Determinism, Counterfactuals, and Decision

Alexander Sandgren & Timothy Luke Williamson


To link to this article: https://doi.org/10.1080/00048402.2020.1764073

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 11 Jun 2020.

Submit your article to this journal

Article views: 1518

View related articles

View Crossmark data
Determinism, Counterfactuals, and Decision

Alexander Sandgren a and Timothy Luke Williamson b

aUmeå University; bAustralian National University

ABSTRACT
Rational agents face choices, even when taking seriously the possibility of determinism. Rational agents also follow the advice of Causal Decision Theory (CDT). Although many take these claims to be well-motivated, there is growing pressure to reject one of them, as CDT seems to go badly wrong in some deterministic cases. We argue that deterministic cases do not undermine a counterfactual model of rational deliberation, which is characteristic of CDT. Rather, they force us to distinguish between counterfactuals that are relevant and ones that are irrelevant for the purposes of deliberation. We incorporate this distinction into decision theory to develop ‘Selective Causal Decision Theory’, which delivers the correct recommendations in deterministic cases while respecting the key motivations behind CDT.

ARTICLE HISTORY
Received 12 September 2019; Revised 23 April 2020

KEYWORDS
causal decision theory; counterfactuals; determinism; compatibilism

1. Introduction
Determinism raises numerous philosophical challenges and, while the challenges for free will and responsibility have been discussed extensively, the challenges for rationality have received far less attention. What does rational decision-making look like when you believe that your choices might be predetermined? Our goal is to begin to answer that question by providing a decision theory that gives sensible advice to agents in deterministic cases (essentially, scenarios in which some states in the decision situation determine that some acts will not be performed). Our starting point is to take Causal Decision Theory (CDT) as a plausible normative theory. That might be surprising, since there has been growing concern that CDT delivers implausible verdicts in deterministic cases. Ahmed [2014a, 2014b], in particular, has argued that CDT is incompatible with determinism. We, however, think that causalism, or at least a generalisation of Lewis’s CDT [1981b], can be squared with belief in determinism.

We first outline the challenge for causalism: if you think that counterfactuals matter for rational choice, as many causalists do, then you go badly wrong by considering what would be the case were you to act in a way in which you are determined not to act. In response, we propose Selective Causal Decision Theory (SDT), which respects the motivations behind CDT while making well-motivated departures from it in deterministic...
cases. The heart of our proposal is that rational agents give no deliberative weight to outcomes that they take to involve a law-violation. This policy is readily formalised in the Lewisian framework for CDT, delivers sensible verdicts, and sheds light on compatibilism, especially the role of counterfactuals in rational deliberation.

2. Causal Decision Theory

CDT is an expected utility theory. The basic idea behind such theories is that each act that you might perform will bring about one of several possible outcomes, to which you assign utilities and probabilities. The expected utility of an act is the sum of the probability-weighted utilities of its possible outcomes, and you should act to maximise expected utility.

What marks out CDT is that the probability used in an outcome’s probability-weighting is a causal probability: it captures the degree to which the act is likely to cause the outcome. Here we work with Lewis’s [ibid.] version of this theory (some moves that we make therefore not be available to other causalists).

To begin, let \( A \) denote an act that is an option available to you. The outcomes are then the most fine-grained propositions that you care about (in particular, you are indifferent between the ways in which an outcome might be realised). Next, we define dependency hypotheses. This is the key innovation of Lewis’s view, and they will form an essential part of our proposal. Let a dependency hypothesis \( K \) be a maximally specific proposition about how outcomes depend on your acts.\(^1\) Following Lewis, we characterise dependency hypotheses as sets of counterfactuals, each specifying, for each option, which outcome would be brought about were that option taken. In this way, Lewis builds causal dependencies into state descriptions. Call the outcome caused by \( A \) in \( K \) \( o_{A,K} \). Let \( u \) be a utility function (a real-valued function representing preferences) and \( C \) be a credence function (a probability function representing degrees of belief). CDT says that you ought to maximise causal expected utility, defined thus:\(^2\)

\[
U(A) = \sum_i C(K_i) \cdot u(o_{A,K_i})
\]

Dependency hypotheses represent the various ways that your acts might cause outcomes, and so the Lewisian takes credences in dependency hypotheses as the right way of probability-weighting such that expected utility reflects causal efficacy.\(^3\)

CDT is typically motivated by Newcomb’s Problem [Nozick 1969]:

NEWCOMB’S PROBLEM. You are presented with two boxes. One box is transparent and contains $1,000 (henceforth \( k \)). The other box is opaque and contains either $0 or $1,000,000 (henceforth \( m \)). You can take either just the opaque box (‘One-boxing’) or both boxes (‘Two-boxing’). The prize in the opaque box is determined as follows: a predictor with a strong track record (say, 99%) yesterday placed \( m \) if they predicted that you Two-box and \( k \) if they predicted that you One-box.

---

\(^1\) This rules out more fine-grained states than required: dependency hypotheses capture just what is required to specify act-outcome dependencies.

\(^2\) Lewis allows that each option might bring about multiple outcomes in each dependency hypothesis. In that case, each counterfactual specifies the counterfactual chance of each outcome. We then replace \( u(o_{A,K}) \) with a function \( V \) measuring the value of performing \( A \) in \( K \) (\( V \) is the weighted average of the utilities of possible outcomes of \( A \) in \( K \)). Since the cases here involve each option causing a single outcome given each dependency hypothesis, we work with the slightly simplified version of the theory.

\(^3\) See Cartwright [1979: 435] for discussion of how partition-choice helps us to choose effective strategies.
Assuming that more money is preferred to less, the causalist recommends Two-boxing, by reasoning as follows: nothing that you now do affects the contents of the opaque box (the predictor’s decision was made yesterday), and so Two-boxing guarantees you an extra $1,000. You should therefore take the extra box: turning down $1,000 is a bad way of utility maximising!

CDT recommends Two-boxing. Let $K_1$ be the dependency hypothesis \{One-box $\rightarrow 0$, Two-box $\rightarrow m$\} and $K_2$ be the dependency hypothesis \{One-box $\rightarrow k$, Two-box $\rightarrow m + k$\}. We model this as follows:

<table>
<thead>
<tr>
<th></th>
<th>$K_1$</th>
<th>$K_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-box</td>
<td>0</td>
<td>$m$</td>
</tr>
<tr>
<td>Two-box</td>
<td>$k$</td>
<td>$m + k$</td>
</tr>
</tbody>
</table>

CDT recommends Two-boxing because each outcome associated with Two-boxing is better than the corresponding outcome associated with One-boxing. And CDT respects causal dominance: it never recommends an act that does worse than another in each dependency hypothesis. CDT respects this principle regardless of what evidence the dominated act provides. In NEWCOMB’S PROBLEM, Two-boxing provides strong evidence that the opaque box is empty, while One-boxing provides strong evidence that it contains a fortune. So, Two-boxers tend to walk away from NEWCOMB’S PROBLEM poor, while One-boxers walk away rich. But the causalist takes this to be a red herring, because you know that you will do better by Two-boxing.

NEWCOMB’S PROBLEM teaches us to causally promote the good. Lewis achieves this by calculating expected utility relative to dependency hypotheses. In this way, the Lewisian is committed to taking counterfactuals seriously in practical deliberation.

### 3. Counterfactuals and Deliberation

The above discussion highlights our commitment to causally promoting the good. To see the tension with determinism, consider the following. Simone is convinced that the universe follows deterministic laws. She is then asked to bet on the truth of some proposition $H$ about the past, to which she assigns credence .99. If $H$ is true and she bets, Simone wins a car; if $H$ is false and she bets, Simone loses $10. Being a genius, Simone knows that $H$, together with the laws of nature determines that she will not bet. What should Simone do? Clearly, she should not take the bet. After all, she knows that winning the bet would involve a law-violation, and she understands enough metaphysics to know that she is not free to break the laws.

This sketch of Simone’s situation needs to be filled out in several ways. But here we want to highlight that Simone will go badly wrong by employing a particular kind of counterfactual reasoning. Suppose that she reasons as follows:

If $H$ is true, then I would win a car if I bet. If $H$ is false, then I would lose $10 if I bet. I am very confident that $H$ is true and that I cannot affect $H$, so the small probability of losing a measly $10 is offset by the high probability of gaining a car, were I to bet. Therefore, I should bet.

This is bad reasoning, since Simone can only lose $10 by betting. So, the above counterfactual reasoning is not a good guide to choice. But counterfactual reasoning lies at the
heart of causalism. Indeed, Lewisian CDT is really a way of formalising counterfactual reasoning in the expected utility framework! This will create serious problems for CDT. Our core claim is that Simone’s case does not show that counterfactual reasoning is irrational. Rather, her case shows that there are particular kinds of irrational counterfactual reasoning. Cases like NEWCOMB’S PROBLEM convince us that counterfactuals are indispensable: you need to think about what each act would achieve, not simply to seek good news. What we need, then, is a principled reason for Simone to avoid the counterfactual reasoning that led her to bet on H, while allowing her to reason counterfactually in NEWCOMB’S PROBLEM.

We now diagnose the fault in Simone’s reasoning and provide such a principled reason. She is very confident of the following counterfactual: ‘if I were to bet, I would win a car.’ But that counterfactual is irrelevant to what she should do. Why? Because she knows that, if H is true, the laws determine that she does not bet. So, worlds in which she wins her bet must have different laws to ours. It is therefore futile for Simone to bet on H in the hope of bringing about a state in which H is true and she wins her bet; that would be to act wishfully in hope of changing the laws. (At the very least, it would be to act in a way that can only yield good results given different laws; we are happy to call that wishful.) Since H is a proposition about the past that determines that Simone will not bet, and she sensibly believes that she cannot change the past, Simone knows that she cannot lawfully win her bet. So, even if it is true that Simone would win her bet if she took it, the truth of H means that she should not take that counterfactual as a reason to bet—that would be wishful. Simone should therefore set aside the possibility of winning a car: from the deliberative perspective, she can either lose $10 or walk away.

So, a counterfactual can be true but irrelevant for the purposes of deliberation when you are certain that making its consequent true given its antecedent would violate the actual laws. Simone’s mistake was in taking an irrelevant counterfactual to be relevant.4

More generally, the fact that doing A would bring about O is relevant only in so far as that fact tells us that A is a good means of bringing about the end O. That is just to say that we care about acts instrumentally. And if you are certain that facts outside your influence make it the case that the actual world is not an (A ∧ O)-world, then it is futile to attempt to cause O by doing A. The mere fact that some nearby worlds are (A ∧ O)-worlds then provides you with no reason for or against doing A, which means that the truth of the counterfactual A ⊢ O provides you with no reason for or against doing A. Certain facts about the actual world trump facts about merely possible worlds, and so not every counterfactual is reason-providing.

Nothing here implies that there is anything wrong with counterfactual deliberation per se. You can ask what would happen if you bet on a roulette wheel (or Two-boxed), without worrying as to what you are determined to bring about. (At least, those with

---

4 We have refrained from commenting here on Simone’s abilities. Many compatibilists will say that Simone is able to bet despite being determined not to do so (e.g. Lewis [1981a] and List [2014]; more generally, anyone who thinks that Simone’s abilities are grounded in certain modal facts, such as facts about her dispositions or what she would do if she chose differently, will probably say that Simone is able to bet). We are happy to accept that. Regardless of what she is able to do, we simply require that Simone should give no deliberative weight to outcomes that would certainly violate the actual laws. Even if there are senses in which she can act differently, and those senses are useful in analysing, say, when she is morally responsible, we only insist that those facts should not lead to the absurd verdict that Simone should bet. That is, we want to distinguish between Simone’s being able to act in a way that would involve a law-violation, and its being rational for her to act in order to violate the laws. Thanks to Daniel Nolan for pushing us to clarify our position here.
broadly compatibilist commitments should think so.) We do not ordinarily treat the outcome of a roulette wheel as incompatible with any of your options. So, Simone’s case helps us to identify which counterfactuals guide action; it does not show that no counterfactuals do so.

Our goal is to formalise this reasoning and to develop a decision theory that gives Simone sensible advice. We spell out problem cases shortly, and explain formally why CDT goes wrong, but first we present our positive proposal. The relevant/irrelevant counterfactual distinction can be incorporated into CDT, meaning that we can uphold the policy of causally promoting the good without running into problems with determinism.5

4. Selective Causal Decision Theory

We propose a three-step theory—Selective Causal Decision Theory. Instead of simply calculating causal expected utility, you first restrict the set of outcomes that are to be given weight in deliberation (Step 1), after which you adjust your credence in the remaining outcomes (Step 2), and then calculate causal expected utility relative to those adjusted credences (Step 3). Steps 1 and 2 achieve what we proposed in section 3, and ensure that only relevant counterfactuals play an action-guiding role.

(1) Identify act-state combinations not worth taking seriously for the purposes of deliberation (in decision matrices, we will ‘grey-out’ the outcomes associated with such act-state combinations).

(2) For option A, let $D_A$ be the disjunction of dependency hypotheses $K$ such that $A \land K$ is not worth taking seriously for the purposes of deliberation. Define the renormalized credence for $A$, denoted $C_{RA}(\cdot)$ as this: $C_{RA}(K_i) = C(K_i | \sim D_A)$.

(3) You may do $A$ just in case there is at least one $K$ such that $A \land K$ is worth taking seriously for the purposes of deliberation and $A$ has maximal renormalized causal expected utility, denoted $U_R$, calculated thus:6

$$U_R(A) = \sum_i C_{RA}(K_i) \cdot u(o_{A,K_i})$$

Since dependency hypotheses are sets of counterfactuals, Step 2 amounts to disregarding irrelevant counterfactuals. Step 3 is an expected utility calculation. Step 1, however, needs to be spelled out more precisely. Why are some act-state combinations not worth taking seriously? One reason is familiar from Simone’s case: you should give no weight to outcomes that could only be brought about with a law-violation. So, you should give no weight to $A \land K$ if you think that $A$ involves a law-violation given the truth of $K$.

Note an important distinction here. Agents who believe in the truth of determinism

---

5 Some causalists have suggested that Simone should simply bet (call these die-hard causalists). What can we say to the die-hards? (1) It is a counterintuitive position: in so far as we are engaged in something like a process of reflective equilibrium, our theory seems to do better at tying together our theoretical commitments and considered judgments. (2) The die-hards must explain why Simone’s betting is not wishful. If she is not free to break the laws, then in what sense ought she to do something that can only yield good results with a law-violation? (3) Even if the die-hards are right, then we have good news for squaring decision theory and determinism: CDT is the correct compatibilist theory! Nonetheless, we think that the considerations raised here make the die-hard position look unattractive.

6 Again, in chancy cases, replace $u(o_{A,K_i})$ with $V(K_i \land A)$. 
will typically think that some outcomes can only be brought about with a law-violation, although they will not know which outcomes these are. Step 1 changes nothing for those agents, and they should treat all outcomes seriously. Ordinary compatibilist reasoning following, say, Lewis [1981a] applies in ordinary cases. You consider the various outcomes that you might bring about, knowing that bringing about some of those outcomes would involve a law-violation. Only when you know that a particular outcome involves a law-violation should you cease to give that outcome weight.

Note also that an outcome’s involving a law-violation is a sufficient condition for giving it no weight, but might not be necessary. For instance, if you meet an oracle or time-traveller who tells you that you will not survive tomorrow’s battle, then you might consider disregarding outcomes where you survive the battle. Theological cases involving predetermination might be similar (if I know that God has foreordained that I do not win my bet, the fact that I win my bet in nearby worlds provides no reason to bet). There are probably other kinds of case that follow this pattern. We will not commit to saying that you must grey-out outcomes in cases involving oracles and divine foreordination. If you do not think that those cases require you to disregard outcomes, then SDT and CDT yield the same verdicts; if, however, you do think that you should disregard outcomes in these cases, then SDT can capture this fact. Having said that, we will continue to focus on law-violations for concreteness and simplicity.

SDT generalises CDT. In ordinary decisions where you take each outcome seriously, \( U_R(A) = U(A) \) for all \( A \). So, SDT and CDT coincide in ordinary cases, including the cases that typically motivate CDT. Causal expected utility matters insofar as you give weight to each outcome featured in the expected utility calculation. When that condition fails, SDT yields more sensible verdicts than CDT does.

5. Three Cases

5.1 Betting on the Past

Ahmed [2014a, 2014b] provides the following counterexample to CDT:

**BETTING ON THE PAST.** You must choose between two bets—\( A_1 \) and \( A_2 \). \( A_1 \) pays out \$10 if \( P \) and costs \$1 if \( \neg P \). \( A_2 \) pays out \$2 if \( P \) and costs \$10 if \( \neg P \). \( P \) is the proposition that the actual universe at some past time was in state \( H \) and that the laws are \( L \). You know that \( H \land L \) determines that you take \( A_2 \) and \( \neg(H \land L) \) determines that you take \( A_1 \).

<table>
<thead>
<tr>
<th>( P )</th>
<th>( \neg P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_1 )</td>
<td>10</td>
</tr>
<tr>
<td>( A_2 )</td>
<td>2</td>
</tr>
</tbody>
</table>

To see that this case is problematic for CDT, note that it is structurally analogous to NEWCOMB’S PROBLEM, and so CDT recommends \( A_1 \) for the same reasons why it recommends Two-boxing. In particular, the truth of \( P \) is causally independent of your choice and specifies act-outcome dependencies. So, \( \{P, \neg P\} \) are dependency hypotheses. By causal dominance then, \( U(A_1) > U(A_2) \). So, CDT recommends \( A_1 \), CDT recommends \( A_1 \), CDT recommends \( A_1 \),

---

7 Of course, Calvinists might have other reasons not to gamble.
8 We have not yet addressed what kind of epistemic position you must be in to disregard some outcome. Is certainty, or confidence, or knowledge of a law-violation required? For simplicity, we first assume that agents are certain about, and know, which outcomes involve law-violations; Section 8 considers loosening this assumption.
which is incorrect since the only ways in which $A_1$ can do better than $A_2$ involve the laws or the past being different.

SDT correctly recommends $A_2$. To see this, note that, if $P$ is actually true, then you are determined not to take $A_1$. Even though you know that if $P$ is true and you were to act differently, you would win $10, you also know that doing so would involve non-actual laws. So, even though $o_{A_1,P}$ and $o_{A_2,\neg P}$ would be brought about were you to act differently than determined, it is wishful to do so. Trying to win $10 by taking $A_1$ is a misguided attempt to causally promote the good, and SDT respects this:

$$UR(A_1) = C(\neg P|\neg P) \cdot 1 = 1$$
$$UR(A_2) = C(P|\neg P) \cdot 2 = 2$$

This highlights the flexibility of SDT. SDT allows you to ignore specific outcomes without giving credence 0 to the dependency hypothesis associated with that outcome. CDT, however, can only give 0 weight to an outcome if the associated dependency hypothesis gets credence 0 (standard CDT can only grey-out entire columns in a decision table). It is this flexibility that allows SDT to break the parity between NEWCOMB’S PROBLEM and BETTING ON THE PAST. CDT was forced to treat those two cases equivalently. But SDT distinguishes between them: the counterfactual ‘if I were to Two-box, I would do better’ is true and relevant, while the counterfactual ‘if I were to bet on $P$, I would do better’ is true but irrelevant. SDT uses this distinction in order to provide sensible advice.

5.2 Simone

Let’s return to Simone’s case:

SIMONE. Simone is offered a bet, which she can Accept or Refuse. If she accepts, she wins a car (valued at $10,000) if $H$ is true, but $-10 if $H$ is false. Simone knows that $H$ specifies a past state of the world that determines that she will not bet.

<table>
<thead>
<tr>
<th></th>
<th>$H$</th>
<th>$\neg H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept</td>
<td>10,000</td>
<td>-10</td>
</tr>
<tr>
<td>Refuse</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

SDT recommends that Simone refuses:

$$UR(A_{Accept}) = C(\neg H|\neg H) \cdot -10 = -10$$
$$UR(A_{Refuse}) = 0$$

This contrasts with standard CDT, which says that Simone’s high credence in $H$ means that she should accept because of her high unconditional credence in $H$.

---

9 To see this, note that $P$ refers to the truth of $H \wedge L$ at the actual world, which is taken to be a rigid designator (we could clarify this by writing $(H \wedge L)_{\eta_0}$). Even if you were to act such that $H$ or $L$ were false, $(H \wedge L)_{\eta_0}$ remains true. Therefore, you cannot make a difference to $P$. See Ahmed [2014a: 669–71] for discussion.

10 Again, note that $H$ specifies act-outcome dependencies and is not causally affected by acts, and so $(H, \neg H)$ are dependency hypotheses.

11 A complication: Simone’s choice provides evidence about $H$, so she faces a case of decision instability (if she accepts, the expected utility of accepting is lower than refusing; if she refuses, the expected utility of refusing...
5.3 Uncertain of Determinism

SDT also works for agents who give some credence to determinism being false. Consider: this case

UNCERTAIN OF DETERMINISM. You must choose between two bets, $A_1$ and $A_2$. The dependency hypotheses are $P$ (which determines that you take $A_2$), $Q$ (which determines that you take $A_1$), and $R$ (which does not determine anything). The payoffs are:

<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>10</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>$A_2$</td>
<td>2</td>
<td>-10</td>
<td>-10</td>
</tr>
</tbody>
</table>

SDT says the following:

$$U_{R}(A_1) = C(Q|\neg P) \cdot 1 + C(R|\neg P) \cdot 10$$

$$U_{R}(A_2) = C(P|\neg Q) \cdot 2 + C(R|\neg Q) \cdot -10$$

This illustrates another advantage of SDT: its advice here is sensitive to your beliefs about $P$ and $Q$. CDT, however, will recommend $A_1$ regardless of your credences in $P$ and $Q$. But that is wrong. For example, if $C(P|\neg Q) \approx 1$, then recommending $A_1$ in an attempt to bring about the impossible $A_1 \land P$ is misguided. SDT gives advice that is appropriately sensitive to your credences in these more complicated deterministic scenarios.

6. The No Decision Response

In so far as they have addressed them, causalists have typically argued that deterministic cases are not genuine decisions [Sobel 1988: 6–10; Joyce 2016]. Proponents of this response argue that there is something incoherent in applying decision theory to deterministic cases. Sobel [1988: 6–7], for example, argues thus:12

the agent … cannot consistently so much as even think that both actions are open … He must, if consistent think that only one is (though, if he is consistent, he cannot be sure which one that is, unless and until he is sure what he is going to do).

If this is right, then you cannot think that both options are open in BETTING ON THE PAST, and so it is a non-decision. CDT therefore does not give the wrong recommendation in BETTING ON THE PAST, since decision theory only applies in genuine decisions. With no decision to be faced, there are no facts about what you ought to do. This ‘No Decision’ response vindicates CDT, provided that deterministic cases really do fall outside the purview of decision theory.

The first thing to say is that, as stated, this argument relies on a strong form of incompatibilism. Sobel takes your choice’s being predetermined to imply that you cannot think that more than a single option is open. But if that reasoning holds in BETTING ON THE PAST, then it seems to hold whenever determinism is true.

---

12 Although Sobel discusses predictors incapable of error, his comments apply equally to deterministic cases.
Those convinced of determinism never think that more than a single option is compatible with the past and laws of nature. Requiring multiple open options, in Sobel’s sense, threatens to undermine basic compatibilism. Indeed, Sobel says that treating yourself as free in infallible-predictor Newcomb-cases involves a contradiction: on the one hand, by treating yourself as free, you commit yourself to the idea that ‘a false prediction on the predictor’s part [is] itself entertainable, and at least in this sense a possibility’ while maintaining ‘that a false prediction on the predictor’s part is not a possibility’ [1988: 6]. But if determinism is true, then there is always, in a sense, an infallible predictor in the background, and so it looks like agents face a contradiction whenever they treat themselves as free.

Indeed, the compatibilist should simply reject Sobel’s reasoning. Sobel claims that in deterministic cases you ‘cannot consistently so much as even think that both actions are open’. But denying such claims is stock-in-trade for compatibilists! As already mentioned, we want to allow that there is a real sense in which you can do multiple things even if, in fact, there is one thing that you are determined to do. The compatibilist already thinks that determinism is compatible with your having multiple open options, and it is unclear why they should deny that claim in BETTING ON THE PAST. Conversely, if the causalist wants to utilise Sobel’s response, then they must deny this claim. Surely that would get the direction of argument wrong. SDT, however, requires no substantive metaphysical commitments in order to get the right verdict in NEWCOMB’S PROBLEM.

Joyce defends the No Decision response, although he does so in a way that avoids implicit incompatibilism. He [2016: 226] makes a similar claim to Sobel’s:

If Alice faces [a deterministic case] then, whatever she does, she could not have done otherwise, and perforce, could not have done better. So, [deterministic cases are] a wash when it comes to questions about what Alice should do.

But Joyce frames things more carefully. He also states that ‘an agent who deliberates about a decision which is framed so that each state entails a single act (and outcome) is engaging in an epistemic exercise, not an agential one’ ([ibid.]; first emphasis ours). The thought is that agents face free choices relative to the framing of a decision situation. Joyce can maintain that in most situations you should not be using states rich enough to determine which act you take. For example, when deciding whether to carry an umbrella, you could calculate expected utility relative to states like ‘Rain and the laws determine that you take an umbrella’, but you would not face a genuine choice relative to that framing of the decision. On the other hand, you would face a genuine decision relative to coarser partitions (like {Rain, No Rain}). Of course, you still do not face a genuine decision in BETTING ON THE PAST: the value of outcomes there depends on the truth of deterministic hypotheses, so the states must include deterministic information. In this way, Joyce can treat BETTING ON THE PAST as a non-decision without collapsing every choice into a non-decision.

We think that there are good reasons to prefer SDT to this sophisticated No Decision response.

First, even the sophisticated response is incompatible with a kind of compatibilism that we find independently plausible. Joyce is a compatibilist in the sense that determinism is compatible with genuine decisions, provided that the value of outcomes does not depend on the truth of propositions that determine your choice. But why would your
caring about the truth of a deterministic hypothesis affect whether you face a decision? When the compatibilist treats themselves as facing decisions in ordinary cases, they foreclose the possibility of making a choice but subsequently pleading ‘Well, I didn’t really choose that; the universe chose, and I merely discovered what the universe decided.’ Why should you then adopt that epistemic stance in cases like BETTING ON THE PAST? Although you are determined to take $A_1$ or $A_2$, you do not know which, and so you can coherently treat both options as open and as the objects of deliberation. It seems plausible, and in the spirit of compatibilism, that you can take this agential stance while caring about the truth of deterministic theses. SDT allows for this, which is an advantage.

Second, Joyce’s response involves denying what Ahmed calls ‘Soft Determinism’ [2014a: 667–9]. This is the view that (a) determinism is true, and (b) decision theory should tell rational agents what to do [ibid.: 667]:

The point of decision theory is to apply to the ‘decisions’ that you … actually face, whether or not those ‘decisions’ should prove on further investigation to have been free in the incompatibilist’s sense.

We agree with Ahmed: there are better and worse courses of action in deterministic cases, and we want decision theory to rank better courses of action over worse ones. More complicated cases highlight this problem for the No Decision response:

BETTING ON THE PAST 2. You must choose between two bets—$A_1$ and $A_2$. The dependency hypotheses are $P$ (which determines $A_2$), $Q$ (which determines $A_1$), and $R$ (which determines $A_1$). The payoffs are these:

<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>0</td>
<td>−100</td>
<td>200</td>
</tr>
<tr>
<td>$A_2$</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The No Decision response treats this as a non-decision. But surely we can distinguish between the advice that we would give to Barbara (who is practically certain of $Q$) and to Ian (who is practically certain of $R$). If we said to both Ian and Barbara ‘Do what you will, the correct decision theory can say no more’, they would rightly point out that such advice is insensitive to their disagreement about the world. Because of that disagreement, Barbara should take $A_2$ and Ian should take $A_1$. It is true that there are outcomes that agents should set aside as they deliberate, but this does not mean that we cannot provide them with guidance.

Finally, the No Decision response gets things wrong in the following case:

PARTIAL DETERMINATION. You must choose between $A_1$, $A_2$, and $A_3$. The dependency hypotheses are $P$ (which determines $\neg A_1$), $Q$ (which determines $\neg A_2$), and $R$ (which determines $\neg A_3$). The payoffs are these:

<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A_2$</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>$A_3$</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

SDT says that $U_R(A_2) = U_R(A_3) = 10$ while $U_R(A_1) = 0$. Therefore, we say that both $A_2$ and $A_3$ are permissible.
Saying that this is a non-decision is incorrect. Even by Joyce’s and Sobel’s lights, each state allows for a genuine choice. It would therefore be incorrect to say that there are no facts about what you should do in this case. But it would also be misguided to give weight to every outcome represented since that would lead to taking \( A_1 \) (given sufficiently high credence in \( P \)). And taking \( A_1 \) is absurd, no matter how confident you are in \( P \).\(^{13}\) The only sensible policy is to calculate expected utility based on just the non-greyed-out outcomes. The No Decision response is too coarse-grained to handle intermediate cases like PARTIAL DETERMINATION.

One possible move on behalf of the No Decision response is to say that, while you do face a choice in PARTIAL DETERMINATION, you only find out which choice that is after learning which dependency hypothesis is true. For example, if \( P \) is true, then you faced a choice between \( A_2 \) and \( A_3 \) all along. But this yields no more action-guiding advice than just calling the case a non-decision. After all, you cannot tell which options you face a choice between while deliberating. But then you cannot know which outcomes to exclude from deliberation, leaving us with the initial dilemma: either provide no advice or provide advice based on all outcomes. Neither option is satisfactory in PARTIAL DETERMINATION.

Although we reject the No Decision response, it is worth highlighting an important methodological agreement between our position and Joyce’s version of that response: we think that the framing of a decision problem matters. Our disagreement concerns how it matters. While Joyce thinks that building deterministic information into state-descriptions changes whether you face a decision, we think that it changes the nature of your decision (by affecting which outcomes are worth taking seriously). So, in cases like SIMONE where deterministic information must be built into state descriptions, the correct response is to restrict the set of outcomes worth taking seriously, rather than to give up on agency altogether. Nonetheless, we agree that the carving up of states plays an important role in handling deterministic cases, which is reflected in the fact that we calculate expected utility relative to dependency hypotheses.

7. Avoiding Evidentialism

Next, we want to consider whether SDT deserves to be called a causal decision theory. You might think that, in Step 1 of SDT, we are covertly appealing to the evidence that an act provides to evaluate that act. That is, we grey-out \( o_{A,K} \) because \( A \) provides evidence against \( K \) (albeit evidence of the particularly strong kind that \( K \) is impossible). This would count as evaluating acts based on their news-value. But the hallmark of causalism is that acts are not evaluated based on their news-value: when news-value and causal efficacy diverge, it is causal efficacy that determines what you should do. So, if SDT is taking account of news-value, that is an ad hoc compromise.

\(^{13}\) Although this case again involves instability, we can still cause problems for deliberative versions of CDT. Say that you have initial credences: \( C(P) = C(Q) = C(R) = \frac{1}{3} \), \( C(P|A_1) = C(Q|A_2) = C(R|A_3) = 0 \), and \( C(Q|A_1) = C(R|A_1) = C(P|A_2) = C(R|A_2) = C(P|A_3) = C(Q|A_3) = \frac{1}{2} \) (i.e. each act makes certain that the incompatible dependency hypothesis does not hold, and leaves the remaining two hypotheses as equally likely). Then, using Joyce’s [2012] dynamics, you have the equilibrium: \( U(A_1) = U(A_2) = U(A_3) = \frac{100}{9} \), with \( C(P) = \frac{1}{19} \), \( C(Q) = C(R) = \frac{9}{19} \), \( C(A_1) = \frac{17}{19} \) and \( C(A_2) = C(A_3) = \frac{12}{19} \). Either \( A_1 \) is permissible or you ought to perform a mixed act weighted towards \( A_1 \).

ALEXANDER SANDGREN AND TIMOTHY LUKE WILLIAMSON
It is important to stress the rationale behind Step 1 of SDT. We do not grey-out $o_{A,K}$ because of your conditional credence $C(o_{A,K} | A) = 0$. That would count as ad hoc evidentialism. Instead, we grey-out outcomes because it is futile to attempt to bring them about—doing so would involve a law-violation. Certain facts about the actual world play a structuring role in our theory. To take the possibility of determinism seriously is to give up on being free to break the laws, which means that your beliefs about what is nomologically possible form fixed points in your deliberation. Of course, when you know what the laws determine, your acts may have news-value as well. But this does not mean that we disregard outcomes because of that news-value. In BETTING ON THE PAST, for example, you disregard the possibility of winning $10 because you respect the laws: knowing that you are determined not to win $10, you structure your deliberation around that fixed point. Step 1 of SDT does not appeal to your conditional views; rather, it appeals to your unconditional views about what is nomologically possible. And that is importantly different to caring about news-value.

This defuses a potential objection from Ahmed. While discussing BETTING ON THE PAST, he claims [2014a: 678–9] that any theory deserving to be called ‘causal’ is forced to give some weight to greyed-out outcomes. If this is correct, our theory no longer counts as causal. Why does Ahmed think that the causalist is committed to giving non-zero weight to greyed-out outcomes? It is because the hallmark of CDT is the use of counterfactual reasoning, and counterfactual reasoning involves thinking about possible worlds that certainly differ from our own. In BETTING ON THE PAST then, the causalist is supposedly forced to ask ‘Given $P$, what would the world be like if I took $A_1$?’, even though they know that $P$ makes the actual world one in which they do not take $A_1$. Ahmed concludes [ibid.: 679]:

CDT regards 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.

Ahmed is right that unadorned CDT regards certainly non-actual worlds as open to free agents, but SDT makes no such requirement. This is the crux of our disagreement with Ahmed. He regards the kind of counterfactual just described as definitive of causalism. We think that a more nuanced position is required: counterfactual thinking is definitive of causalism at the level of expected utility calculations. The core of causalism is that expected utility is a matter of expected causal efficacy. But, before calculating expected utility, the causalist is entitled to use non-counterfactual reasoning to restrict the set of outcomes that go into the expected utility calculation. The actual world matters, even for the causalist! Proponents of SDT care about causation, although they deny that all outcomes are worth causally promoting.14

It is important not to miss the wood for the trees here: in all ordinary cases, SDT agrees with CDT (including in cases often taken as characteristic of CDT, such as

---

14 Ahmed [2015] raises a distinct worry for views like ours. He argues that every Newcomb case can be viewed as a weighted lottery between a certainly correct predictor and a certainly incorrect predictor. He argues that views like ours will then recommend One-boxing in every Newcomb case (since, when framed as a weighted lottery between certainly correct and incorrect predictors, we will have to grey-out outcomes such that SDT agrees with Evidential Decision Theory). This does not seem to be a problem for our view, since Ahmed’s argument relies on calculating expected utility relative to the partition: (Predictor Certainly Correct, Predictor Certainly Incorrect). But these states are not dependency hypotheses, and so we cannot use them in SDT’s expected utility calculation. This again highlights the importance of framing: when determinism is involved, not just any states will do.
NEWCOMB’S PROBLEM). In cases where SDT builds on CDT, it does so in a way that respects the causalist intuition that news-value is irrelevant to decision-making. SDT still deserves to be called causal.

7.1 A Slippery Slope towards Evidentialism?

You might worry that our view opens the door to evidentialist reasoning, even if it does not itself rely on evidentialist reasoning. In particular, you might worry about an objection raised by Seidenfeld [1984] (see also Sobel [1988]). Seidenfeld objects to theories that treat Newcomb cases in which $C(\text{Predictor Correct}) = 1$ differently to those in which $C(\text{Predictor Correct}) = 1 - \epsilon$ for any $\epsilon > 0$, since he thinks that such an $\epsilon$-decrease cannot make a difference to what you ought to do. Now, this objection does not directly target our view, since we grey-out based on your views about what is nomologically possible, not merely your confidence in the predictor’s accuracy. But a similar objection might arise: we distinguish between cases in which you are certain that some outcome involves a law-violation and cases in which you are merely confident. This raises a question that we have hitherto ignored: what kind of epistemic situation must you be in to disregard some outcome? We now turn to that point.

8. Analysing Futility

We say that you should disregard an outcome when it is futile to attempt to bring it about. But when are you entitled to do this?

The easiest cases are those in which you are rationally certain that some outcome involves a law-violation. If you are rationally certain that you cannot lawfully cause some outcome, then your deliberation should be structured around that fact.

But what about cases involving less-than-complete certainty? Consider the following:

BETTING ON THE PAST 3. You are choosing between two bets on $P—A_1$ and $A_2$. You are confident but not certain (your credence is .99) that $P$ determines that you will take $A_2$, and you are confident but not certain (your credence is .99) that $\neg P$ determines that you will take $A_1$.

<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$\neg P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>$A_2$</td>
<td>2</td>
<td>-10</td>
</tr>
</tbody>
</table>

Given that you are not certain that any outcome involves a law-violation, should you treat this case like BETTING ON THE PAST or NEWCOMB’S PROBLEM? That is an interesting question, and we want to remain broadly neutral, partly because we

15 There are various ways that this could be so: you could be .99 certain that some deterministic theory holds; or you could be certain that some system of laws holds that is deterministic apart from the occasional indeterministic event (so, though certain of the laws, you are only .99 certain of what they determine). Our proposed solutions treat these versions of the case the same. There is a related case—the one in which you know that $P$ determines that the chance of $A_2$ is .99. We will not settle what you should do in that case, since it is a case of thoroughly going indeterministic laws; such cases certainly raise challenges, although, as our focus is on determinism, they must be set aside for future work. Thanks to a referee for pushing us to clarify this point.
have no firm intuitions about this case, and partly because a full argument for either position would take us beyond the scope of this paper. Instead, we sketch three ways of analysing futility. SDT can incorporate any of these analyses and so deliver different verdicts in BETTING ON THE PAST 3.

First, we could adopt a strict analysis of futility. On this account, for an outcome to not be worth taking seriously, you must have credence 1 that it involves a law-violation. This means that we treat BETTING ON THE PAST 3 like an ordinary Newcomb case: every outcome is worth taking seriously, which means that you should take $A_1$.

Some might worry that the strict analysis is too strict. After all, it is extraordinarily unlikely that your doing better by taking $A_1$ is compatible with the laws. Does this mean that you ought not to take $A_1$?

We are not sure. In so far as there is an intuition that you ought not to take $A_1$, we are unsure as to how much weight to put on that intuition. And the proponent of the strict analysis can point to a difference between BETTING ON THE PAST and BETTING ON THE PAST 3: your winning $10 without violating the laws is possible in the former but not the latter. (This explains why an $\epsilon$-decrease in credence might be significant, contra Seidenfeld. A shift from $C(\neg X) = 1$ to $C(\neg X) = 1 - \epsilon$ can signal a shift from X’s being impossible to being possible.) True, it is unlikely that $P$ is true and that you take $A_1$ in BETTING ON THE PAST 3, but there is nothing incoherent involved in taking $A_1$ and winning $10$ (it is just unlikely). And causalists already think that we need to take into account unlikely act-state combinations when deciding between options (like Two-boxing when the predictor guessed One-box). In BETTING ON THE PAST, you are certain that you cannot lawfully win $10, and so the fact that you do not win $10 should act as a fixed point as you structure your deliberation. But you are not certain of that fact in BETTING ON THE PAST 3, and so it might make sense to treat winning $10$ as a live possibility. Causalists should not be misled by the fact that taking $A_1$ provides strong evidence against $P$; that is just the kind of news that the causalist sets aside in NEWCOMB’S PROBLEM, and they should set it aside here.

A second approach would be to adopt a threshold analysis of futility. For an outcome to not be worth taking seriously, we might insist only that you have high credence that it involves a law-violation.

The threshold view faces the standard worry that thresholds can look arbitrary. Say that we set the threshold at .99. Then, we might ask, what is the real difference between $C$(Law Violation) = .99 and $C$(Law Violation) = .9899? How could that miniscule decrease in confidence affect whether you take some option seriously? And why pick .99 in the first place?

At this point, the defender of the threshold analysis can make use of the moves that are made in response to Sorites sequences. We could say that our inability to work out where the threshold is does not show that there is no threshold. Or we could say that

---

16 What if some possible propositions get credence 0 (e.g. maybe you assign credence 0 to a fair coin landing ‘heads’ indefinitely)? Then we might want an even stricter strict analysis, one requiring something more than credence 0 for futility. For example, we might insist that $O_{AX}$’s lawful occurrence is doxastically impossible (there is no doxastically possible world at which $O_{AX}$ is true and the actual laws hold). This would mean not greying-out outcomes that might be compatible with the laws, even if you have credence 1 that they violate the laws (say, because there are worlds where the outcome occurs lawfully, although they are as unlikely as a fair coin landing ‘heads’ indefinitely). Thanks to a referee for pushing us to clarify this point.

17 We take Joyce [2016] to advocate a kind of threshold account, although his threshold condition differs slightly from ours and it is intended for when a situation counts as a decision.
there is a vague threshold. This strikes us as quite plausible. We can point to paradigm cases where some outcomes are futile (BETTING ON THE PAST), and we can point to paradigm cases where no outcomes are futile (NEWCOMB’S PROBLEM). Between those cases, there might be a range of indeterminate cases. If you are .98 confident that your winning big by Two-boxing involves a law-violation, then perhaps it is indeterminate whether you ought to Two-box. The concept of freedom is tricky, and it seems plausible that we might sometimes be neither determinately free nor determinately unfree to bring about some outcomes. Clearly, more needs to be said here. We simply want to point out that defenders of the threshold view will be able to draw on the tools developed elsewhere to help deal with problematic threshold concepts.

The final strategy that we consider is a knowledge-based analysis of futility: you should treat some outcome as not worth taking seriously when you know that it involves a law-violation. On this view, it is not partial belief but is instead knowledge that determines the fixed points around which deliberation should be structured.

Some might be uncomfortable with introducing a concept like knowledge into decision theory. But Weatherson [2012] argues that decision theorists cannot ignore knowledge. On his view, you know p if and only if it is legitimate to write p as an outcome in your decision table [ibid.: 77] (similarly, you know that a state does not obtain if and only if you can legitimately leave that state off the decision table). So, knowledge plays a crucial role in decision theory: given that agents like us are rarely certain, knowledge helps us to understand what goes into our decision tables in the first place. Now, Weatherson does not talk about the kinds of cases that we are considering here. But it seems natural to extend his account to supplement SDT; indeed, he is concerned with how decision problems should be structured, and we are arguing that one structuring principle is that you should disregard law-violating outcomes. So, we might suggest, you can legitimately grey-out an outcome if you know that the outcome would involve a violation of the actual laws. We will not try to give a further analysis of knowledge here, but note that this approach may take into account whether your beliefs are justified, how you came to have your beliefs, the stakes of the case, and so on.

The strict analysis will say that you ought only to disregard outcomes in BETTING ON THE PAST. The threshold analysis will say that you should disregard outcomes in both BETTING ON THE PAST and BETTING ON THE PAST 3 (given a choice of threshold below .99). The knowledge-based analysis will say that it depends on what you know in each case. We have argued that each of these strategies is plausible, although it is beyond the scope of this paper to argue that any one strategy is best. What matters is that, no matter which option is taken, SDT can be supplemented with a plausible account of futility.

9. Conclusion

We began by noting that determinism raises serious questions about rationality. By formulating a decision theory compatible with determinism, we have been able to shed light on many of those questions. One key lesson is that determinism does not rule

---

18 This suggests related approaches: belief approaches, justified-belief approaches, knowledge-of-a-certain-quality approaches (for those who think that knowledge can be graded: e.g. Hetherington [2001]), etc. Hopefully, it is clear how other concepts could be substituted into this analysis.
out a counterfactual model of deliberation, if we can distinguish between relevant and irrelevant counterfactuals. While CDT cannot capture this distinction, SDT can do so while respecting the motivation for CDT. Another key lesson is that we can settle decision-theoretic questions without being forced into making strong metaphysical assumptions. We can accept that Simone faces choices, has abilities, and so on, without abandoning a broadly causal decision theory. So, while determinism raises numerous challenges, and we do not pretend to have solved all of them, we have shown that rational agents can deliberate sensibly while taking seriously the possibility of determinism. And that is progress.

Disclosure Statement

No potential conflict of interest was reported by the authors.

ORCID

Alexander Sandgren http://orcid.org/0000-0001-5599-7598

Timothy Luke Williamson http://orcid.org/0000-0002-9920-602X

References


19 Not even the decision-theoretic challenges. Ahmed [2013] argues that causalists cannot endorse determinism, while cases involving foreknowledge of chancy processes raise separate challenges (see Rabinowicz [2009], Price [2012], and Bales [2016]). Future work is needed to show how SDT responds to those challenges.

20 Thanks to Edward Elliott, Torfinn Huvenes, Alan Hájek, Kalle Grill, Sofia Jeppsson, Boris Kment, Daniel Nolan, Wlodek Rabinowicz, Toby Solomon, Katie Steele, Jeremy Strasser, Caroline Torpe Touborg, James Willoughby, several anonymous referees, and audiences at an ANU Philsoc Seminar and the Swedish Congress of Philosophy, for invaluable help and suggestions.