Causal Decision Theory, Two-Boxing, and Deliberation-Compatibilism: a Reply to Sandgren and Williamson*

Toby Charles Penhallurick Solomon

Abstract

The possibility of predetermined choices raises a challenge for Causal Decision Theory [Ahmed 2014b]. Sandgren and Williamson [2021] have recently proposed a response—*Selective Causal Decision Theory*—they hope will avoid Ahmed's counterexamples, maintain (a particular kind of) compatibilism, and endorse universal Two-boxing in Newcomb's Problem—CDT's *raison d'etre*. Their proposal does an admirable job of satisfying the first two desiderata. However, in this reply I will raise several worries about whether it can satisfy the third.

1 Introduction

Causal decision theory (CDT) endorses Two-boxing in NEWCOMB'S PROBLEM (see below) because it follows *causal*, rather than evidential, dependence. One way to capture this in expected utility calculations is to weight utilities with the *unconditional* probability of each Dependency Hypothesis: "a maximally specific proposition about how outcomes depend [causally] on your acts" [Sandgren and Williamson 2021: 2]. Given an agent's credence function Cr, utility function U, and the partition $\mathbb{K}$ of Dependency Hypotheses, we have:

$$EU(A) = \sum_{K_i \in \mathbb{K}} Cr(K_i)U(A \wedge K_i)$$

Arif Ahmed [2014b] argues that cases like the following are counterexamples 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

[0]This is an Accepted Manuscript of an article accepted for publication by Taylor & Francis in the *Australasian Journal of Philosophy*, available online:
http://www.tandfonline.com/10.1080/00048402.2021.1968448

> H and that the laws [of nature] are L. You know that $H \wedge L$ determines that you take A_2 and $\neg(H \wedge L)$ determines that you take A_1 [Sandgren and Williamson 2020: 6]

Ahmed argues that it is never rational to take A_1, because it is certain that you will either take A_1 and lose \$1, or take A_2 and gain \$2. Winning \$10 on A_1, or losing \$10 on A_2, are *contradictions*. But, *prima facie*, CDT advises you to take A_1: it endorses the partition of Dependency Hypotheses represented in Table 1, relative to which A_1 dominates A_2.

	P	$\neg P$
A_1	\$10	−\$1
A_2	\$2	−\$10

Table 1

Sandgren and Williamson (S&W) [2021] have proposed a variation on CDT to avoid Ahmed's counterexamples: *Selective Causal Decision Theory* (SDT). Ahmed's counterexamples arise because standard, subjunctive, definitions of Dependency Hypotheses lend weight to outcomes that are "not worth taking seriously" [S&W 2021: 5]—such as those that require violation of the laws of nature (laws). SDT selectively rules out these outcomes in a three step process:

1. For each option, A, identify any outcomes that are not worth taking seriously—and call the disjunction of these D_A.

2. For each option, define a probability function, P_A, by conditionalising Cr on $\neg D_A$: $P_A(K_i) = Cr(K_i|\neg D_A)$. This has the effect of assigning zero probability to the outcomes not worth taking seriously and normalising the remaining probabilities.

3. Calculate the expected utility of A using P_A instead of Cr:

$$EU(A) = \sum_{K_i \in \mathbb{K}} P_A(K_i)U(A \wedge K_i) = \sum_{K_i \in \mathbb{K}} Cr(K_i|\neg D_A)U(A \wedge K_i)$$

SDT avoids endorsing A_1 because neither the outcome of A_1 if P—winning \$10—nor the outcome of A_2 if $\neg P$—losing \$10—are worth taking seriously. SDT therefore assigns zero probability to these outcomes (and normalises), giving $EU(A_1) = -\$1$ and $EU(A_2) = \$2$. Unfortunately, SDT will have a harder time securing Two-boxing in NEWCOMB'S PROBLEM—a non-negotiable for causalists.

I will proceed as follows: First I outline why SDT will sometimes endorse One-boxing in NEWCOMB'S PROBLEM. Then, section 3 considers a hypothetical response from S&W—supported by the text—and argue that it fails because it is doubly *ad hoc*. Finally, section 4 argues that even if SDT can secure Two-boxing in NEWCOMB'S PROBLEM, it cannot secure what causalists really care about: respect for the kind of reasoning that justifies Two-boxing.

2 SDT and Newcomb's Problem

NEWCOMB'S PROBLEM goes like this:

> **NEWCOMB'S PROBLEM**: You are presented with two boxes. One box is transparent and contains $1,000 (henceforth [$1k]). The other box is opaque and contains either $0 or $1,000,000 (henceforth [$1m]). 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 $0 if they predicted that you Two-box and [$1m] if they predicted that you One-box [S&W 2021: 2].

S&W analyse the case with the Dependency Hypotheses represented in Table 2, relative to which Two-boxing dominates One-boxing and is the rational choice.

	Two-boxing predicted = K_1	One-boxing predicted = K_2
One-box	\$0	$\$1m$
Two-box	$\$1k$	$\$1m + \$1k$

Table 2

Rational decision-makers should take into account all the ways the world might be; Table 2 does not do so. The outcome of your choice in NEWCOMB'S PROBLEM might depend not only on what the prediction is—which S&W's Table 2 captures—but also on if, and how, your choice is predetermined—which S&W's Table 2 ignores. Table 3 shows a partition of Dependency Hypotheses taking this dependence into account. The grey outcomes are not worth taking seriously because they require violations of the laws.[1] Note that Table 3 does not *add* columns to Table 2, but rather divides the columns in Table 2 into finer-grained possibilities.

[1] I have followed S&W in assuming that we can assign utilities to these inconsistent option-state pairs. Causalists who reject subjunctive reasoning for defining outcomes, or reject a Lewisian semantics for subjunctives, might argue that no such utility can be assigned. See section 2.2 of my [Solomon 2021] for more on this point. Even if so, SDT can deal with Table 3 since we must treat an undefined utility multiplied by zero probability as equal to zero for calculating expected utilities [Lewis 1981b: 14]. This would only make things worse for S&W below, so I will not comment on it further.

	Two-boxing predicted			One-boxing predicted		
	No predetermination $= K_{1.1}$	Predetermined to One-box $= K_{1.2}$	Predetermined to Two-box $= K_{1.3}$	No predetermination $= K_{2.1}$	Predetermined to One-box $= K_{2.2}$	Predetermined to Two-box $= K_{2.3}$
One-box	\$0	\$0	\$0	$\$1m$	$\$1m$	$\$1m$
Two-box	$\$1k$	$\$1k$	$\$1k$	$\$1m + \$1k$	$\$1m + \$1k$	$\$1m + \$1k$

Table 3

The problem, for S&W, is that with this partition SDT will recommend One-boxing whenever:[2]

$$\begin{aligned} &Cr(K_{2.1}|\neg(K_{1.3} \vee K_{2.3})) + Cr(K_{2.2}|\neg(K_{1.3} \vee K_{2.3})) \\ &\quad > (1 + \frac{1}{1000}) \times [Cr(K_{2.1}|\neg(K_{1.2} \vee K_{2.2})) + Cr(K_{2.3}|\neg(K_{1.2} \vee K_{2.2}))] \\ &\quad\quad + \frac{1}{1000} \times [Cr(K_{1.1})|\neg(K_{1.2} \vee K_{2.2})) + Cr(K_{2.1}|\neg(K_{1.2} \vee K_{2.2}))] \end{aligned}$$

Nothing in SDT rules out such credences. They might arise, for example, if you believe that (predetermined) evidentialists are harder to predict than (predetermined) causalists. We might have $Cr(K_{1.2}|\neg(K_{1.3} \vee K_{2.3})) = 0.2$ and $Cr(K_{2.2}|\neg(K_{1.3} \vee K_{2.3})) = 0.5$, while $Cr(K_{1.3}|\neg(K_{1.2} \vee K_{2.2})) = 0.6$ and $Cr(K_{2.3}|\neg(K_{1.2} \vee K_{2.2})) = 0.1$. If the rest of your credence is evenly spread—you have no more reason to believe the you will One-box than Two-box if your choice is not predetermined—then:

$$\begin{aligned} EU(\text{One-box}) &= (0.15) \times \$0 + (0.2) \times \$0 + (0.15) \times \$1m + (0.5) \times \$1m \\ &= 0.65 \times \$1m \\ &= \$650,000 \\ EU(\text{Two-box}) &= (0.15) \times \$1k + (0.6) \times \$1k + (0.15) \times (\$1m + \$1k) + (0.1) \times (\$1m + \$1k) \\ &= (0.75) \times \$1k + (0.25) \times (\$1m + \$1k) \\ &= \$251,000 \end{aligned}$$

SDT will sometimes endorse One-boxing in NEWCOMB'S PROBLEM if Table 3 is an acceptable partition of Dependency Hypotheses.

3 Is Table 3 Too Fine-Grained?

While S&W don't directly discuss Table 3, we can construct a more or less hypothetical response on their part, as follows. Table 3 requires us to distinguish between outcomes on

[2]This inequality is calculated simply by applying SDT to Table 3, grouping terms, and then dividing both sides by \$1m.

the basis of whether they are worth taking seriously or not. If instead outcomes are only unique up to utility values then the leftmost three columns in Table 3 really represent one and the same Dependency Hypothesis. And similarly for the rightmost three columns. This interpretation is suggested by S&W's definition of outcomes as:

> ...the most fine-grained propositions that you care about (in particular, you are indifferent between the ways in which an outcome might be realised) [S&W 2020: 2].

And when they tell us:

> 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 broadly compatibilist commitments should think so.) [S&W 2020: 4].

This would suggest that Table 2 captures the finest-grained Dependency Hypotheses rational decision-makers should care about.

However, this response leaves us wondering: what should rational decision-makers do when they are uncertain whether an outcome is worth taking seriously? Applying expected utility maximisation to uncertainty about whether outcomes in Table 2 are worth taking seriously will lead us straight back to Table 3. S&W tell us:

> 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 [S&W 2020: 6].

But the picture of NEWCOMB'S PROBLEM we are left with cannot be acceptable to decision theorists; it is doubly *ad hoc*. First, it requires agents to care about whether outcomes are worth taking seriously in BETTING ON THE PAST but does not allow them to do so in NEWCOMB'S PROBLEM. The idea that outcomes are unique only up to utility value is plausible on the standard assumption that everything an agent cares about is encoded in their utility function. But S&W have given up on that assumption by requiring that rational decision-makers take into account whether an outcome is worth taking seriously in BETTING ON THE PAST without reference to their utility function. There is no difference between the cases that could justify treating them differently.

Second, S&W's picture requires agents to deal with uncertainty about whether outcomes are worth taking seriously by appealing to knowledge, while they apply expected utility maximisation to all other uncertainty about matters of fact. But there is no difference between the kinds of facts involved that could justify treating them differently. In a slogan: rational decision making is expected utility maximisation all the way down, if it is expected utility maximisation at all.[3]

Now, S&W might suggest these distinctions are not *ad hoc* because:

[3]Or at least expected utility maximisation goes as far down as bounded rationality will allow us to go. But it is clear that we—actual human agents—can go as far as Table 3 (we just have!). So there is no objection on the grounds of bounded rationality here.

> ...we think that the framing of a decision problem matters... While [Joyce 2016] 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) [S&W 2020: 11].

But how can the *description* of a problem affect which outcomes are worth taking seriously, or whether we should apply expected utility maximisation to our uncertainty about that? Rational decision-makers are not limited to the state-descriptions that we write down in philosophy papers; a rational decision-maker facing NEWCOMB'S PROBLEM should take into account the possibility some outcomes are not worth taking seriously, as Table 3 does. The content of the problem, not the presentation, is what matters.

Finally, we need to ask whether S&W's appeal to compatibilism can provide the required motivation. Unfortunately, exactly how they intend compatibilism to motivate the required picture is somewhat opaque. S&W's most concrete statement of the kind of compatibilism they have in mind is a quote from Ahmed:

> 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 [Ahmed 2014a: 667].

This is true, but irrelevant: decision theory always assesses decisions against the agent's perspective, not how the actual world proves to be on further investigation. The relevant question is whether rational decision-making is compatible with non-zero credence that a choice is predetermined—that is, whether *deliberation-compatibilism* is true [Pereboom 2014: ch. 5].[4]

S&W [2020: sec. 6] are right to criticise *No Decision* responses that violate deliberation-compatibilism by ruling out rational decision-making whenever predetermination is a possibility. But it is not clear how an appeal to deliberation-compatibilism can support Table 2 over Table 3. SDT gives us verdicts either way; either way it applies to the decision that you actually face in NEWCOMB'S PROBLEM. Though we might say it applies more accurately in Table 3 because no information is ignored. Nor does the appeal to Lewis's [1981a] compatibilism help: Lewis is there providing an analysis of 'can'. But Table 3 doesn't rely on any specific claims about what an agent can do; only on claims about what agents care about and when they should use expected utility maximisation. Compatibilism cannot tell us whether to One-box or Two-box, and so it is unclear how it could directly support the picture of NEWCOMB'S PROBLEM that S&W require.

Perhaps, instead, S&W intend to motivate that picture more obliquely by arguing that combining such a picture with SDT is the only way to capture all three of their desiderata: avoiding Ahmed's counterexamples, allowing for deliberation-compatibilism, and securing Two-boxing in NEWCOMB'S PROBLEM. To show this, however, they would need to

[4]Pereboom [2014: 106] defines deliberation-compatibilism as the view that: rational decision-making (deliberation) is compatible with *belief* that one's actions are (causally) predetermined. Decision theory trades in credence—not belief—but if rational decision-making is compatible with any non-zero credence in a proposition, it is compatible with belief in that proposition. Hence, we can charitably generalise deliberation-compatibilism to require compatibility with any non-zero credence.

show that there are no alternative ways of securing all three desiderata. There is at least one candidate[5]: Rational decision-makers apply CDT on the supposition, perhaps subjunctive, that their choice is not predetermined. That is, rational decision-makers assign zero credence—*qua* probability for expected utility calculation—to their choice being predetermined. This allows CDT to avoid Ahmed's counterexamples because they presuppose rational decision-makers give non-zero credence to their choice being predetermined. And it secures Two-boxing in NEWCOMB'S PROBLEM because the supposition that our choice is not predetermined is equivalent to the supposition that only $K_{1.1}$ and $K_{2.1}$ in Table 3 receive positive credence—in which case Two-boxing dominates One-boxing even with the finer-grained Table 3. And it allows us to maintain deliberation-compatibilism because—unlike No Decision responses—it does not require that rational decision-makers assign zero credence—*qua* degree-of-belief—to their choice being predetermined. We could say they *assume*, or *act-as-though*, or *pretend*, etc., their choice is not predetermined. This is not the place to mount a full defence of this alternative, but it is clear that S&W need to tell us more for compatibilism to motivate the otherwise *ad hoc* claims their picture of NEWCOMB'S PROBLEM relies on.

S&W want to argue that all rational decision-makers care about whether outcomes are worth taking seriously in BETTING ON THE PAST, but that none do so in NEWCOMB'S PROBLEM. And that, unlike uncertainty about any other matter of fact, we should not apply expected utility maximisation to uncertainty about whether our outcomes are worth taking seriously, but should instead appeal to what we know. Fans of expected utility maximisation, and causalists specifically, should not be happy with this picture.

4 SDT and the Bonus Money Newcomb Problem

Now, suppose that one of the above objections fails and S&W can justify restricting rational decision-makers to Table 2 and hence secure Two-boxing in NEWCOMB'S PROBLEM. S&W are still not out of the woods, because defenders of CDT do not care merely about securing Two-boxing in NEWCOMB'S PROBLEM; they could do that by stipulation if that were all they wanted. Rather, they care about capturing the kind of reasoning that justifies Two-boxing in NEWCOMB'S PROBLEM and relevantly similar cases. Unfortunately for S&W, the reasoning that justifies Two-boxing in NEWCOMB'S PROBLEM applies equally to cases where SDT cannot secure Two-boxing on any assumptions. Consider the following variation on NEWCOMB'S PROBLEM:[6]

> **BONUS MONEY NEWCOMB PROBLEM (BMNP):** Everything is as in NEWCOMB'S PROBLEM, except: at the same time as the prediction, a bonus of $100,000 (henceforth $100k) was placed in the opaque box if and only if the past and laws entail that you will One-box.

[5]Which we might consider an alternative reading of [Joyce 2016]. See also the FFCD solution in section 3 of my [Solomon 2021].

[6]See section 4 of my [ibid.] for more on this case. SDT is a version of what I there call the PPII solution.

Causalists should endorse Two-boxing, even with the addition of the bonus: exactly the same reasoning that justifies Two-Boxing in NEWCOMB'S PROBLEM applies here. Quoting S&W:

> 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! [S&W 2021: 3]

But the decision whether to place the bonus in the box was also made yesterday. If the contents of the opaque box are out of your control in NEWCOMB'S PROBLEM—as S&W argue—then so is whether you get the bonus in the BMNP. There is no difference between the facts about the prediction and the facts about the bonus that could justify treating them differently—indeed, they might be the very same facts. Whatever the contents of the opaque box now, Two-boxing guarantees you an extra $1000—so you should take both boxes. The standard argument for Two-Boxing in NEWCOMB'S PROBLEM therefore justifies Two-boxing in the BMNP.

	Two-boxing predicted		One-boxing predicted	
	Predetermined to One-box $= K_1$	¬(Predetermined to One-box) $= K_2$	Predetermined to One-box $= K_3$	¬(Predetermined to One-box) $= K_4$
One-box	$\$100k$	\$0	$\$1m + \$100k$	$\$1m$
Two-box		$\$1k$		$\$1m + \$1k$

Table 4

However, SDT will sometimes endorse One-boxing in the BMNP. The Dependency Hypotheses are as in Table 4. The grey squares are not worth taking seriously since you only get the bonus if the past and laws are inconsistent with Two-boxing. Applying SDT we find that:

$$EU(\text{One-box}) = Cr(K_1) \times (\$100k) + Cr(K_3) \times (\$1m + \$100k) + Cr(K_4) \times (\$1m)$$
$$EU(\text{Two-box}) = Cr(K_2|\neg(K_1 \vee K_3)) \times (\$1k) + Cr(K_4|\neg(K_1 \vee K_3)) \times (\$1m + \$1k)$$

And there is nothing in SDT to stop EU(One-box) being greater than EU(Two-box). For example, let $Cr(K_3) = 0.7$ and $Cr(K_1) = Cr(K_2) = Cr(K_4) = 0.1$. Then $EU(\text{One-box}) = \$880,000$. But if $Cr(K_2|\neg(K_1 \vee K_3)) = 0.2$ and $Cr(K_4|\neg(K_1 \vee K_3)) = 0.8$ then $EU(\text{Two-box}) = \$801,000$. Hence, SDT sometimes endorses One-boxing in the BMNP.

The addition of the bonus in the BMNP forces SDT to consider a partition of Dependency Hypotheses where some outcomes are worth taking seriously and some are not. Given this, SDT treats whether you get the bonus as dependent on your choice. But the past and the laws—which entirely determine whether you get the bonus—do not depend

causally on your choice. Causalists should prefer Two-boxing in the BMNP, but SDT cannot secure it.

5 Wrapping Up

S&W want SDT to respect "the motivations behind CDT while making well-motivated departures from it in deterministic cases" [S&W 2020: 1]. The problem, in a nutshell, is that the main motivation for CDT—NEWCOMB'S PROBLEM—is a deterministic case. As is any case where the decision-maker is uncertain if their choice is predetermined. Rational decision-makers respond to such uncertainty by expected utility maximisation. And, if SDT is the correct way to maximise expected utility, we will find that One-boxing is sometimes rational in NEWCOMB'S PROBLEM. Furthermore, SDT endorses One-boxing in the BMNP, but the argument that justifies Two-boxing in NEWCOMB'S PROBLEM applies equally well there. Even if SDT can secure Two-boxing in NEWCOMB'S PROBLEM, it cannot secure what causalists really care about: respect for the reasoning that justifies Two-boxing in NEWCOMB'S PROBLEM.[7]

References

Ahmed, A. 2014a. Causal Decision Theory and the Fixity of the Past, *The British Journal for the Philosophy of Science* 65/4: 665–85.

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

Joyce, J. M. 2016. Review of *Evidence, Decision and Causality* by Arif Ahmed, *The Journal Of Philosophy* 113/4: 224–32.

Lewis, D. 1981a. Are We Free to Break the Laws?, *Theoria* 47/3: 113–21.

Lewis, D. 1981b. Causal Decision Theory, *Australasian Journal of Philosophy* 59/1: 5–30.

Pereboom, D. 2014. *Free Will, Agency, and Meaning in Life*, Oxford: Oxford University Press.

Sandgren, A., and Williamson, T. L. 2021. Determinism, Counterfactuals, and Decision, *Australasian Journal of Philosophy* 99/2: 286–302, DOI: 10.1080/00048402.2020.1764073.

Solomon, T. C. P. 2021. Causal Decision Theory's Predetermination Problem, *Synthese* 198/6: 5623–54, DOI: 10.1007/s11229-019-02425-0.

[7]Thanks for useful discussion and comments must go to Alan Hájek, Brian Hedden, Alex Sandgren, and Timothy L. Williamson. Two anonymous reviewers for this journal provided particularly helpful comments that substantially improved the paper.