Causal Decision Theory is false

*Abstract*. Causal Decision Theory (CDT) cares only about the effects of a contemplated act, not its causes. The paper constructs a case in which CDT consequently recommends a bet that the agent is certain to lose, rather than a bet that she is certain to win. CDT is plainly giving wrong advice in this case. It therefore stands refuted.

***1: The argument***

Billy makes Suzie an offer.

In my pocket (he says) I have a slip on which is written a proposition P. You must choose between two bets. Bet 1 is a bet on P at 10:1 for a stake of $1. Bet 2 is a bet on P at 1: 10 for a stake of $10. So your profits and losses are as in Table 1.

|  |  |  |
| --- | --- | --- |
|  | P | ¬P |
| Bet 1 | $10 | -$1 |
| Bet 2 | $1 | -$10 |

**Table 1**

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 cause you now to take Bet 2.

Let soft determinism be the doctrine that (a) Suzie is free now to take either bet and (b) that at every past time there already existed a determining cause for her now taking whichever bet she actually *does* take. Suppose that soft determinism is true and that Suzie believes it. Suppose also that what Billie says is true and that she believes it. Should she take Bet 1 or Bet 2?

I believe that Causal Decision Theory (CDT) recommends that Suzie take Bet 1. I also believe Suzie *should* take Bet 2. I conclude that Causal Decision Theory is false.

This paper aims to justify those beliefs. But first let me describe the case in more detail.

***2: The argument in more detail***

My description of Suzie’s case is completely open about the betting mechanism, slightly inaccurate as well as vague about soft determinism, and slightly inaccurate about the content of P. Let me now state the first and second more precisely and the second and third more accurately.

*2.1:* *The betting mechanism*. Suzie has a choice between now raising her hand and doing something else. Suppose for simplicity that she will in fact either *raise* her hand or *lower* her hand. Suppose also that she knows this. Then it seems to Suzie that she has the choice between now raising her hand and now lowering her hand.

Suppose also that *raising* her hand signals to Billie that she accepts Bet 1, and that *lowering* her hand signals to Billie that she accepts Bet 2. So her choice between raising her hand and lowering her hand *is* her choice between Bet 1 and Bet 2. Suppose that she knows all of this.

*2.2. Soft Determinism*.

I have just said that soft determinism is the view that (a) Suzie is free to take either bet; and that (b) at every past time there already existed a determining cause for her now taking whichever of Bet 1 and Bet 2 she actually *does* take.

Talk of ‘determining causes’ is a helpful way to get across the general idea. But we can explain soft determinism more precisely and more accurately without it. Soft determinism tells us this about Suzie’s present betting behaviour:

1. **Soft Determinism:** (a) Suzie does freely whatever she will actually do. (b) There is a true historical proposition H about the intrinsic state of the world long ago, and there is a true proposition L specifying the laws of nature that govern @, such that H and L jointly determine what Suzie will do.[[1]](#footnote-0)

(a) is the soft part. (b) is the determinism part. ‘@’ rigidly designates the actual world.

I should explain four other expressions that occur in (1): ‘Suzie does *freely* whatever she will actually do’, ‘about the *intrinsic* *state* of the world long ago’, ‘the *laws of nature* that govern @’ and ‘H and L *jointly determine* what Suzie will do’.

I can’t define ‘Suzie does freely whatever she will actually do’ in any way that satisfies all of our intuitions about ‘free’. But I don’t need to. My argument only cares about soft determinism because it cares about its consequences for CDT. So I only need to give a *necessary* condition for ‘Suzie does freely whatever she will actually do’ that connects (1) with CDT. The relevant necessary condition is that what Suzie now does is free in the sense that CDT should *tell* her what to do.

This is plausible. After all, it is plausible that if you are free to do A and free to do B instead then there is some point in *deliberating* whether to do A or to do B. And it is plausible that CDT should *tell* you which of A and B to do if there *is* some point in deliberating which of A and B to do. So I propose:

1. **Freedom:** Suzie does freely whatever she will actually do only if CDT should tell her what to do.

Lewis once defined an intrinsic property of a *region* as one such that, whenever two possible regions are perfect duplicates, the property belongs to both or neither.[[2]](#footnote-1) Extrapolating from this let us try:

1. **About the intrinsic state of the world at time t:** A proposition H is about the intrinsic state of the world at time t if and only if, whenever two possible worlds w1 and w2 are perfect duplicates at t, H is either true at *both* of w1 and w2 or true at *neither*.

One might object to the consequence of (3) that all necessary truths are about the intrinsic state of the world at all times. That consequence may be counterintuitive but it is harmless for *my* purposes, because in this paper I will only discuss *contingent* historical propositions.

Relatively little turns upon how precisely we explain ‘the laws of nature that govern @’. Lewis’s account will serve perfectly well:

1. **Laws of nature:** The laws of nature that govern @ are those regularities that would come out as axioms in a system that was optimal among (actually) true systems in its combination of simplicity and strength.[[3]](#footnote-2)

One consequence of this definition that *will* be important is that it allows what (1) requires, namely that the laws are *true*. (In fact it entails that the laws are true.) So Suzie knows that any law of our world tells the truth about it.

As for ‘determine’, we might try this: H and L determine what Suzie will do if and only if: *either* it is deducible from (sentences expressing) H and L that Suzie will *raise* her hand, *or* it is deducible that she will *lower* her hand. But there are two objections to this.

First, determination as in (1b) clearly relates H and L to possible *events*, not to propositions or sentences.

Second, no sensible soft determinist would commit herself to the existence of a true proposition or sentence that describes the movement of Suzie’s hand and which is also deducible from H and L. After all H and L probably make no mention of macroscopic bodily movements at all. It is more likely that H describes e.g. distributions of gravitational and electromagnetic potential at a past time and that L describes e.g. the relation between these factors across space and time. So it is unlikely that any sentence describing the motion of Suzie’s hand is deducible from H and L.

I prefer to explain determination in terms of *supervenience*:

1. **Determination:** Propositions P1, P2, … Pn jointly *determine* a possible event E if and only if E occurs at every possible world at which P1, P2, … Pn are true.

According to (1)-(5) then, Suzie’s soft determinism is soft because she thinks that Causal Decision Theory should tell her *whether* to raise her arm. And it is deterministic because she thinks that some historical proposition H, that is true at all worlds that duplicate the state of @ at a distant time t (including @ itself), and some proposition L, that states the laws of @ (and so truly describes @), are such that any world where both are true is one where Suzie does whatever she *actually* does. So if she raises her arm at @ then she raises her arm at *every* world where H and L are both true. If she lowers her arm at @ then she lowers her arm at *every* world where H and L are both true.

*2.3 The Content of P.*

As I described it in the original story P was the statement that something happened *at some time* in the distant past to *cause* Suzie now to *take Bet 2*. That is not quite right: in fact P is more specific about the *time* that it describes, it doesn’t mention *causation* at all, and it mentions Suzie’s *bet* in terms of a bodily motion.

Let T be some particular time in the distant past: let it be Christmas Day, 10 million BC. Then what P says is this:

1. **The proposition P**: There is a true historical proposition H describing the intrinsic state of the world at T, and there is a proposition L describing the laws of nature at @, such that H and L jointly determine that Suzie now lowers her hand.

So P is quite specific about the time that it describes (Christmas Day, 10 million BC). P doesn’t mention causation but only laws of nature (as explained in (4)). And P doesn’t actually describe Suzie’s bet on or against P itself in such terms: it only states that back then at T, there existed nomological antecedents for Suzie’s now lowering her hand. (So it is possible to explain the content of P without mentioning P itself.)

P also involves the concepts of intrinsicality and joint determination. You should read these as explained at (3) and (5) respectively.

Finally, the actual world itself enters into the content of P by means of the rigid designator ‘@’. The rigidity of ‘@’ makes no difference to the truth-value of P evaluated *at the actual world*: the result of replacing ‘@’ in (6) with some flexible but actually co-designative term, like ‘my favourite world’ is to express a proposition P\* that is true if and only if P is true. But it *does* make a difference to the truth-value of P evaluated *at non-actual worlds*. Assuming that my favourite world at any world *w* is *w* itself, P is true and P\* is false at some worlds that match @’s intrinsic state at T but falsify its laws; and P\* is true but P is false at some of those other worlds that do *not* match @’s intrinsic state at T *and* falsify its laws.

This is going to be important because as we shall see, Causal Decision Theory demands that Suzie evaluate P at a range of possible worlds that includes, but does not *only* include, the actual world. In particular the consequence of ‘@’’s rigidity that matters for Causal Decision Theory is:

1. **P’s Modal Profile:** For any possible world *w*, P is true at *w* if and only if: there is a true historical proposition H describing the intrinsic state *of w* at T, and there is a proposition L describing the laws of nature *at @*, such that H and L jointly determine that Suzie now lowers her hand.

Note especially the consequence of (7) that the truth-value of P at a world supervenes on the intrinsic state of that world at time T. This is because it follows from (7) that P is true at any world *w* that contains at T what in the actual world would be nomological antecedents of Suzie’s now lowering her hand, even if she *doesn’t* lower her hand *at w*. Suppose for instance that the laws L of the actual world are such that any L-world at which electron e is in state S on Christmas Day, 10 million BC, is a world at which Suzie now lowers her hand. Then P is true at *every* world *w* at which electron e is in state S on that date, even if Suzie doesn’t exist at *w*.

I don’t see anything philosophically incoherent about offering Suzie a bet on the truth of a proposition that I need a rigid designator to express. After all, there is nothing wrong in offering her a bet on, say, the proposition that Benjamin Franklin was the inventor of bifocal lenses, even though ‘Benjamin Franklin’ is rigid. Of course we are also supposing that Suzie has at least enough knowledge of modal metaphysics to see that and why (7) follows from (6). But that looks harmless. Whatever advice Causal Decision Theory gives to someone who *is* modally literate should be at least as good as its advice to anyone who *isn’t*.

*2.4 The argument again*.

In light of these clarifications and amendments to the original statement of the argument, let me now put it in more detail.

What Billie actually says to Suzie is this:

Raising your hand commits you to betting $1 at 10:1 on the truth of the proposition P as stated at (6). Lowering your hand commits you to betting $10 at 1:10 on the truth of the same proposition P. Your options, my abbreviations for them, the relevant states of nature and the payoffs are as in Table 2.

|  |  |  |
| --- | --- | --- |
|  | P | ¬P |
| A1: You raise your hand | $10 | -$1 |
| A2: You lower your hand | $1 | -$10 |

**Table 2**

What do you do?

I claim that this case refutes Causal Decision Theory. My argument is as follows.

1. **Relevance Premise:** If CDT is true then it gives Suzie correct advice.
2. **Descriptive premise:** CDT advises Suzie to do A1.
3. **Normative premise:** The correct advice to Suzie is to do A2.
4. **Conclusion:** CDT is false

The argument is plainly valid. Moreover the Relevance Premise (8) follows straightforwardly from (1a) and (2). It therefore remains only to defend the Descriptive Premise (9) and the Normative Premise (10).

***3: The Descriptive Premise***

In order to defend the Descriptive Premise (9) I first outline Causal Decision Theory and then argue that it advocates A1.

*3.1 Causal Decision Theory.*

There are at least five formulations of Causal Decision Theory.[[4]](#footnote-3) I will use the attractively simple formulation of Gibbard and Harper.

But first a word about those other formulations. It is plausible that some of them make different recommendations from Gibbard-Harper Decision Theory to some agents in some situations.[[5]](#footnote-4) Still, they *all* agree with Gibbard-Harper Decision Theory about what *Suzie* should do in *her* situation. I cannot argue for that proposition here. But at the end of 3.2 I will briefly say why any Decision Theory that *deserves* the name ‘causal’ *should* agree with Gibbard-Harper Decision Theory about what Suzie should do in her situation.

In what follows let ‘A1’, ‘A2’, … denote, sometimes a possible act that is open to an agent, and sometimes the proposition that she performs that act. In all cases the context will make it clear which is which.

Let ‘N1’, ‘N2’, … denote the elements of some relevant partition over the state of the world. Let *V* (*p*) denote the desirability or news-value of the proposition *p* i.e. of its truth. Let Cr (*p*) denote the agent’s degree of confidence or credence in the proposition p. Let ‘→’ express the *non-backtracking counterfactual*, of which more in a moment.

Then the *utility* *U* of an act Ai is given by:

1. U (Ai) = ∑j Cr (Ai → Nj) . V(Ai ∧ Nj)

That is: U (A) is the sum, over all the elements of the partition, of the agent’s desirability for A given each element Nj of the partition, weighted by the agent’s Cr (A → Nj), that is, her confidence that if she *were* to do A then Nj *would* be true.

Now Causal Decision Theory advises agents to *U-maximize*. It says that if you have a choice between acts then you should do whichever of those acts has the *greatest U-value*.

This account of Causal Decision Theory is not yet adequate for my purposes. I still need to explain ‘non-backtracking counterfactual’, so let me now do that.

By calling ‘→’ a non-backtracking counterfactual I mean to convey what Lewis calls the ‘standard’ interpretation of counterfactual talk[[6]](#footnote-5). Lewis’s semantic framework makes ‘A → C’ true at @ if and only if there are possible worlds at which A and C are both true ((A ∧ C)-worlds) that are ‘closer’ to @ than any world at which A is true and C is *false* (any (A ∧ ¬C)-worlds).[[7]](#footnote-6)

Everything now turns on what it is for one possible world to be *closer* to @ than another. Lewis’s own best-known spelling out of the criteria for closeness[[8]](#footnote-7) has been widely criticized[[9]](#footnote-8); but it has (at least Lewis thought that it had[[10]](#footnote-9)) a widely accepted consequence for counterfactuals about ordinary, everyday acts carried out in ordinary everyday circumstances.[[11]](#footnote-10) Here is the consequence:

1. If A describes the occurrence of a possible event *E* at time *tE* then the intrinsic state of all the closest A-worlds matches that of @ at all times up until shortly before *tE*.[[12]](#footnote-11)

What happens at those A-worlds *after* *tE* depends on whether or not A is true. If A is true then ‘all’ the closest A-worlds just *are* @ and so trivially match @’s intrinsic state at *all* times, not only pre-*tE* times. If A is false then shortly before *tE* there is, at each closest A-world *w*, a ‘small miracle’—that is, an unobtrusive violation of the laws of @. This miracle does just enough to ensure—given the laws of @, to which *w* thereafter conforms—that A is true at *w*.

For instance, let A be the proposition that Nixon presses this button at *t* and let C be the proposition that his clothes are ready at *t’* = *t* plus 1 hour. (It is a button on Nixon’s washing machine.)

If A is true then the closest A-worlds just *are* @ itself. So A → C is true at @ if and only @ itself is a C-world, that is, if and only if Nixon’s clothes are in fact ready at *t’*.

If A is false then Nixon does *not* in fact press this button. And (13) says that the closest worlds at which he presses it are worlds that exactly match @ until shortly before *t*. Then, at each such world *w*, a small miracle occurs—perhaps a neuron in Nixon’s brain that did not fire at @ *does* fire at *w*—and this miracle is just enough to bring it about that Nixon *does* press the button at *w*. Each of the closest A-worlds then unrolls in accordance with the laws of @. And A → C is true at @ if and only if C is true at all such closest A-worlds, that is, if there are A-worlds at which Nixon’s clothes are ready at *t’* that are closer to @ than any at which they are not.

Given this explanation of ‘→’ it is easy to see why a theory that advocates U-maximization should be called *Causal* Decision Theory. Which of two contemplated acts A1 and A2 gets the higher utility is typically going to depend in part upon the differences between Cr (A1 → Nk) and Cr (A2 → Nk) for each element Nk of the partition. But these differences reflect Nixon’s opinion of the *causal dependence* of Nk upon his act. For instance, Nixon is certain just before time *t* that his clothes *would* be ready at *t*’ if he *were* to press this button at *t*; also that they would *not* be ready at *t*’ if he were *not* to press this button at *t*. That reflects his certainty that whether his clothes are ready at *t*’ is *causally* dependent on whether he presses the button at *t*. And it is ultimately because of this causal dependence, and the fact that he wants a clean shirt ready for teatime, that CDT counsels Nixon to press the button at *t*.[[13]](#footnote-12)

*3.2 Causal Decision Theory prefers A1*.

Let us now apply this analysis to the problem at hand. Assuming that Suzie’s utility for a payoff is its cash value in dollars, it follows from (12) and Table 2 that:

1. U (A1) = 10.Cr (A1 → P) - Cr (A1 → ¬P)
2. U (A2) = Cr (A2 → P) - 10.Cr (A2 → ¬P)

Which of A1 and A2 CDT recommends depends on, and only on, which of U (A1) and U (A2) is greater.

Now it might look as though we cannot answer this question until we know the values of Suzie’s Cr (A1 → P), Cr (A1 → ¬P), Cr (A2 → P) and Cr (A2 → ¬P). And I haven’t stipulated enough about the case to settle what these quantities *are*. However, this is not really a problem. It is possible to show that *whatever* they are, Cr (A1 → P) and Cr (A2 → P) are *equal*, as are Cr (A1 → ¬P) and Cr (A2 → ¬P). And it is also possible to show that this suffices for U (A1) > U (A2). I turn now to these two tasks.

My first task is to show that Suzie’s Cr (A1 → P) and her Cr (A2 → P) are equal. Note first that whichever of them is actually true and whichever of them is actually false, *both* A1 *and* A2 describe events that are supposed to occur *now* or in the *near future*. It therefore follows from (13) that the intrinsic states of all of the closest A1-worlds *and* of all of the closest A2-worlds match that of @ at all times up until shortly before now. Hence the intrinsic states of all of the closest A1-worlds and those of all of the closest A2-worlds all match *one another* at all times up until shortly before now.

In particular therefore the common intrinsic state of all of the closest A1-worlds matches that of all of the closest A2-worlds *at time T*, i.e. on Christmas Day, 10 million BC. For T should not count as a time that is ‘shortly before’ the present. (If it does then just let T be an earlier time that doesn’t so count. Such a time surely exists if (13) has any content at all.)

Now recall the consequence of (7) that I mentioned immediately after stating (7). It is that P’s truth-value supervenes on the intrinsic state of the world at T. It follows from this and the foregoing that *either* P is *true* at all of the closest A1-worlds *and* at all of the closest A2-worlds, *or* P is *false* at all of the closest A1-worlds *and* at all of the closest A2-worlds. Hence by Lewis’s semantics we have:

1. ((A1 → P) ∧ (A2 → P)) ∨ ((A1 → ¬P) ∧ (A2 → ¬P))

Now Suzie could easily have gone through all of this reasoning. (There is no harm in supposing that she did. The advice that CDT gives to people who know something about the semantics of counterfactuals should be at least as good as the advice that it gives to everyone else.) So Suzie can be certain of (16). But certainly (16) implies (17) and (18):

1. (A1 → P) ↔ (A2 → P)
2. (A1 → ¬P) ↔ (A2 → ¬P)

So Suzie should be certain of (17) and (18). So if she is rational then her credences must conform to the following:

1. Cr (A1 → P) = Cr (A2 → P)
2. Cr (A1 → ¬P) = Cr (A2 → ¬P)

That completes the first part of my task.

The second task was to infer that U (A1) > U (A2). Substituting (19) and (20) into (15) we get:

1. U (A2) = Cr (A1 → P) - 10. Cr (A1 → ¬P)

And assuming that Suzie’s credence function takes any argument to a non-negative value, it follows from (14) and (21) that U (A1) > U (A2). This completes the second part of my task and establishes that CDT recommends A1, which is the Descriptive Premise (9).

It is possible to see intuitively what I have just established more formally. Which column of Table 2 Suzie is in is something that it is quite beyond her powers to *affect*. After all, it depends only upon what intrinsic state the world was in on Christmas Day, 10 million BC. So it is already completely settled and determinate which column Suzie is in. But *whichever* column she is in, the payoff to A1 is strictly greater than the payoff to A2. In the jargon, A1 *strongly dominates* A2.

Now however exactly you formulate it, anything that is worth calling *Causal* Decision Theory is going to prefer any act to every act that it strongly dominates whenever the columns are known to be *causally* independent of the rows.[[14]](#footnote-13) [[15]](#footnote-14) So Causal Decision Theory inevitably prefers A1 to A2.

***4: The Normative Premise***

It is difficult to *argue* that A2 is preferable to A1 because it seems intuitively so obvious that it is (and everyone on whom I have tried this example agrees that it is.) But let me say a few words to underscore that intuition.

By (1b) Suzie *knows in advance* that anyone who lowers her hand in her situation was determined by L and P to lower her hand; and she knows in advance that anyone who *raised* her hand in this situation was *not* determined by L and P to lower her hand. So she is in a position to know in advance that if she lowers hand now then P is true, and that if she raises her hand then ¬P is true. So she is in a position to know in advance, that anyone who lowers her hand in her situation is *bound* to make a $1 *profit*, and that anyone who raises her hand in that situation is *bound* to make a $1 *loss*.

Furthermore, by playing this game once it would be possible for a bookie to be certain of making $1 out of any hand-raiser. By repeatedly playing this game it would be possible for a bookie to be certain of making an indefinite amount of money out of any hand-raiser. Everyone is in a position to know all of this in advance.

What *could* be better reasons for saying that hand-raising, that is, A1, is practically irrational?

Perhaps the following way of putting the intuition is more forceful. I stress that I am not giving an *argument* for the intuition. I am only presenting the intuition in a way that makes it seem especially compelling, at least to *me*.

Let N be the proposition that the actual laws of nature have always held and will continue to hold for the next five minutes. Now consider the following material biconditionals:

1. (A1 ∧ N) ↔ (A1 ∧ ¬P)
2. (A1 ∧ ¬N) ↔ (A1 ∧ P)
3. (A2 ∧ N) ↔ (A2 ∧ P)
4. (A2 ∧ ¬N) ↔ (A2 ∧ ¬P)

I claim that (22) and (23) are true at all the closest A1-worlds and that (23) and (24) are true at all the closest A2-worlds.

Consider first the left-to-right direction of (22). Suppose that (A1 ∧ N) is true at some closest A1-world *w*. Then N is true at *w*. Now if A1 is false at @ then all of the closest A1-worlds contain an N-falsifying miracle and so N is *false* at *w*. So A1 is true at @. So all of the closest A1-worlds are identical to @. So *w* = @. But by (6), if A1 is true at @ then P is false at @. Hence (A1 ∧ ¬P) is true at @, that is, at *w*. This establishes that the left-to-right direction of (22) is true at all the closest A1-worlds.

Now consider the right-to-left direction of (22). Suppose that (A1 ∧ ¬P) is true at some closest A1-world *w*. Then ¬P is true at w. So by (13), ¬P is true at @. But if ¬P is true at @ then by (1b) A1 is true at @. Hence all of the closest A1-worlds are identical to @. Hence w = @. But clearly N is true at @. Hence (A1 ∧ N) is true at @, that is, at w. This establishes that the right-to-left direction of (22) is true at all the closest A1-worlds.

Hence (22) is true at all the closest A1-worlds. And (22) ⊃ (23) is a logical truth. So (23) is true at all the closest A1-worlds. The same reasoning shows that (24) and (25) are true at all the closest A2-worlds.

So what? The point is that taken together, the closest A1-worlds exhaust all of the worlds that a causalist will regard as relevant to Suzie’s evaluation of A1; and the closest A2-worlds exhaust all of the worlds that a causalist regards as relevant to Suzie’s evaluation of A2. It follows that Table 3 below is just as accurate as Table 2 when it comes to representing Suzie’s situation. It represents the same payoffs to the same actions in the same circumstances at all the possible worlds where this could matter to a causalist.

|  |  |  |
| --- | --- | --- |
|  | N | ¬N |
| A1: raise hand | -$1 | $10 |
| A2: lower hand | $1 | -$10 |

**Table 3**

More precisely: the fact that (22) is true at all the closest A1-worlds implies that the top left-hand entry of Table 3 should match the top right-hand entry of Table 2. The fact that (23) is true at all the closest A1-worlds implies that the top right-hand entry of Table 3 should match the top left-hand entry of Table 2. The fact that (24) is true at all the closest A2-worlds implies that the bottom left-hand entry of Table 3 should match the bottom left-hand entry of Table 2. And the fact that (25) is true at all the closest A2-worlds implies that the bottom right-hand entry of Table 3 should match the bottom right-hand entry of Table 2.

It follows from *this* that anyone who advocates A1 in the problem as presented in Table 2 must advocates A1 in the problem as presented in Table 3.

But Table 3 shows clearly that A1 *is*—that is, it has the payoff structure of—a bet against *the laws of nature*. And A2 is similarly a bet *on* the laws of nature. More precisely: A1 is a bet at 10:1 for a stake of $1 on the proposition that in five minutes’ time the laws of nature *will* have been broken. A2 is a bet at 1:10 for a stake of $10 on the proposition that in five minutes’ time the laws of nature will *not* have been broken. And Suzie is certain that the laws of nature will *not* have been broken in five minutes’ time.

A1 is therefore equivalent to betting *against* a certainty. A2 is equivalent to betting *on* that very certainty. If doing A1 when A2 is available isn’t *obviously* irrational then I don’t know what is. This concludes my case for the normative premise (10).

So all three premises of my argument are true. So the conclusion is true: Causal Decision Theory is false. I turn to four objections.

***5: Objections***

*5.1: Table 1 and Table 2 are misleading*

One might object that Tables 1 and 2 misrepresent the outcomes. They present outcomes as being open to Suzie that are *not* open to her. In particular: assuming that Suzie is certain of determinism she must be certain that she will now raise her hand if and only if some determining cause of that event existed on Christmas Day, 10 million BC. She must also be certain that she will now lower her hand if and only if some determining cause of *that* event existed back at that time.

But if Suzie is certain of these things then she should be certain that neither the top left-hand entries nor the bottom right-hand entries of Tables 1 and 2 represent outcomes that are open in her circumstances. So Suzie’s decision problem really looks like this:

|  |  |
| --- | --- |
|  | Determinism |
| A1: Raise hand | -$1 |
| A2: Lower hand | $1 |

**Table 4**

And (of course) it is now clear that CDT prefers A2 to A1. So my descriptive premise (9) is false.

The formal answer to this objection is that *Causal Decision Theory itself* is committed to treating *all four* entries in Tables 1 and 2 as relevant to Suzie’s deliberation, that is, as outcomes that are open to anyone in her situation.

Consider for instance the top left-hand entry in Table 2. Before she knows what she’ll do Suzie has non-zero credence in (A2 → P). By (19) this is also her credence in (A1 → P). So her credence in (A1 → P) is non-zero. And *this* credence is the weight that she gives to the payoff in the top left-hand box of Table 2 when evaluating her U (A1) in accordance with equation (12); that weight is therefore non-zero. U-maximization, that is, Causal Decision Theory, is therefore committed to giving non-zero weight to the possibility represented by the top left-hand entry in Table 2. Exactly parallel reasoning applies to for the bottom right-hand entry of Table 2. And nobody was ever in any doubt that the bottom left-hand and top right-hand entries in Tables 1 and 2 should get non-zero weight. Exactly the same reasoning goes for Table 1. So CDT is committed to giving non-zero weight to all four options. To insist on Table 5 is to reject the Causal theory altogether.

So I *am* entitled to assume that Tables 1 and 2 correctly represent Suzie’s options. I am entitled to assume it because Causal Decision Theory is committed to it, and the point of this paper is to argue against Causal Decision Theory.

Less formally: Causal Decision Theory regards the worlds that are open to a free agent as those that *would* obtain were she to act otherwise than she actually does, even if those worlds are *certainly* *non-actual*. For instance, if P is true then as far as Causal Decision Theory is concerned, the closest (A1 ∧ P)-worlds are just as open to Suzie as the closest (A2 ∧ P)-worlds, even though the closest (A1 ∧ P)-worlds are certainly not actual. And if ¬P is true then not only the closest (A1 ∧ ¬P)-worlds but also the closest (A2 ∧ ¬P)-worlds are open to Suzie, even though the closest (A2 ∧ ¬P) worlds are certainly not actual. Since she does not know which of P and ¬P is true she must weight these two possible option-sets by Cr (P) and Cr (¬P) respectively. So the causalist is committed to giving all four outcomes non-zero weight.

*5.2 The agency theory of causation*.

One might alternatively argue that CDT only *seems* to recommend A1 because it only *seems* that whether or not P is true is causally independent of what Suzie does. On the correct theory of causation this is in fact *not* the case.

I have in mind the theory that what is definitive of the causal relation is its relevance to control and manipulation. Causes are primarily causal *handles*. That theory makes B causally dependent upon A whenever the statistical correlation between A and B survives the hypothesis that A is open to a free agent’s direct manipulation.[[16]](#footnote-15)

To give an idea of its content: the theory implies that EPR-style violations of the Bell inequalities involve temporally backwards causation. The statistical correlation between the prior state of the emitting apparatus and the subsequent setting of the detector survives the hypothesis that the latter is under the experimenter’s direct control. So the *later* setting of the detector has a causal influence on the *earlier* state of the electrons it detected, that is, their state on emission from the apparatus.[[17]](#footnote-16)

It also implies that Suzie’s act is causally relevant *to*—as well as being caused *by*—the state of @ in the distant past. Hence (13) fails. At least it fails if the counterfactual ‘→’ is supposed to be tracking causal dependence, which it must do if (a) the definition of U is as in (12) and (b) U-maximization just is *Causal* Decision Theory properly so-called. Since (13) fails, so too does my argument for (17) and (18). Instead we have:

1. Cr (A1 → P) = 0
2. Cr (A1 → ¬P) = 1
3. Cr (A2 → P) = 1
4. Cr (A2 → ¬P) = 0

But from Table 2, (12), and (26)-(29) it follows that CDT prefers A2 to A1.

This is not the place to discuss the agency theory at the length it deserves. For what it’s worth, my own view is that our everyday causal concept probably has the shape that it does as a result of *several* independent pressures. The need to mark out what is open to manipulation is certainly one of these. But the need to mark out certain types of continuous or quantity-preserving processes is another, the need to make predictions is a third, and the need to give explanations is probably a fourth. An account of causation that focuses on just one of these pressures is therefore unlikely to yield the everyday causal concept. It is more likely to yield one of at least three possible distillates of it.

If the agency theory tells me that certain outlandish quantum phenomena involve backwards causation then I think I could still accept that it is a theory of causation, that is, an analysis of a pre-existing concept. But if the agency theory is going to save CDT from preferring A1 then it must be implying that if determinism is true, then *every* action has a temporally backwards effect upon every one of those past events that was a determining cause of it. It must be implying that human actions everywhere and always—and in contrast with other sorts of event—are the causes as well as the effects of each of their determining causes. Any theory that implies this seems to me not to be talking about *causation* at all.

Of course it is dogmatic to legislate over just how far a theory of ‘causation’ would have to stretch credulity before it becomes reasonable not to call it a theory of causation at all. If you think that a theory of causation properly so-called *can* have this amazing consequence then I have no more objections to make. You needn’t take my argument to show that CDT is false. You will take it to show that *if* CDT is true then everything that we do has a causal influence upon the distant past.

That conclusion still seems to me to be novel and interesting and so worth pointing out. You are welcome to draw it.

*5.3 Newcomb’s Problem*.

The next objection concedes that CDT recommends A1 and concedes that A1 definitely loses money. ‘But what is so interesting about that? The problem is structurally analogous to Newcomb’s problem[[18]](#footnote-17) with a predictor who is certainly unerring. And it is well known that in Newcomb’s problem Causal Decision Theory recommends an option (“two-boxing”) that definitely loses money, at any rate relative to its alternative (“one-boxing”). So what does your example teach us that we could not have learnt from Newcomb’s problem?’

Two replies. First: there are many reflective people whose intuition is that CDT makes the *right* recommendation in Newcomb’s problem, even when the predictor is certainly unerring.[[19]](#footnote-18) But *everybody* on whom I have tried out Tables 1 and 2—in philosophy seminars and in pubs—prefers A2 to A1. (In particular the causalists all preferred A2 to A1 but insisted that CDT does not after all recommend A1. That is why I started section 5 with replies to two objections of that sort.)

So the charge that CDT gives the wrong advice in Suzie’s situation certainly has more extensive appeal—though perhaps indeed no more justice—than the charge that CDT gives the wrong advice in a Newcomb situation.

Second: it is perhaps *especially* easy to dismiss that version of Newcomb’s Problem that involves unerring prediction on the grounds that it involves an obviously *fantastic* example. So CDT makes the wrong recommendation in science fictional cases: so what?

You might reply that for Newcomb’s problem to refute CDT it is not necessary that a supernatural predictor be realistic but only that the agent *believe* in one. But only a *space cadet* could seriously believe that a supernatural predictor has correctly predicted his next act and rewarded him accordingly.[[20]](#footnote-19) So CDT gives the wrong recommendation to space cadets: again, so what?

By contrast some of my best friends are soft determinists, and *they* are not space cadets; so CDT *ought* to makes the right recommendation to *them*. So the argument from Newcomb’s problem to ¬CDT is vulnerable to an objection that does not obstruct the argument in this paper.

*5.4 Drop soft determinism*

The causalist might also claim that my argument shows not that *CDT* is false but that *soft determinism* is false. Either Suzie’s bet is nomologically undetermined by its past or she does not bet *freely* at all, at least not in any sense of ‘freely’ that satisfies (2). So my Relevance Premise (8) is false.

If that response *were* enough to get around the difficulty then it would be tempting to bill my argument as a novel refutation of *soft determinism* rather than as a novel refutation of CDT.

But it’s hardly clear that it *is* enough. It is easy to imagine *stochastic* laws governing @ that enforce, say, Cr (A1 ↔ Q) = 20% and Cr (A1 ↔ ¬Q) = 80% upon any agent who takes their evidence seriously and whose next act is certainly A1 or A2, for some Q that describes the intrinsic state of @ at T. Then we can arrange the following payoffs:

|  |  |  |
| --- | --- | --- |
|  | Q | ¬Q |
| A1 | $3 | -$1 |
| A2 | $1 | -$3 |

**Table 5**

Then CDT recommends A1 over A2; but it is easily verified (by a similar argument to that in Section 4) that A1 is equivalent to taking a bet at 3:1 on a proposition in which the laws of nature dictate that one’s credence be 20% (irrational), whereas A2 is equivalent to taking a bet at 1:3 on a proposition in which the laws of nature dictate that one’s credence be 80% (rational). So CDT *still* recommends betting against the laws of nature.[[21]](#footnote-20)

The causalist who takes this line is therefore committed to saying that whatever empirical evidence there is, *either* for determinism *or* for stochastic laws of the sort that I have just mentioned, is *ipso facto* evidence that CDT simply does not apply to the kind of agents that we actually are.

But the point of Decision Theory is to apply to the ‘decisions’ that you and I (and Suzie) actually face, *whether or not* those ‘decisions’ should prove on further investigation to have been free in the incompatibilist’s sense. If CDT falls short of that requirement then we should prefer a Decision Theory that does not.[[22]](#footnote-21)

***6: Conclusion***

So CDT is false. But where did it go wrong?

You are about to put money on the result the next time Nadal plays Federer on clay. But first you consult the oracle. The oracle tells you two things: (a) that Federer will win 95% of their future matches; and (b) that Nadal will win all of their future matches *on clay*. What should you do?

It would be a mistake to bet on Federer in the light of (a). That would be to base your decision upon a broader but *less* relevant reference class when a narrower but *more* relevant reference class is available. Instead, you should take (b) to be a decisive reason for betting on bet on Nadal.

The mistake that Causal Decision Theory makes is analogous to that of betting on Federer. It ranks A1 and A2 by their relative performances across the broader but less relevant reference class comprising some *neighbourhood* of A1-worlds and A2-worlds. It does not rank them by their relative performance across the narrower but more relevant reference class comprising just *one* world, the *actual* world. So it is bound to get things wrong in a case like Suzie’s, where A1 certainly outperforms A2 across the *broader* class but A2 certainly outperforms A1 across the *narrower* one.

References

Bar-Hillel, M. and A. Margalit. 1972. Newcomb’s paradox revisited. *British Journal for the Philosophy of Science* 23: 295-304.

Bennett, J. 2003. *A Philosophical Guide to Conditionals*. Oxford: OUP.

Edgington, D. 2004. Counterfactuals and the benefit of hindsight. In P. Dowe and O. Noordhof (eds), *Cause and Chance: Causation in an Indeterministic World*: 12-27. London: Routledge.

Elga, A. 2000. Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science Supplementary Volume 68*: S313-S324.

Gibbard, A. and W. Harper 1978. Counterfactuals and two kinds of expected utility. In C. Hooker, J. Leach and E. McClennen, (eds.), *Foundations and Applications of Decision Theory*: 125-62. Dordrecht: Riedel.

Horgan, T. 1981. Counterfactuals and Newcomb’s problem. *Journal of Philosophy* 78: 331-55.

Jeffrey, R. 1983. *The Logic of Decision*. Second edition. Chicago: Chicago University Press.

Joyce, J. 1999. *Foundations of Causal Decision Theory*. Cambridge: CUP.

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

———. 1979a. Counterfactual dependence and time’s arrow. Nous 13: 455-76. Reprinted in his *Philosophical Papers Volume II* (1986). Oxford: OUP.

———. 1979b. Prisoners’ dilemma is a Newcomb problem. *Philosophy and public Affairs* 8: 235-40. Reprinted in his *Philosophical Papers Volume II* (1986). Oxford: OUP.

———. 1980. A subjectivist’s guide to objective chance. In R. Jeffrey, ed., Studies in Inductive Logic and Probability, Volume II: 263-93. Reprinted in his *Philosophical Papers Volume II* (1986). Oxford: OUP.

———. 1981a. Are we free to break the laws? *Theoria* 47: 113-21. Reprinted in his *Philosophical Papers Volume II* (1986). Oxford: OUP.

———. 1981b. Causal decision theory. *Australasian Journal of Philosophy* 59: 5-30. Reprinted in his *Philosophical Papers Volume II* (1986). Oxford: OUP.

———. 1986. Events. In his *Philosophical Papers Volume II*: 241-269. Oxford: OUP.

McDermott, M. Counterfactuals and access points. *Mind* 108: 291-334.

Menzies, P. and H. Price 1993. Causation as a secondary quality. *British Journal for the Philosophy of Science* 44: 187-203.

Price, H. 1996a. *Time’s Arrow and Archimedes’ Point: New Directions for the Physics of Time*. Oxford: OUP.

———. 1996b. A neglected route to realism about quantum mechanics. *Mind* 103: 303-36.

——— and B. Weslake. 2009. The time-asymmetry of causation. In H. Beebee, C. Hitchcock and P. Menzies (eds), *The Oxford Handbook of Causation*: 414-43. Oxford: OUP.

Skyrms, B. 1980. *Causal Necessity*. New Haven: Yale UP.

Sobel, J. 1988. Infallible predictors. *Philosophical Review* 97: 3-24.

———. 1989. Partition theorems for causal decision theories. *Philosophy of Science* 56: 71-93.

1. Cf. Lewis 1981a: 291-2. Lewis’s own example concerns what the agent *did*, not what the agent *will* do. But I cannot see that I am straying from Lewis’s reading of soft determinism. If I am, then Lewis must be holding that determinism says that past acts of mine *were* determined by events in their pasts but that future and present acts *are* not, or *will* not be. I am not even sure that that position is diachronically consistent. Consistent or not it makes the future undetermined, so it doesn’t deserve to be called a form of determinism, so it can’t be what Lewis had in mind. [↑](#footnote-ref-0)
2. Lewis 1986: 263. Lewis notes (*ibid*. n14) that here ‘we are dealing with a substantial circle of interdefinables, and so have a choice of alternative primitives’. [↑](#footnote-ref-1)
3. Lewis 1979a: 55. Furthermore: (a) ‘@’ and ‘actually’ are my interpolations. I am assuming that the result is still faithful to Lewis. (b) (3) is actually Lewis’s definition of *fundamental* laws. The other laws, the *derived* laws, are the other theorems of that optimal system of which the fundamental laws are axioms. Nothing hangs on this: soft determinism about fundamental laws is equivalent to soft determinism about (fundamental laws + derived laws). (c) See Lewis 1980: 123ff. for more on ‘simplicity and strength’. [↑](#footnote-ref-2)
4. Gibbard and Harper 1978; Skyrms 1980; Lewis 1981b; Sobel 1989; Joyce 1999. [↑](#footnote-ref-3)
5. For instance Lewis’s theory in effect modifies the Gibbard-Harper theory by invoking, not counterfactual dependences of *states of nature* on acts, but counterfactual dependences of *chance distributions over states of nature* on acts. This certainly makes a difference to the theory. But it doesn’t make a difference to the theory’s advice to Suzie. See Lewis 1981b: 325-335. [↑](#footnote-ref-4)
6. Lewis 1979a: 34-5. [↑](#footnote-ref-5)
7. Lewis 1973: 13-19. [↑](#footnote-ref-6)
8. Lewis 1979a: 47-48 [↑](#footnote-ref-7)
9. See for instance McDermott 1999: 301-5 and Bennett 2003: 296-8. [↑](#footnote-ref-8)
10. For one argument that (13) does not follow from Lewis’s official account of closeness see Elga 2000. For an argument that Lewis’s criteria make (13) *fail* at the micro-level see Price 1996a: 146-52. [↑](#footnote-ref-9)
11. At 1979a: 35 Lewis lists circumstances that he considers exceptional. [↑](#footnote-ref-10)
12. Lewis 1979a: 44. Edgington (2004: 21) implicitly agrees with this much of Lewis’s theory. So does Bennett, who regards (13) as belonging to one amongst more than one equally acceptable ways to sharpen the vague thought behind an ordinary speaker’s counterfactual (2003: 218). [↑](#footnote-ref-11)
13. Obviously this crude equation of counterfactual and causal dependence ignores pre-emption. Let A = Assassin shoots and let B = Victim dies. Assassin knows that his equally proficient backup will shoot if and only if Assassin doesn’t. So Assassin has Cr (A → B) = Cr (~A → B). But intuitively Assassin’s shot is causally relevant to Victim’s death. However *that* sort of causal dependence—the sort that cuts finer than counterfactual dependence—is arguably of a kind that decision theory *should* ignore. When comparing shooting with not shooting for the purposes of deciding which to do, Assassin *should* be comparing scenarios that agree on Victim’s death, and the counterfactual independence of the death on the shooting is enough to ensure this.

    So counterfactual dependence doesn’t quite encompass causal dependence. But it *does* encompass as much of causal dependence as deliberators ought to care about. [↑](#footnote-ref-12)
14. It is easy to prove this for the Gibbard-Harper Decision Theory.

    A1 strictly dominates A2 (Supposition)

    The columns are certainly causally independent of the rows (Supposition)

    V (A1 ∧ Nj) > V (A2 ∧ Nj) for each j (from (i))

    Cr (A1 → Nj) = Cr (A2 → Nj) for each j (from (ii))

    Σj Cr (A1 → Nj). V(A1 ∧ Nj) > Σj Cr (A1 → Nj).V (A2 ∧ Nj) (from (iii))

    Σj Cr (A1 → Nj). V(A1 ∧ Nj) > Σj Cr (A2 → Nj).V (A2 ∧ Nj) (from (iv), (v))

    U (A1) > U (A2) (from (vi), (12)) [↑](#footnote-ref-13)
15. Of course Causal Decision Theory would *not* conform to this version of the dominance principle, and it would *not* recommend A1, if the counterfactuals in my definition (12) of U were interpreted as ‘backtracking’ counterfactuals (as in Horgan 1981). But then it would not deserve the name of *Causal* Decision Theory. See Lewis 1981b: 328. [↑](#footnote-ref-14)
16. Menzies and Price 1993 is a sophisticated recent exposition of the theory. [↑](#footnote-ref-15)
17. See Price 1996b. [↑](#footnote-ref-16)
18. Nozick 1969. [↑](#footnote-ref-17)
19. E.g. Lewis (1979b: 303-4) and Sobel (1988: 105-7). [↑](#footnote-ref-18)
20. Both the term ‘space cadet’ and the point that this sentence makes are due to Richard Jeffrey (1983: 25). [↑](#footnote-ref-19)
21. More generally, all that we need for the problem to arise is that the laws of nature dictate Cr (A1 ↔ Q) = *n* and Cr (A1 ↔ ¬Q) = 1 – *n*, where 0 < *n* < 1 – *n* < 1. Then consider following payoff arrangement:

    |  |  |  |
    | --- | --- | --- |
    |  | Q | ¬Q |
    | A1 | $ *x* | -$1 |
    | A2 | $1 | -$ *x* |

    CDT will prefer A1 to A2 whenever *x* > 1; but rationality prefers A2 to A1 whenever *x* > (1 – *n*)/*n*. So CDT gets it wrong for any *x* such that (1 – *n*)/*n* > *x* > 1. But we are assuming that 1 – *n* > *n*; hence such an *x* is guaranteed to exist. [↑](#footnote-ref-20)
22. Note that there *are* other Decision Theories that meet this requirement. In particular Evidential Decision Theory *without* ratificationism (EDT) meets it. EDT is also immune to the argument of this paper because it recommends A2 to Suzie. For expositions of EDT and defences of it in connection with Newcomb’s problem, see Bar-Hillel and Margalit 1972, and Price and Weslake 2009 (section 4). [↑](#footnote-ref-21)