favour of, or against, doing A what would happen in some state S_2 with zero subjective probability—in which case we are counting in favour of, or against, A something which should not count in favour of, or against, *anything* we could do. This will lead to the wrong results in cases of the kind introduced by Arif Ahmed (2013, 2014a; 2014b, Ch. 7).

Before we examine these options more closely, note that there is no obvious reason that the relevant outcome of counterpossibles cannot take the first option in some cases and the second option in others. We might suggest that this will allow us to avoid the counterexamples. But it can only do so if the relevant account of counterpossibles always generates epistemically impossible outcomes in every Bonus Money Newcomb Problem, and generates epistemically possible outcomes in every case like Ahmed's Bet (introduced below). I cannot prove that no such account can be produced, but it seems implausible that any such account can be produced that will not be either unacceptably ad hoc or for which some minor variation of these cases would produce a counterexample. We move on now to discussing the two options.

4.2.1 Epistemically Possible Subjunctive Outcomes

The first option is to use an account of counterpossibles which generates outcomes for inconsistent act-state pairs, A and S$_i$, which are epistemically possible, and therefore must have a different past or laws of nature to S$_i$. That is, we read '$\Rightarrow$' in such a way that $\{W|$ If $(A \wedge S_i \Rightarrow p)$ then p is true at $W\}$ is a set of epistemically possible worlds. But we can then see that this set, call it $\mathbf{W}_i$, must be the same as $\{W|$ If $(A \wedge W_i \Rightarrow p)$ then p is true at $W\}$. And since the members of $\mathbf{W}_i$ are, by hypothesis, epistemically possible we can define a state S$_k$ (remember that states are sets of worlds) which is just $\mathbf{W}_i$. That is to say, if the relevant account of counterpossibles generates outcomes for performing an act A in a state which is inconsistent with A whose members are epistemically possible worlds then it generates outcomes which are the same as the outcomes of performing A in some other state S$_k$.[17]

This is a problem because it allows us to show that defining outcomes in this way is

[17] In case you are suspicious that S$_k$ need not be part of any partition involving S$_i$ note that, given partition invariance, we can run the above argument in terms of the maximally fine grained partition to show that the relevant account of counterpossibles must generate outcomes for inconsistent act-world pairs which are the same as the outcome of performing that act in some epistemically possible world. This will be enough for our purposes below, with the appropriate changes in notation. We could alternatively define a new partition by reference to S$_i$—namely, the partition $\{S_i, \neg S_i\}$—which is guaranteed to make the relevant equivalence claims true. And, again given partition invariance, showing that there is a problem for this partition is all we need.

mathematically equivalent to defining causal probabilities in one of the ways which meets the three requirements in §4.1. And, as we have seen, any way of doing that allows the construction of Bonus Money Newcomb Problems where the resulting account of CDT endorses one boxing—which it should not do.

Intuitively the problem is that there is no difference between moving the credence we apply to one outcome to a different outcome (as redefining causal probabilities does) and moving the utility of the outcomes in the opposite direction (as using counterpossibles in this way does).

More formally: Let S$_i$ be a state in which we are predetermined not to do A, S$_k$ a state that is consistent with doing A, and P_I a causal probability function as defined by the three conditions on causal probabilities introduced in §3—namely, one on which the past and laws are *always* causally independent of our choices. Then we will find that:

$$\begin{aligned} EU(A) &= ... + P_I(S_i \| A)U(A \wedge S_i) + P_I(S_k \| A)U(A \wedge S_k) + ... \\ &= ... + P_I(S_i \| A)U(A \wedge S_k) + P_I(S_k \| A)U(A \wedge S_k) + ... \\ &= ... + (P_I(S_i \| A) + P_I(S_k \| A))U(A \wedge S_k) + ... \end{aligned}$$

But this is the same as defining causal probabilities such that $P_D(S_i \| A) = 0$ and $P_D(S_k \| A) = P_I(S_j \| A) + P_I(S_i \| A)$ (P_D because this causal probability function allows the past and laws to *depend causally* on our choices). And any such causal probability function satisfies the three requirements in §4.1: it will be normalised (as long as P_I was, which we can assume it was), it assigns zero probability to the inconsistent act-state pairs, and it does not assign $P_D(S_n \| A) = 0$ when S$_n$ entails that we will do A (unless P_I did, which we can assume it did not). With such a definition we would have:

$$\begin{aligned} EU(A) &= ... + P_D(S_i||A)U(A \wedge S_i) + P_D(S_k||A)U(A \wedge S_k) + ... \\ &= ... + 0 \times U(A \wedge S_i) + P_D(S_k||A)U(A \wedge S_k) + ... \\ &= ... + (P_I(S_i||A) + P_I(S_k||A))U(A \wedge S_k) + ... \end{aligned}$$

We can therefore map *any* account of counterpossibles that generates outcomes in this way to an account of causal probabilities satisfying the three requirements in §4.1. And any such account of causal probabilities must endorse one boxing in some Bonus Money Newcomb Problems—forcing us to give up the standard reasoning in Newcomb's problem that lead us to CDT in the first place.

Any account of counterpossibles that generates outcomes for inconsistent act-state pairs which consist of epistemically possible worlds is, therefore, equivalent to—and faces the same costs as—defining causal probabilities as in horn one. Of course, it might (again) be possible to make an argument that this is the right result, but the resulting decision theory will not be CDT as we know it.

4.2.2 Epistemically Impossible Subjunctive Outcomes

Arif Ahmed (2013, 2014a; 2014b, Ch. 7) has provided a series of examples involving predetermination which he argues are counterexamples to CDT.[18] Ahmed's analysis of the verdict CDT gives in these cases is (more or less implicitly) based on assuming a subjunctive account of the outcomes in CDT, and in particular a subjunctive account which assigns epistemically impossible worlds to some outcomes. In this section I will examine a slightly modified version of one of Ahmed's examples and show that using counterpossibles to generate epistemically impossible outcomes give the wrong result. The example goes as

[18] Actually Ahmed's examples generally involve reference to *determinism*, but since all my choices must be predetermined if the world is deterministic this will not make any difference to us.

follows (adapted from (Ahmed 2014a, 666)):

> **Ahmed's Bet:** In my pocket, Billy says, I have a slip on which is written a proposition, P. You must choose between two bets. Bet 1 is a bet which pays \$10 if P is true and costs \$1 if P is false. Bet 2 is a bet which pays \$1 if P is true and costs \$10 if P is false. Before you choose whether to take bet 1 or bet 2, I should tell you what P is. It is the proposition that the state of the world yesterday was such as to determine that you now take bet 2.

If we understand the outcomes of act-state pairs subjunctively, and allow them to be epistemically impossible worlds then it is very natural to think that the outcome of taking bet 1 if P is true is that we will get \$10 (this is epistemically impossible because it would require a violation of the laws of nature) and that the outcome of taking bet 2 if P is false is loosing \$10. We will then have the decision matrix in Table 3.

	I am now determined to take bet 2 (P)	I am now determined to take bet 1 (¬P)	The outcome of my decision is not yet determined (¬P)
Take bet 1	\$10	-\$1	-\$1
Take bet 2	\$1	-\$10	-\$10

Table 3: Decision matrix for Ahmed's Bet.

Whether or not we are already determined, if at all, to take bet 1 or bet 2 is independent of what we choose now: the truth of P is entirely settled by facts about the past and the laws of nature over which we have no control now. According to CDT we can, therefore, rely on dominance reasoning. And it is obvious that bet 1 will dominate bet 2, it is better no matter how the world turns out to be. We should, on this understanding of subjunctive outcomes, take bet 1 no matter what our credences are.

But, Ahmed argues, this is the wrong result[19]: if we have a high enough credence that our choice in this decision is predetermined then we should take bet 2. We are certain that if

[19] I am somewhat modifying Ahmed's actual argument here since he argues on the basis that we are *certain* that our choice is predetermined. The modification does, I think, no harm.

our choice is predetermined then taking bet 2 will get us \$1 because the only epistemically possible outcome in which our choice is predetermined and we take bet 2 is one in which P is true. And we are certain that if our choice is predetermined then taking bet 1 will loose us \$1 because the only epistemically possible outcome in which our choice is predetermined and we take bet 1 is one in which P is false. Hence, if our choice is predetermined it is better to take bet 2. Then, if we are sufficiently sure that our choice is predetermined we should take bet 2 (specifically if our credence that our choice is predetermined either way is greater than $\frac{9}{11}$). Ahmed's reasoning here seems solid—it is very intuitive to think that we should take bet 2 if we are sure enough that our choice is predetermined. But this contradicts the result that using the suggested account of counterpossibles gives us: that we should take bet 1 no matter what our credences are.

Using an account of counterfactuals which generates epistemically impossible outcomes will lead us to the wrong result in this case, and cases like it. It does so because it weighs in favour of some actions outcomes which we are certain will never obtain. Decision-making is about deciding what it is best to do in the actual world as it stands, and, therefore, what the consequences of an action will be in an epistemically impossible world should not weigh in favour of, or against, it.

We see then that neither of the obviously plausible ways of using counterpossible subjunctive conditionals to define the outcome of inconsistent act-state pairs is successful. If the relevant account of counterpossibles generates epistemically possible outcomes then it leads to the same failure as allowing the past and laws of nature to depend on our choice. If the relevant account of counterpossibles generates epistemically impossible outcomes then it leads to the wrong results in cases like Ahmed's Bet. Generating outcomes using counterpossibles is either a surreptitious way of accepting that the past or the laws of nature are causally dependent on our choices, or it brings into consideration possibilities which are irrelevant to working out what we should do in the actual world—namely epistemically impossible ones.

4.3 Horn Three: Assuming Our Choices Are Not Predetermined

The final way to solve the predetermination problem is to simply insist that predetermination is not a possibility at all in decision-making. Indeed there is a long tradition of such insistence, often traced back to Kant's claim that we must *act under the idea of freedom* (Nelkin 2011, 117). Specifically, we can solve the predetermination problem by assuming that every world in which the outcome of our current decision is predetermined has zero probability. That is, assigning $P(S_i) = 0$ whenever S_i entails that we will perform ϕ for any act ϕ at all. The problematic act-state pairs would then all have zero probability, solving the predetermination problem. At the same time, we avoid endorsing one boxing in any Bonus Money Newcomb Problem—including the standard Newcomb problem—because CDT only does so when the probability $P(S_1) \neq 0$, but S_1 is a state which entails that we will one box and is, therefore, assigned zero probability if we grasp this third horn of the trilemma. There are, however, heavy costs to pay for ruling out the possibility that our choices are predetermined in decision-making. I will examine these now.

A rational agent can, and very likely must, have a non-zero credence that the outcome of their current decision is predetermined. This possibility is compatible with any set of evidence (including our actual evidence which suggests a non-zero credence in deterministic theories like Bohmian mechanics) and, as such, should not be ruled out a priori. How then can we claim that predetermination is not a possibility in decision-making? There are two basic strategies: we could deny the very plausible claim that credences must always be proportioned to our evidence, or we could deny that credences, understood as degrees of belief, are always the probabilities that we should use in decision-making.

There is little need, I take it, to discuss here the cost of denying the principle that one's credences should be proportioned to one's evidence. This principle is eminently plausible and denying it is a very high cost to pay in order to solve the predetermination problem. Note especially that we are not merely saying it is acceptable to treat as false, for the purposes of decision-making, some proposition which we give very low credence to. One might be sympathetic to the idea that, at least for non-ideal agents like us, it is acceptable to treat as false in our decision-making some claims with a low enough probability, because doing so will reduce cognitive burden (and thus the probability of mistakes) without substantially increasing the likelihood that we will make the wrong decision. For example, I can plausibly ignore the possibility that all the world's cats will suddenly develop super intelligence and rise up to enslave humanity when I am planning for my future—the probability of such an occurrence is extremely low and entertaining it is very likely a waste of time for a cognitively limited agent like me. But the possibility of predetermination is not like this, because scientific evidence might rationally lead us to assign the possibility of predetermination a very high credence. The suggestion that CDT requires that the probability of predetermination is zero might require that the probability of what we take to be *far and away the most likely scenario* is zero. The suggestion is not that we ignore a particular possibility while our credence in it is very low, as with respect to the possibility of a super-intelligent cat uprising, but that we ignore a particular possibility *no matter what our credence in it is*. (This applies as well to the second strategy of divorcing credences and the probabilities used in decision-making.) Giving up the idea that our credences should be proportioned to our evidence in such an extreme manner is a very high price to pay to solve the predetermination problem.

The difficulties that will face divorcing credences and the probabilities used in decision-making are somewhat more subtle. The primary challenge is simply giving a detailed account of theoretical and practical reasoning on which they employ distinct probability functions. This is especially difficult because separating credences and the probabilities used in decision-making forces us to give up on many elements of the standard (Bayesian) picture of the relationship between theoretical and practical reasoning. The success of this picture is therefore an objection to making such a distinction.

For example, separating credences and the probabilities used in decision-making will force us to change how we understand the representation theorems which are often taken to be central to decision theory. Such representation theorems should still be sound, but they will not be able to tell us (directly) about an agent's epistemic state—they will only be able to tell us about the probabilities that the agent uses in their decision-making. We might use these probabilities to infer things about an agent's epistemic state, but we will no longer have a tight link between an agent's preferences or choices and their degrees of belief.

As another example, separating credences and the probabilities used in decision-making will undermine Dutch book arguments for probabilism about credences. If I can, rationally, escape a Dutch book by merely making sure the probabilities I use in decision-making obey the probability axioms then I have no need to make sure my degrees of belief also obey the probability axioms (at least not on account of Dutch books). Even worse, the pressure to use probabilities in decision-making which satisfy the probability axioms may not be as strong as the pressure to have probabilistic credences was: the probabilities used in decision-making are, presumably, to be understood pragmatically, but there are several pragmatic ways to avoid Dutch books other than making sure the probabilities I use conform to the probability axioms (for example, simply refusing to take the bets in a Dutch book).

The final problem for this solution is that it endorses the same option in Ahmed's Bet as the solution above in §3. According to this solution the only possible state in Ahmed's Bet is one where P is false, and it is obvious we should take bet 1 in Ahmed's Bet if P is false—doing so will, with certainty, loose us only \$1, whereas taking bet 2 would, with certainty, loose us \$10. However, there is some reason to think that a proponent of this solution can undermine Ahmed's intuitions, which rely on reasoning about what we should do if we are sufficiently sure that our choice is predetermined; something that the proponent of this solution can claim is necessarily mistaken. Undermining Ahmed's arguments, or accepting that the right thing to do in such cases is very counterintutive, are small prices to pay in comparison to the other costs that this solution faces.

Insisting that it is not a possibility that our choices are predetermined when we are engaged in decision-making about them will force us to give up either a very plausible understanding of the relationship between credences and evidence, or a very plausible understanding of the link between theoretical and practical rationality. Either is a high price to pay to solve the predetermination problem.

5 Conclusion

The possibility that our choices are predetermined is a problem for causal decision theory: The possibility of predetermination gives rise to inconsistent act-state pairs in our decision problems—guaranteed by partition invariance to occur even when predetermination is not salient—and intuitively there is no fact of the matter about what the outcome of these inconsistent act-state pairs would be. But their causal probability is, *prima facie*, not zero. Hence, our expected utilities are undefined whenever it is a possibility that our choice is predetermined, and it appears that this is always possibility. We see then that the intuitive tension between believing that one's choices are predetermined and engaging in decision-making is not an illusion—it can be made precise.

We have canvassed the three ways to solve this problem and seen that their most plausible implementations (and some less plausible ones) all face serious costs. Causal decision theorists are faced with a trilemma:

Horn One: Define causal probabilities such that $P(S_i||A) = 0$ whenever S_i entails that we will not do A—that is, accept that performing an act can *cause* the past and laws not to be a particular way. Doing so will force us to endorse one boxing in some Bonus Money Newcomb Problems and, therefore, to give up the reasoning that justifies two boxing in Newcomb's problem—the resulting decision theory will not be causal decision theory as we know it.

Horn Two: Use counterpossible conditionals to define the outcome of doing A in state S$_i$, $\mathbb{O}(A \wedge S_i)$, so that it is non-empty even when S_i entails that we will not do A. Doing so is either equivalent to a version of horn one, or it forces us to weigh in favour of some acts outcomes which we are certain will not obtain, and thereby reach the wrong conclusion in cases of the kind discussed by Arif Ahmed (2014b, Ch. 5).

Horn Three: Define the probabilities used in decision-making so that $P(S_i) = 0$ whenever S$_i$ entails that we will not do ϕ for *any act* ϕ under consideration in the present decision—that is, assume, for the purposes of CDT, that our present choice is not predetermined. Doing so forces us to either give up the idea that credences should be proportioned to our evidence, or the standard picture of a unified probability function for theoretical and practical reasoning.

I have not here attempted to adjudicate which horn will be least painful for the causal decision theorist to grasp. Whatever the answer turns out to be, we can see that it is not an easy choice. Causal decision theorists have a predetermination problem.

References

Ahmed, Arif. 2013. “Causal Decision Theory: A Counterexample.” *Philosophical Review* 122, no. 2 (January 1): 289–306.

———. 2014a. “Causal Decision Theory and the Fixity of the Past.” *The British Journal for the Philosophy of Science* 65 (4): 665–685.

———. 2014b. *Evidence, Decision and Causality.* Cambridge: Cambridge University Press.

Cantwell, John. 2010. “On an Alleged Counter-Example to Causal Decision Theory.” *Synthese* 173, no. 2 (March): 127–152.

———. 2013. “Conditionals in Causal Decision Theory.” *Synthese* 190:661–679.

Edgington, Dorothy. 2011. “Conditionals, Causation, and Decision.” *Analytic Philosophy* 52, no. 2 (June): 75–87.

Fernandes, Alison. 2016. “Varieties of Epistemic Freedom.” *Australasian Journal of Philosophy* 94 (4): 736–751.

Hájek, Alan. 2016. “Deliberation Welcomes Prediction.” *Episteme* 13 (04): 507–528.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory.* Cambridge studies in probability, induction, and decision theory. Cambridge ; New York: Cambridge University Press.

Lewis, David. 1981. “Causal Decision Theory.” *Australasian Journal of Philosophy* 59, no. 1 (March): 5–30.

McKenna, Michael, and Derk Pereboom. 2016. *Free Will: A Contemporary Introduction.* Routledge contemporary introductions to philosophy. New York, NY: Routledge.

Nelkin, Dana Kay. 2011. *Making Sense of Freedom and Responsibility.* Oxford: Oxford University Press.

Nozick, Robert. 1969. “Newcomb’s Problem and Two Principles of Choice.” In *Essays in Honor of Carl G. Hempel,* edited by Nicholas Rescher, 114–146. Dordrecht: Springer Netherlands.