Escaping the Cycle

J. Dmitri Gallow †

Abstract: I present a decision in which causal decision theory appears to violate the *independence of irrelevant alternatives* (IIA) and *normal-form extensive-form equivalence* (NEE). I show that these violations lead to exploitable behavior and long-run poverty. These consequences appear damning, but I urge caution. Causalists can dispute the charge that they violate IIA and NEE by carefully specifying when one decision is a *subdecision* of another.

As I'll understand it here, the *independence of irrelevant alternatives* (IIA) says that adding an additional, irrelevant, option to the menu can't transform an impermissible choice into a permissible one. An old story attributed to Sidney Morgenbesser illustrates the seeming irrationality of violating this principle: asked to decide between steak and chicken, a man says "I'd rather have the steak". The waiter tells him that they also have fish, to which he responds: "Oh, in that case, I'll have the chicken". This behavior looks irrational, and a principle like IIA explains why. The principle is quite plausible; all else equal, we should want a theory of rational choice which vindicates it.

The principle I'll call *normal-form extensive-form equivalence* (NEE) says that, so long as you're certain to not change your beliefs or desires, and you're certain to remain rational, if it's permissible to choose an option other than X, then, if you're given the decision to either have X or go on to choose amongst the other options, it is permissible to choose to leave X behind.[1] If, in a decision between chicken, steak, and fish, it's permissible for you to order the steak, then, in a decision between the fish and a decision between chicken and steak, it's permissible to decline the fish. Like IIA, this principle is very plausible; all else equal, we should want a theory of rational choice which vindicates it.

† Thanks to Kevin Dorst, Daniel Drucker, James Shaw, Rohan Sud, and two anonymous reviewers for helpful feedback on this material.

1. This is a weakened version of the principle usually called 'normal-form extensive-form equivalence'; it only infers something about 'extensive-form' permissibility from 'normal-form' permissibility, and it only does so in special conditions. For this reason, it is a bit uncomfortable to name the principle an 'equivalence', but I'll stick to this terminology nonetheless.

Draft of November 19, 2020; Word Count: 10,155
✉: dmitri.gallow@gmail.com

Here, I'll present a decision—called 'Utility Cycle', for reasons which will become clear—in which orthodox causal decision theory (CDT) appears to violate both IIA (§2.1) and NEE (§2.2). In minor variants of Utility Cycle, these violations lead causalists to engage in exploitable behavior like paying to have options presented to them in a certain order, and paying to change their choice once it's been made, for no apparent reason (§2.3). These consequences look bad. Some will see them as a reason to reject CDT. But I will urge caution. Principles like IIA and NEE concern two decisions, where the first is a *subdecision* of the second. In rough outline, a decision, $\mathscr{D}$, is a subdecision of another, $\mathscr{D}^*$, just in case the states of nature in the two decisions are relevantly similar and the available options in $\mathscr{D}$ are relevantly similar to a subset of the options in $\mathscr{D}^*$. Then, for instance, IIA says that, if it is impermissible to choose X in $\mathscr{D}$, and $\mathscr{D}$ is a subdecision of $\mathscr{D}^*$, then it is impermissible to choose the option corresponding to X in $\mathscr{D}^*$. So in order to show that CDT violates principles like IIA or NEE, we must make some assumptions about what it takes for one decision to be a subdecision of another. Given a natural assumption about when one decision is a subdecision of another, CDT will violate IIA and NEE. But I'll suggest an alternative approach to causalists which allows them to satisfy the principles (§§3–4). I'll additionally counsel causalists to defend their exploitable behavior as an intrapersonal tragedy of the commons: agents incapable of binding themselves to a course of action can be led to predictable financial ruin through a series of rational actions, in just the same way that society may be predictably led to collective tragedy through a series of individually rational actions.

1 Causal Decision Theory

1.1 Desire. I will assume that, when you face a decision,[2] you have some set of available *options* $\mathcal{O} = \{X_1, X_2, \ldots, X_N\}$ between which you must choose. When making this choice, there is some set of *states of nature* $\mathcal{K} = \{K_1, K_2, \ldots, K_M\}$, which, for all you know, may obtain.[3] Exactly one of the K_i obtains, though you know not which; nor are you in any position to influence which obtains.[4] Though you do not know which K_i obtains, you do have opinions,

2. Terminology: English uses 'decision' and 'choice' ambiguously to refer both to the situation in which you must select one of a set of options, and the action of selecting one of those options. To avoid confusion, I will reserve 'decision' for the situation you face, and 'choice' for the act of selection. Thus: you *face* a decision, and you *make* a choice.

3. Throughout, I'll use letters like 'X' and 'K' to stand both for options and states and the proposition that you've chosen those options and that those states obtain. Context will disambiguate.

4. Nothing much will hang upon how we think about states of nature, but if we can spot ourselves conditional excluded middle, then I am happy to go along with Lewis (1981a) in taking them to be conjunctions of conditionals specifying which outcome each option counterfactually implies.

$\mathcal{D}(Row \wedge Col)$	K_L	K_M
L	100	0
M	110	10

$\mathcal{P}(Row \mid Col)$	L	M
K_L	90%	10%
K_M	10%	90%

Table 1: Desires and Probabilities for Newcomb. The matrix on the left shows how strongly you desire choosing the row option while in the column state. The matrix on the right shows the probability that you are in the row state, given that you've chosen the column option.

represented with a probability function, $\mathcal{P}$, defined over both $\mathcal{O}$ and $\mathcal{K}$. Finally, we can represent your desires with a function, $\mathcal{D}$, which says how strongly you desire that you select each option, in each state of nature. I assume that, for any option $X \in \mathcal{O}$,

$$\mathcal{D}(X) = \sum_K \mathcal{P}(K \mid X) \cdot \mathcal{D}(X \wedge K)$$

$\mathcal{D}(X)$ tells us how good you would expect things to be, were you to learn that you have chosen X. If $\mathcal{D}(X)$ is high, then you should be glad to learn that you've chosen X—low, and you should be sad to learn that you've chosen X.

1.2 Newcomb. Some—known as *evidential decision theorists*—think that $\mathcal{D}(X)$ provides a measure of the *choiceworthiness* of an option X.[5] Causal decision theorists disagree, because of cases like the following:

> Newcomb
> You are on a game show. Before you are two boxes, labelled 'L' and 'M' (for 'less' and 'more'). You may take one, and only one, of the boxes. Money was placed in the boxes on the basis of a reliable prediction. If it was predicted that you would take L, then \$100 was placed in box L and \$110 was placed in box M. If it was predicted that you would take M, then \$0 was placed in box L and \$10 was placed in box M. These predictions are 90% reliable—that is, conditional on you selecting box X, the chance that it was predicted that you would select X is 90%. But nothing you do now will affect how much money is in the boxes.

We can represent this decision with the two matrices shown in table 1. There are two relevant states of nature. Either it was predicted that you would take box L, 'K_L', or it was predicted that you would take box M, 'K_M'. I suppose that your desires are linear in dollars, so that the degree to which you desire each option in each state are as shown in the $\mathcal{D}$-matrix on the left of table 1. The matrix on the right says: given that you choose box L, you're 90% likely to be

5. For defenses of evidential decision theory, see Jeffrey (1965, 2004) and Ahmed (2014b).

in state K_L and 10% likely to be in state K_M. And, given that you choose box M, you're 10% likely to be in state K_L and 90% likely to be in state K_M.

In Newcomb, you should be happier to learn L than M, since

$$\begin{aligned} \mathcal{D}(L) &= \mathcal{P}(K_L \mid L) \cdot \mathcal{D}(L \wedge K_L) + \mathcal{P}(K_M \mid L) \cdot \mathcal{D}(L \wedge K_M) \\ &= 90\% \cdot 100 + 10\% \cdot 0 \\ &= 90 \end{aligned}$$

$$\begin{aligned} \text{while} \quad \mathcal{D}(M) &= \mathcal{P}(K_L \mid M) \cdot \mathcal{D}(M \wedge K_L) + \mathcal{P}(K_M \mid M) \cdot \mathcal{D}(M \wedge K_M) \\ &= 10\% \cdot 110 + 90\% \cdot 10 \\ &= 20 \end{aligned}$$

So evidential decision theorists advise you to take box L. But notice that, no matter what was predicted, taking box M will get you strictly more money. In each state of nature, taking box M will get you \$10 more than taking L will. Notice also: if you were to learn which prediction was made, you would be happier to learn M than L, and evidential decision theorists would advise you to take M—*no matter what* you learned. If you were to learn K_L, you'd desire M more than L. And if you were to learn K_M, you'd desire M more than L. Evidential decision theorists therefore violate a principle of deontic reflection: they recommend options which they know your better informed, future self will wish you had not chosen.[6]

We may dramatize this violation of deontic reflection in the case of Newcomb. Suppose that the evidential decision theorist faces Newcomb, and they are playing, not for themselves, but rather for a poor orphan boy, Oliver. While they are not allowed to look in the boxes, Oliver is. He is there with them as they choose. He is allowed to offer the evidentialist advice about which box to choose, but he is not allowed to tell them the contents of the boxes. He looks inside, and says: 'Please, choose box M'. (Of course he does—the evidentialist knew that's what he'd say, no matter what he saw.) The evidential decision theorist ignores Oliver's advice, and chooses box L instead. They tell him: 'If you were able to tell me what the boxes contain, I would agree with you, and I would choose M, no matter what you told me. But, since you haven't told me what's in the boxes, I must take box L.' At this point, the producers of the game show—who are really pulling for Oliver—intervene. They say: 'If you allow him, Oliver may tell you what the boxes contain.' The evidential decision theorist does not allow him. They say: 'If I allow you to tell me what's in the boxes, then I will end up taking box M. But currently, I think that's worse than choosing L. So I think it's better for me to not know.' The producers try a different tack. They say: 'Alright, if you don't listen to what Oliver has to say

6. See Arntzenius (2008)

about the contents of the boxes, then we'll take \$60 away from whatever Oliver wins (perhaps leaving him with a bill to pay).' The evidential decision theorist knows that, if they listen to Oliver, they'll take box M. They desire taking M with a strength of \$20. On the other hand, if they don't listen, they'll take box L. They desire taking L with a strength of \$90. Minus the \$60 lost by not listening, not listening is desired with a strength of \$30. So, in order to keep Oliver quiet, they'll take \$60 away from him.[7]

Imagine yourself as Oliver, pleading with the evidential decision theorist to take the box that you can see contains an additional \$10. They are choosing only for your benefit. You are telling them that M is the box which will most benefit you. They believe you. They know that box M will benefit you the most. Yet they refuse to take it. They moreover refuse to take the information you are trying to give them, even though they know that this information is not in any way misleading, that it will teach them what is objectively in your best interest, and that their learning this information is objectively in your best interest. To keep themselves from learning this information, they are willing to take \$60 away from you—though, again, their only concern is maximizing *your* welfare. Does this look like the behavior of a rational agent? The causal decision theorist thinks not, and I agree. And so I think that $\mathcal{D}$ does not give an adequate measure of the choiceworthiness of an option. You should not always choose the option which you'd be happiest to learn that you'd chosen. Sometimes, you should be sad to learn that you're choosing rationally.

1.3 Utility. According to the orthodox causal decision theorist, we should measure the choiceworthiness of an option, X, not by looking at how glad you'd be to learn that you have selected it, $\mathcal{D}(X)$, but rather by looking at the degree to which you expect X to *bring about* your desired ends. For each $K \in \mathcal{K}$, $\mathcal{D}(XK)$ is the degree to which X would bring about your desired ends, were you to choose it in the state K. So the quantity

$$\mathcal{U}(X) \stackrel{\text{def}}{=} \sum_{K} \mathcal{P}(K) \cdot \mathcal{D}(X \wedge K)$$

tells us how desirable you expect choosing X to *make* the world.[8]

The difference between $\mathcal{D}$ and $\mathcal{U}$ is that, in $\mathcal{D}$, we conditioned the probability function $\mathcal{P}$ on the proposition that you choose X. In $\mathcal{U}$, we do not. Your choice may give evidence that a state of nature obtains, but it does nothing to

7. See Wells (2019).

8. This is Skyrms's formulation of causal decision theory. There are alternatives—see, *e.g.*, Sobel (1994), Lewis (1981a), Joyce (1999), and especially Rabinowicz (1982) and Rabinowicz (2009). The differences between these version of CDT won't make a difference to anything I have to say here.

bring that state about (that's what it is for K to be a state *of nature*). According to causalists, the fact that an option makes a desired state more likely doesn't speak in its favor if it doesn't causally affect whether that state obtains or not.

Just as you may evaluate the utility of an option, X, from the perspective you currently occupy, so too may you evaluate the utility of X from the perspective you would occupy, were you to choose another option, Y. (I mean: the perspective you would occupy after learning *only* that you'd chosen Y, and before learning anything else.) From this perspective, your probability for each state K would be $\mathcal{P}(K \mid Y)$, and

$$\mathcal{U}_Y(X) \stackrel{\text{def}}{=} \sum_K \mathcal{P}(K \mid Y) \cdot \mathcal{D}(X \wedge K)$$

would be the utility of X. Given the quantities $\mathcal{U}_Y(X)$, for each pair of options X and Y, we may calculate $\mathcal{U}(X)$ as follows.

$$\mathcal{U}(X) = \sum_Y \mathcal{U}_Y(X) \cdot \mathcal{P}(Y)$$

In a decision between two options, X and Y, both of the following situations are possible:

> Self-Undermining Decision
> Once chosen, each option would have a lower utility than the alternative
>
> $$\mathcal{U}_X(Y) > \mathcal{U}_X(X) \quad \text{and} \quad \mathcal{U}_Y(X) > \mathcal{U}_Y(Y)$$
>
> Self-Reinforcing Decision
> Once chosen, each option would have a higher utility than the alternative
>
> $$\mathcal{U}_X(X) > \mathcal{U}_X(Y) \quad \text{and} \quad \mathcal{U}_Y(Y) > \mathcal{U}_Y(X)$$

This can lead CDT's verdicts to change as you make up your mind about what to do. In a self-undermining decision, once you follow CDT's advice and intend to choose the option it called rational, it will change its mind and call your choice irrational. In a self-reinforcing decision, if you disregard its advice and do what it deemed irrational, CDT will change its mind and call you rational for doing so.[9]

9. *Cf.* Gibbard & Harper (1978), Richter (1984), Weirich (1985), Harper (1986), Egan (2007), Joyce (2012), Hare & Hedden (2016), Armendt (2019), and Williamson (forthcoming).

I believe that cases like these give us reason to doubt CDT. I defend a heterodox revision of causal decision theory whose verdicts do not depend upon your option probabilities. But these kinds of decisions won't be relevant to the arguments against CDT which I'll introduce below.[10] For those arguments, I need only appeal to the following, minimal commitment of CDT, which is also endorsed by heterodox causalists like myself:[11]

Minimal CDT In a decision between two options, X and Y, if X's utility would exceed Y's, whichever you chose,

$$\mathcal{U}_X(X) > \mathcal{U}_X(Y) \quad \text{and} \quad \mathcal{U}_Y(X) > \mathcal{U}_Y(Y)$$

then X is required and Y is impermissible.

(Thus: I distinguish between CDT and **Minimal CDT**. The latter is strictly weaker than the former; **Minimal CDT** only applies in decisions between two options, where the decision is neither self-undermining nor self-reinforcing.) In Newcomb, **Minimal CDT** tells us that M is required and L is impermissible. You know that, no matter what was predicted, M will get you \$10 more than L does. So, if you choose L, then the utility of M will exceed the utility of L by 10 ($\mathcal{U}_L(M) = 100$ and $\mathcal{U}_L(L) = 90$). And, if you choose M, then the utility of M will exceed the utility of L by 10 ($\mathcal{U}_M(M) = 20$ and $\mathcal{U}_M(L) = 10$). So the utility of M will exceed the utility of L, whichever box you happen to take.

2 Utility Cycle, and Three Objections to Minimal CDT

Consider the following decision:[12]

> Utility Cycle
> Before you are three boxes, labeled 'A', 'B', and 'C'. You may take one and only one of the boxes. The contents of the boxes were decided on the basis of a prediction about how you would choose. If it was predicted that you would choose A, \$100 was left in B and a bill for \$100 was left in C. If it was predicted that you would choose B, \$100 was left in C and a bill for \$100 was left in A. If it was predicted you would choose C, \$100 was left in A and a bill for \$100 was left in B. These predictions are 80% reliable.

Your desires and probabilities for this problem are shown in table 2. Which

10. Though I'll return to these kinds of decisions in §5.

11. **Minimal CDT** is accepted by Wedgwood (2013), Barnett (ms), Spencer (forthcomingb), Gallow (2020), and Podgorski (forthcoming), as well as by *deliberational* causal decision theorists like Skyrms (1990), Arntzenius (2008), and Joyce (2012, 2018).

12. *Cf.* Ahmed (2012) and Hare & Hedden (2016).

$\mathcal{D}(Row \wedge Col)$	K_A	K_B	K_C
A	0	−100	100
B	100	0	−100
C	−100	100	0

$\mathcal{P}(\ Row\ \vert Col)$	A	B	C
K_A	80%	10%	10%
K_B	10%	80%	10%
K_C	10%	10%	80%

Table 2: Desires and Probabilities for Utility Cycle

option has the highest utility depends upon how likely you think you are to select each option. Let 'a', 'b', and 'c' be your probabilities that you will take box A, B, and C, respectively. Then:[13]

$$\mathcal{U}(A) = 70(c - b) \qquad \mathcal{U}(B) = 70(a - c) \qquad \mathcal{U}(C) = 70(b - a)$$

So, for illustration: if you're most likely to take A, and more likely to take B than C ($a > b > c$), then B will have the highest utility; if you're most likely to take B, and more likely to take C than A ($b > c > a$), then C will have the highest utility; and if you're more likely to take C than A, and more likely to take A than B ($c > a > b$), then A will have the highest utility.

Suppose now that you face a decision between just A and B—C is taken off of the menu (note, however, that even though you are guaranteed to not take C, there is still a 10% probability that it was falsely predicted that you'd take C). In that case, your probability for C, c, is constrained to be zero, and the utilities for A and B are:

$$\mathcal{U}(A) = 70a - 70 \qquad \mathcal{U}(B) = 70a$$

(To get these formulae, just set $c = 0$ in the equations '$\mathcal{U}(A) = 70(c - b)$' and '$\mathcal{U}(B) = 70(a - c)$', and then set $b = 1 - a$.) No matter the value of a, B will have a higher utility than A. So **Minimal CDT** says that, in a decision between A and B, B is required and A is impermissible. Suppose, on the other hand, that A is removed from the menu, and you face a decision between B and C. In that case, your probability for A, a, is constrained to be zero, and the utilities of B and C are:

$$\mathcal{U}(B) = 70b - 70 \qquad \mathcal{U}(C) = 70b$$

Again, no matter the value of b, the utility of C will exceed the utility of B. So **Minimal CDT** says that, in a decision between B and C, C is required and

13. I will spare the reader the tedium of deriving everything explicitly in the main text. For those who wish to check the math, some advice: multiply the matrix $\mathcal{D}(Row \wedge Col)$ by the matrix $\mathcal{P}(Row|Col)$. This gives the matrix $\mathcal{U}_{Col}(Row)$, of the utility of the row option, from the perspective you'd occupy immediately after choosing the column option. The identity $\mathcal{U}(Row) = \sum_{Col} \mathcal{U}_{Col}(Row) \cdot \mathcal{P}(Col)$ can then be used to easily calculate the unconditional utilities, $\mathcal{U}(Row)$.

B is impermissible. Similarly, if B is removed from the menu, and you face a decision between C and A, the utilities of C and A will be:

$$\mathcal{U}(C) = 70c - 70 \qquad\qquad \mathcal{U}(A) = 70c$$

The utility of A will exceed the utility of C, no matter the value of c. So **Minimal CDT** says that, in a decision between C and A, A is required and C is impermissible.

2.1 The Independence of Irrelevant Alternatives. If we assume that Utility Cycle is not a rational dilemma (*i.e.*, if we assume that *some* option is permissible), then **Minimal CDT** appears to lead to a violation of a principle known as *the independence of irrelevant alternatives* (or just '**IIA**').

IIA: If, in a decision between X and Y, X is not permissible, then, in a decision between X, Y, and Z, X is not permissible.[14]

According to **Minimal CDT**, *every* option in Utility Cycle is impermissible in a one-on-one decision with some alternative. So, if *some* option is permissible,[15] then it looks like we will have a violation of **IIA**. For illustration: suppose that A is a permissible choice in Utility Cycle. By **Minimal CDT**, in a decision between A and B, A is impermissible. So A is not a permissible choice when you face the restricted menu $\{A, B\}$, but it *is* a permissible choice when you face the larger menu $\{A, B, C\}$. And this contradicts **IIA**. The same goes if we say that B or C is permissible instead. For **Minimal CDT** says that B is impermissible on the restricted menu $\{B, C\}$, and C is impermissible on the restricted menu $\{C, A\}$.

2.2 Normal-Form Extensive-Form Equivalence. Utility Cycle also seems to show that **Minimal CDT** violates a weak principle of *normal-form extensive-form equivalence* (or just '**NEE**').

NEE: If you are certain to remain rational and your beliefs and desires are certain to not change, then, if it is permissible to not choose X in a decision between X, Y, and Z, then, in a decision between X and a decision between Y and Z, it is permissible to not choose X. (See figure 1.)

14. We should sharply distinguish **IIA** from other principles that go by that name. For instance, Podgorski (forthcoming) calls the following principle 'the independence of irrelevant alternatives': Your preference between X and Y is a function of $\mathcal{U}_X(X)$,$\mathcal{U}_X(Y)$,$\mathcal{U}_Y(X)$, and $\mathcal{U}_Y(Y)$ alone. This principle is logically independent from the one I'm calling '**IIA**'. See Ray (1973) for more on conflicting uses of 'independence of irrelevant alternatives' in social choice theory.

15. By the symmetry of the case, we should conclude that *every* option is permissible, but we need not assume this in order to make the present point.

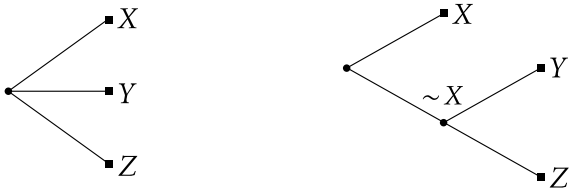
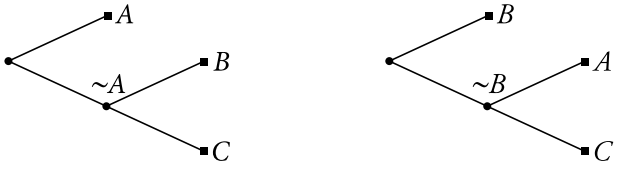


FIGURE 1: **NEE** says that, if it is permissible to not choose X in the decision between $X, Y,$ and Z on the left, then it is permissible to not choose X in the decision between X and a decision between Y and Z on the right (so long, that is, as you are certain to retain your beliefs, desires, and rationality).



(a) A OR ~A (b) B OR ~B

FIGURE 2

The antecedent of **NEE** is important. Suppose you think that, if you were to forego X, your beliefs or desires might change before you decide between Y and Z. Then, it may be rational to choose X now in order prevent your not-entirely-trustworthy future self from choosing against your current interests. Likewise, if you fear that your future self will not choose rationally, this could give you additional reason to select X at stage one. However, restricted to cases where you are certain to retain your beliefs, desires, and rationality, **NEE** is very plausible.

Consider now the following two decisions (see figure 2):

> A OR ~A
> Money was distributed between boxes A, B, and C as in UTILITY CYCLE. At stage 1, you may either take box A or not. If you take box A, then you receive its contents. If you don't, then at stage 2, you decide between B and C. You are certain to retain your beliefs, desires, and rationality throughout.
>
> B OR ~B
> Money was distributed between boxes A, B, and C as in UTILITY CYCLE. At stage 1, you may either take box B or not. If you take box B, then you receive its contents. If you don't, then at stage 2, you decide between A and C. You are certain to retain your beliefs, desires, and rationality throughout.

Assume **Minimal CDT**, and assume also that you know you will abide by **Min-**

imal CDT throughout any sequential decisions. Then, in A OR ∼A, if you choose ∼A at stage 1, at stage 2, you will choose C, and you know this at stage 1. So, at stage 1, you are deciding between A and C. So A is required at stage 1. In B OR ∼B, if you choose ∼B at stage 1, then, at stage 2, you will choose A, and you know this at stage 1. So, at stage 1, you are deciding between B and A. So B is required at stage 1.

We can now argue that, assuming *some* option is permissible, **Minimal CDT** violates **NEE** in UTILITY CYCLE. For, in a decision between A, B, and C, B is either permissible or it is not. Suppose it is. Then, **NEE** says that ∼A is permissible in A OR ∼A. **Minimal CDT** on the other hand, says that ∼A is impermissible, contradicting **NEE**. Suppose on the other hand that B is impermissible. Then, it is permissible to not choose B (this follows because we've assumed that some option is permissible). In that case, **NEE** says that ∼B is permissible in B OR ∼B. **Minimal CDT**, on the other hand, says that ∼B is impermissible, contradicting **NEE**. Either way, **Minimal CDT** contradicts **NEE**.

2.3 Predictable Long-run Poverty. **Minimal CDT**'s advice in UTILITY CYCLE may be exploited to lose you money in the long run. Suppose that, instead of taking a box yourself, you select a box with the aid of an assistant. You tell the assistant which box to take, but it is the assistant who makes the final selection. (You keep the money. Note also that the reliable predictions are now about which box your assistant will end up selecting.) By the symmetry of the case, you see no reason to favor any box over the others, and you tell your assistant to take box A. Before your assistant departs, they get an idea. They say: 'Are you sure? I'll give you an opportunity to change to box B (but not box C—I'm taking that off the menu). In exchange for changing your mind, I'll require \$60.' (You are certain that they will take this choice to be final, they will take the box you choose, and that there's no longer any way to get them to take C.) At this point, you face a new decision: not between A, B, and C, but instead between sticking with A and switching to B and losing \$60. If a is your probability for taking A, then the utilities of the available options are:

$$\mathcal{U}(A) = 70a - 70 \qquad \mathcal{U}(B) = 70a - 60$$

In this new decision, switching to B will have a higher utility than sticking with A, no matter whether you take A or switch to B. So **Minimal CDT** says to hand your assistant \$60 to have them take B instead. But you could have had B in the first place, for free. How could your assistant's offer give you reason to switch?

Nothing changes if we suppose that you know in advance that your assistant will make you an offer of this kind. Suppose you know all of the following in advance: at stage 1, you will make an *initial selection*. Then, at stage 2, your assistant will give you the opportunity to switch for \$60. If your initial selection

at stage 1 is box A, then at stage 2, your assistant will make you an offer to switch to B for \$60. If your initial selection at stage 1 is box B, then at stage 2, your assistant will make you an offer to switch to C for \$60. And if your initial selection at stage 1 is box C, then at stage 2, they'll make you an offer to switch to A for \$60. In each of these cases, paying to switch will have a higher utility than sticking with your initial selection, and **Minimal CDT** will require you to pay to switch. So long as you abide by **Minimal CDT** at stage 2, there's no initial selection you could make at stage 1 which would prevent your future self from paying to switch at stage 2.

Note that, if you pay to switch, then you will likely end up losing money overall. You have an 80% chance of breaking even, a 10% chance of winning \$100, and a 10% chance of losing \$100—so you have an expected return of \$0. And you've just handed over \$60. In the long run in which you make this choice over and over again, with your assistant offering the trade each time, you will lose \$60 on average. In contrast, someone who refuses the assistant's offer to switch will break even, on average. Note also that every series of choices permitted by **Minimal CDT** is *causally dominated* by another series of choices. Whatever box you end up with after paying \$60 to switch, you could have had that box's contents for free by simply making it your initial selection and refusing to switch.

Causalists are used to making less money in certain decisions. For instance, anyone who takes box M in Newcomb will predictably make less money, over the long run, than someone who takes box L. The usual causalist reply is convincing: this is true, but only because those who take L will typically be *provided* with more money than those who take box M. Being afforded greater opportunities for wealth is no sign of rationality; nor is being afforded fewer opportuntities for wealth a sign of irrationality. So predictable poverty in Newcomb is no sign of irrationality.[16] A comparable defense is not available here. In this case, it was not an unfortunate environment which led to your poverty. Over the long run, someone who was indifferent between A and B in a decision between the two would never pay to switch, and they would predictably end up making more money in the long run, in exactly the same environment.

Minimal CDT will advise you to pay to have options presented to you in a certain order—even when you're certain to retain your beliefs, desires, and rationality throughout. For instance, consider Pay or A:

> Pay or A
> Money is distriuted between boxes A, B, and C as in Utility Cycle. At stage 1, you may either pay \$60, P, or not, $\sim P$. If you pay,

16. See, *e.g.*, Gibbard & Harper (1978), Lewis (1981b), Joyce (1999), Bales (2018), and Wells (2019).

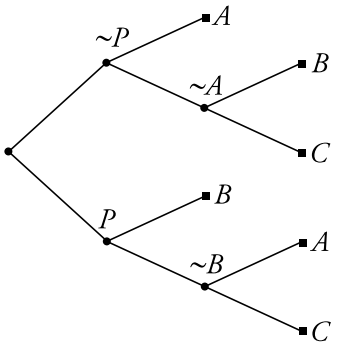


Figure 3: Pay or A

> then, at stage 2, you will face the decision B or ~B. If you do not, then, at stage 2, you will face A or ~A. (See figure 3.)

If you know that you abide by **Minimal CDT**, you will choose *A* in A or ~A. So, if you don't pay, you will end up choosing *A*. If you abide by **Minimal CDT**, you will choose *B* in B or ~B. So, if you pay, you will end up choosing *B*. So, at stage 1, you face a decision between paying $60 and taking box *B* and not paying and taking box *A*. This is the same choice you faced with your assistant. And, again, **Minimal CDT** tells you to pay the $60.

Again, paying likely leads to you losing money overall. Whether you play A or ~A or B or ~B, the expected return is $0. So in the long run in which you choose to pay in Pay or A over and over again, you will lose $60 on average. Again, someone who was indifferent between *A*, *B*, and *C* in a decision between any two would predictably make more money when facing exactly the same decision in exactly the same circumstances. The series of choices advised by **Minimal CDT**—pay, then take *B*—is causally dominated. No matter what was predicted, another series of choices—don't pay, then refuse *A*, then take *B*—makes $60 more. So this predictable poverty does not appear to be a consequence of poor opportunities.

3 Subdecisions and Utility Profiles

These consequences of **Minimal CDT** look bad. These do not appear to be the choices of a rational agent. It's natural to see the foregoing as an argument against **Minimal CDT**. However, I want to urge caution. Though I reject orthodox causal decision theory, I believe that the weaker claim **Minimal CDT** is correct, and I accept what it says about Utility Cycle.[17] Defenders of

17. That is to say: I accept what CDT says about the decision between any two options in Utility Cycle. In a decision between *A, B,* and *C*, I say you should be indifferent between all three, and that this does not depend upon your option probabilities. See Gallow (2020) for details.

Minimal CDT could reject the principles **IIA** and **NEE**;[18] or they could insist that Utility Cycle is a rational dilemma in which no option is permissible.[19] These moves are available, but I think there's a more attractive option. In my view, the lesson causalists ought to draw from the case is this: the options A and B are importantly different when they appear on the menu $\{A, B\}$ than when they appear on the larger menu $\{A, B, C\}$. For this reason, a decision between A and B is not a *subdecision* of the decision between A, B, and C. Properly understood, the independence of irrelevant alternatives only says that, if A is impermissible in a decision between A and B, and the decision between A and B is a *subdecision* of the decision between $A, B,$ and C, then A is impermissible in the decision between $A, B,$ and C. However, since the former decision is *not* a subdecision of the latter, the independence of irrelevant alternatives does not apply.

3.1 Subdecisions. **IIA** says that, if you add an *irrelevant* option, Z, to the menu $\{X, Y\}$, this shouldn't transform an impermissible option into a permissible one. However, not every additional option is truly *irrelevant* to your decision between the original options X and Y. Some apparent counterexamples to **IIA** are not genuine counterexamples, because they involve new considerations which are relevant to how you should evaluate the options X and Y. For instance, consider the following putative counterexample to **IIA**: You arrive at the boss's house for dinner. If she offers you soda or beer, you're disposed to opt for soda (you don't want to come off like a drunkard). If she offers you soda, beer, or whiskey, you're disposed to opt for beer (you don't want to come off as either too straight-laced or too intemperate).[20] Do these choice dispositions violate **IIA**? No. What you value in your drink choice is the signal it sends to your boss, and what signal it sends can depend upon the alternatives she offers you. Additionally, if she offers you whiskey, this provides you with important information about how that signal will be received. This should change the way that you evaluate the options of beer and soda, and it makes your decision between them relevantly different. In other words: the decision you face when presented with the options of soda or beer is not a *subdecision* of the decision you face when presented with the options of soda, beer, or whiskey.

When I argued in §2 above that **Minimal CDT** violated **IIA** and **NEE**, I was implicitly assuming that the decision between box A and B was a subdecision of the decision between boxes $A, B,$ and C. That is: I was implicitly assuming that including the option of taking box C doesn't make any difference to how

18. *Cf.* Wedgwood (2013), who rejects **IIA**.

19. *Cf.* Harper (1986).

20. *Cf.* Sen (1993).

you should choose between A and B. In this section, I'd like to get a bit more explicit about that assumption. In general, I'll take for granted that, once we know (1) the set of available options, (2) the set of relevant *states of nature*, (3) your desire function, and (4) your probability function, we know everything we need to know in order to decide which options are rationally permissible. That is: when it comes to rational choice, these four components give us all of the relevant information. At this point, I don't need to take a stand on whether all of the information they give us is relevant (though, truth be told, I think they give us more information than we need). If $\mathcal{O}$ are your options, $\mathcal{K}$ the states of nature, $\mathcal{D}$ your desires and $\mathcal{P}$ your probabilities, then I'll say that your decision is *characterized by* the quadruple $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$.

Notice that a decision may be characterized by *multiple* such quadruples. For instance, if $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ characterizes your decision and $\mathcal{D}'$ is a positive linear transformation of $\mathcal{D}$, then $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}', \mathcal{P} \rangle$ will also characterize your decision. The differences between $\mathcal{D}$ and $\mathcal{D}'$ are like the differences between measuring temperature in Fahrenheit and in Celsius. Both $\mathcal{D}$ and $\mathcal{D}'$ measure the same quantity—namely, how well satisfied your desires are—but they measure this quantity in different units. When it comes to characterizing your decision, these differences are irrelevant.

Take two decisions, $\mathscr{D}$ and $\mathscr{D}^*$. In general terms, I was assuming in §2 above that $\mathscr{D}$ is a subdecision of $\mathscr{D}^*$ iff there are two quadruples, $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ and $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$, such that (1) $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ characterizes $\mathscr{D}$ and $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$ characterizes $\mathscr{D}^*$, and (2) there is some way of associating each option $X \in \mathcal{O}$ with a unique option $X^* \in \mathcal{O}^*$, and each state $K \in \mathcal{K}$ with a unique state $K^* \in \mathcal{K}^*$, such that (a) in $\mathscr{D}$, the degree to which you desire selecting each option $X \in \mathcal{O}$ in each state $K \in \mathcal{K}$ is the degree to which, in $\mathscr{D}^*$, you desire selecting the corresponding option $X^* \in \mathcal{O}^*$ in the corresponding state $K^* \in \mathcal{K}^*$; and (b) in $\mathscr{D}$, the probability you give to each state $K \in \mathcal{K}$, conditional on each option $X \in \mathcal{O}$, is the same as the probability you give to the corresponding state $K^* \in \mathcal{K}^*$, conditional on the corresponding option $X^* \in \mathcal{O}^*$ in the decision $\mathscr{D}^*$. Call this 'the natural view' of when one decision is a subdecision of another.

The Natural View $\mathscr{D}$ is a *subdecision* of $\mathscr{D}^*$ iff there are two quadruples $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ and $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$, such that

(1) $\mathscr{D}$ is characterized by $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ and $\mathscr{D}^*$ is characterized by $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$; and

(2) there is an injection, $f : \mathcal{O} \to \mathcal{O}^*$ and an injection $g : \mathcal{K} \to \mathcal{K}^*$ such that, for each $X \in \mathcal{O}$ and each $K \in \mathcal{K}$,[21]

21. An *injection* $f : \mathcal{O} \to \mathcal{O}^*$ is a function from $\mathcal{O}$ to $\mathcal{O}^*$ which maps each $O \in \mathcal{O}$ to a *unique* $O^* \in \mathcal{O}$. That is: for every $X, X' \in \mathcal{O}$, if $X \neq X'$ then $f(X) \neq f(X')$. Likewise, for every $K, K' \in \mathcal{K}$, if $K \neq K'$, then $g(K) \neq g(K')$.

(2a) $\mathcal{D}(X \wedge K) = \mathcal{D}^*(f(X) \wedge g(K))$; and
(2b) $\mathcal{P}(K \mid X) = \mathcal{P}^*(g(K) \mid f(X))$.

This is a natural and plausible way of saying when one decision is a subdecision of another. According to it, the decision you face between beer and soda is importantly different when they are the only options on the menu and when you have the additional option of whiskey. That is: according to **The Natural View**, whiskey is not an *irrelevant* alternative. For, when you are offered whiskey as an alternative, this changes your opinions about which signals beer and soda will send, which will change the degree to which you desire choosing beer and soda in some state of nature (violating condition (2a)). And, according to **The Natural View**, the decision you face between boxes A and B when they are the only options on the menu is *not* importantly different from the decision you face between A and B when C is also included on the menu. That is, according to **The Natural View**, the decision between A and B will count as a subdecision of the decision between A, B, and C.

Some causalists may wish to say that your option probabilities also play an important role in determining when one decision is a subdecision of another. They may wish to add to **The Natural View** the additional requirement that your probability distribution over options is the same in both $\mathscr{D}$ and $\mathscr{D}^*$. That is, they may wish to add the additional proviso under condition (2) above:

(2c) $\mathcal{P}(X) = \mathcal{P}^*(f(X))$.

This would reconcile **Minimal CDT** with **IIA**, but at the price of trivializing the latter. In a decision between X and Y, your probabilities for X and Y will necessarily sum to 100%. Then, the only way for this to be a subdecision of a decision between X, Y, and Z would be for your option probability for Z to be 0%. On this proposal, so long as you always leave open that you'll select each available option, no decision of yours will ever be a subdecision of any other, and principles like **IIA** will impose no constraint at all.

Alternatively, we may wish to say that your (unconditional) state probabilities help to determine when one decision is a subdecision of another. That is, we may suggest adding to **The Natural View** the requirement that each state $K \in \mathcal{K}$ has the same unconditional probability in the decision $\mathscr{D}$ as its associated state $g(K)$ has in the decision $\mathscr{D}^*$. That is, we may wish to add the following under condition (2) above:

(2d) $\mathcal{P}(K) = \mathcal{P}^*(g(K))$

This suggestion does not trivialize **IIA**, though it has other undesirable consequences. To appreciate these consequences, first note that the law of total probability tells us that each K's unconditional probability is a weighted average of its probability *conditional on* each option, with weights given by your option

probabilities. That is, your probability that the state K obtains, $\mathcal{P}(K)$, is equal to the weighted sum $\sum_X \mathcal{P}(K \mid X) \cdot \mathcal{P}(X)$. And note that, while your conditional probabilities $\mathcal{P}(K \mid X)$ are fixed, your option probabilities, $\mathcal{P}(X)$, will in general change as you deliberate about what to do. For, if you know that you're rational and you reason yourself to the conclusion that you should choose X, you thereby give yourself evidence that you *will* choose X. So your probability that you will choose X, $\mathcal{P}(X)$, will go up. Let us narrow our attention to decisions in which $\mathcal{P}(K \mid X) \neq \mathcal{P}(K \mid Y)$, for every pair of distinct options X and Y. Call a decision like this 'interesting'. In interesting decisions, changes in your option probabilities automatically lead to changes in your (unconditional) state probabilities. So, if your decision is interesting, then deliberating about what to do will change your state probabilities.

For this reason, if we were to add (2d) to **The Natural View**, it would turn out that the decision you face post-deliberation is different from the one you face pre-deliberation. To appreciate this, first note that every decision counts as an (improper) subdecision of itself. Contraposing: if $\mathscr{D}$ is *not* a subdecision of $\mathscr{D}^*$, then $\mathscr{D}$ must be distinct from $\mathscr{D}^*$. Now, if we were to add (2d) to **The Natural View**, then, in an interesting decision, the decision you would face *after* deliberation (once your option probabilities, and therefore, your state probabilities, had changed) would not be a subdecision of the one you face *before* deliberation. So you would face a *different* decision after deliberation than you faced before.

This is odd. It's natural to think that, if you face a decision, then you will continue to face that decision until you choose one of the available options. But the present proposal would force us to disagree; it says that another way to exit a decision is by deliberating about which option to choose. Deliberation is a way of exiting a decision without making a choice.[22] I find it difficult to make sense of this idea. In my view, what it is to *face a decision* is to be forced to choose

22. Notice: that's not the same as saying that one of the available *options* is to deliberate about which option to choose. I am happy to allow that you can sometimes choose to deliberate. But I deny that deliberation is *always* chosen. For I am inclined to say that rational choice requires deliberation. Making the choice which is rational is not the same as choosing rationally; to choose rationally, you must make the right choice for the right reasons. And making a choice for the right reasons requires deliberation, in some good sense of the word 'deliberation' (see Pettit, 2010). If I were then to add that deliberation is always chosen, it would follow that it is impossible to always choose rationally. For suppose you have chosen rationally in a decision $\mathscr{D}_0$. Then, you must have deliberated about what to choose in $\mathscr{D}_0$. But then, you must have *chosen* to deliberate about what to choose in $\mathscr{D}_0$. This choice must have been made in a prior decision, $\mathscr{D}_1$, in which deliberation about $\mathscr{D}_0$ was an option. In order for your choice in $\mathscr{D}_1$ to have been rational, you must have deliberated about *it*. So you must have chosen to deliberate about what to do in $\mathscr{D}_1$. And this requires a prior decision, $\mathscr{D}_2$. The reasoning iterates. To stop the regress, we should accept that sometimes, the action of deliberation is not chosen. Then, the odd consequence of adding (2d) to **The Natural View** is that—*even in cases where deliberation is not chosen*—deliberating about which option to choose is a way of exiting your decision. In those cases, it is a way of exiting a decision without making a choice.

from amongst the available options; if you were able to exit a decision without making a choice, then it seems to me that it is not a decision you truly faced in the first place. But suppose, just for the sake of argument, that adjusting your option probabilities can mean exchanging one decision for another. Suppose you face an interesting decision, and you know that you're going to deliberate about which option to select. Then, you know that, by the time your deliberation is finished, you will no longer face your current decision. Instead, because your options probabilities will have shifted, at the conclusion of your deliberation, you will face a *different* decision. If that's so, then it seems to me that deliberating about your current decision would be a waste of time. Wouldn't that time be better spent thinking about what to do in the decision you *will* face, at deliberation's end? (Since you don't know which direction your deliberation will take, you won't know which decision that is. You might end up inclining towards X, and you might end up inclining towards Y. But, either way, once it comes time to choose, you won't be facing your current decision. Isn't it better, then, to think about the decisions you could end up facing at the moment of choice?) It seems to me that the answer is 'yes'. On this view, deliberation about what to do in your current decision is deliberation about a decision you will no longer face once deliberation has ended—and that is deliberation wasted. Better to think ahead to the decision or decisions you will end up facing at the moment of choice. But once I've convinced myself of this, it becomes difficult for me to understand the sense in which you *face* your current decision at all. It seems to me that a decision which it is irrational for you to deliberate about is not a decision you currently face. Since it's irrational for you to deliberate about the decision you currently face, it is not a decision you currently face. All of the foregoing reasoning strikes me as reasonable, but it has led to a contradiction: you do not currently face the decision you currently face. So I'm inclined to reject the starting assumption which led me there. That is: I'm inclined to resist adding (2d) to **The Natural View**. (All of these considerations apply with equal force to the proposed (2c), as well.)

In sum: I wish to respond to **Minimal CDT**'s apparent violation of **IIA** in Utility Cycle by insisting that the decision between boxes A and B is not a subdecision of the decision between boxes A, B, and C. However, I don't think that we should make this case by appealing to to either your option probabilities or your state probabilities. In §3.2, I'll offer the causalist a different account of when one decision is a subdecision of another. I will then explain, in §4, how this account allows causalists to dispute the charge that they violate **IIA** and **NEE** in Utility Cycle.

3.2 Utility Profiles. The two suggestions from the previous subsection—adding (2c) or (2d) to **The Natural View**—had the consequence that deliberat-

ing about what to do can exchange one decision for another. For the reasons rehearsed above, I wish to avoid this consequence. So I want my criterion for one decision being a subdecision of another to only appeal to properties of your decision which remain fixed throughout deliberation. One property like this is the conditional probability of each state, K, given each option X, $\mathcal{P}(K \mid X)$. This is one reason why **The Natural View** is so natural.

But there are other important properties of your decision which do not change over the course of deliberation. In particular: the utilities you *would* assign to each option, *were* you to learn that you'd chosen any of the available options—that is, the values $\mathcal{U}_Y(X)$, for each X and Y—do not change as you deliberate. And my suggestion to causalists is that they specify when one decision is a subdecision of another partly in terms of these quantities. If this is done carelessly, it can end up trivializing principles like **IIA** and **NEE**. In a decision between X and Y, there are *two* potential post-choice perspectives from which to evaluate the utilities of X and Y—these are the perspectives encoded in $\mathcal{U}_X$ and $\mathcal{U}_Y$. In a decision between X, Y, and Z, there are *three*: $\mathcal{U}_X$, $\mathcal{U}_Y$, and $\mathcal{U}_Z$. If this is enough for the decision between X and Y to not count as a subdecision of the decision between $X, Y,$ and Z, then it will be impossible for a decision between X and Y to be a subdecision of a decision between $X, Y,$ and Z. And **IIA** would never apply.

So let us proceed carefully. Suppose you face a decision $\mathscr{D}$, with the available options $\mathcal{O}$. Let $\langle X_1, X_2, \ldots, X_N \rangle$ be an ordered tuple of options from $\mathcal{O}$. Then, I will define the *utility profile* of this tuple, given the options in $\mathcal{O}$—which I'll write '$\mathcal{U}_{\mathcal{O}}(\langle X_1, X_2, \ldots, X_N \rangle)$'—to be the set of tuples of utilities which are assigned to the options $\langle X_1, X_2, \ldots, X_N \rangle$, from any of the perspectives you would occupy after having selected one of the options $Y \in \mathcal{O}$.

$$\mathcal{U}_{\mathcal{O}}(\langle X_1, X_2, \ldots, X_N \rangle) \stackrel{\text{def}}{=} \{\langle \mathcal{U}_Y(X_1), \mathcal{U}_Y(X_2), \ldots, \mathcal{U}_Y(X_N) \rangle \mid Y \in \mathcal{O}\}$$

That is: the utility profile of $\langle X_1, X_2, \ldots, X_N \rangle$, in a decision with available options $\mathcal{O}$, is a set containing all of the opinions you could have about the utilities of the options $X_1, X_2, \ldots, X_N$, after having chosen any particular option $Y \in \mathcal{O}$.

The suggestion, then, is this: one decision, $\mathscr{D}$, is a subdecision of another, $\mathscr{D}^*$, iff there are two quadruples, $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ and $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$, such that (1) $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ characterizes $\mathscr{D}$ and $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$ characterizes $\mathscr{D}^*$; and (2) there is some way of associating each option $X \in \mathcal{O}$ with a unique option $X^* \in \mathcal{O}^*$ such that the utility profile of the options in $\mathscr{D}$ is exactly the same as the utility profile of the corresponding options in $\mathscr{D}^*$. If $\mathcal{O} = \{X_1, X_2, \ldots, X_N\}$, then the utility profile of the options in $\mathscr{D}$ will be $\mathcal{U}_{\mathcal{O}}(\langle X_1, X_2, \ldots, X_N \rangle)$. And, given some function, f, which associates each $X \in \mathcal{O}$ with a unique $X^* \in \mathcal{O}^*$, the utility profile of 'the corresponding options' in $\mathscr{D}^*$ will be $\mathcal{U}_{\mathcal{O}^*}(\langle f(X_1), f(X_2), \ldots, f(X_N) \rangle)$. If the former set is identical to the latter, then I say that $\mathscr{D}$

is a subdecision of $\mathscr{D}^*$.

Subdecision $\mathscr{D}$ is a *subdecision* of $\mathscr{D}^*$ iff there are two quadruples, $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ and $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$, such that:

(1) $\mathscr{D}$ is characterized by $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ and $\mathscr{D}^*$ is characterized by $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$; and

(2) there is an injection, f, from $\mathcal{O} = \{X_1, X_2, \ldots, X_N\}$ to $\mathcal{O}^*$, such that:

$$\mathcal{U}_{\mathcal{O}}(\langle X_1, X_2, \ldots, X_N \rangle) \;=\; \mathcal{U}_{\mathcal{O}^*}(\langle f(X_1), f(X_2), \ldots, f(X_N) \rangle)$$

Notice that the utility profile of $\langle X_1, X_2, \ldots, X_N \rangle$ does not vary with your option probabilities. So, if we accept **Subdecision**, we won't end up saying that the decision you face at the beginning of deliberation is different from the one you face at deliberation's end. Nor does **Subdecision** trivialize principles like **IIA**. Take a mundane example: the utilities of steak and chicken do not depend upon whether you order steak, chicken, or fish. Whatever item you select, the utility of steak is 5 utiles, and the utility of chicken is 10 utiles. Then the utility profile of ⟨*steak, chicken*⟩ will be $\{\langle 5, 10 \rangle\}$ whether or not fish is on the menu. So we may say fish is a genuinely *irrelevant* alternative, and that the decision between chicken and steak is a subdecision of the decision between chicken, steak, and fish. Then, **IIA** will say that, if it is not permissible to choose steak over chicken, it's not permissible to choose steak over chicken and fish.

Additionally, in the cases where CDT and EDT part ways, options on two different menus can share a utility profile. Suppose that, in Newcomb, we include an additional box, labeled '*O*', which is guaranteed to contain the same amount of money as *L*. If it was predicted that you'd take *L*, then there is \$100 in both *L* and *O* and \$110 in *M*. If it was predicted that you'd take either *O* or *M*, then there's \$10 in *M* and nothing in either *L* or *O*. As in the original Newcomb, these predictions are 90% reliable.[23] This additional option will not affect the utility profile of $\langle L, M \rangle$. For the perspective on the utilities of *L* and *M* which you'd have after learning that you'd chosen *O* is precisely the same as the perspective on the utilities of *L* and *M* you'd have after choosing *M*. So $\mathcal{U}_{\{L,M\}}(\langle L, M \rangle) = \mathcal{U}_{\{L,M,O\}}(\langle L, M \rangle) = \{\langle 90, 100 \rangle, \langle 10, 20 \rangle\}$. And we will be able to say that a decision between *L* and *M* is a subdecision of the decision between *L*, *M*, and *O*. Thus, *O* will be an *irrelevant* alternative, and **IIA** will tell us that, if *L* is impermissible to select in the original Newcomb, it is also impermissible to select when *O* is included on the menu of options.[24]

23. That is: conditional on your choosing either *M* or *O*, you're 90% sure that there's \$0 in *L* and *O* and \$10 in *M*; and, conditional on your choosing *L*, you're 90% sure that there's \$100 in *L* and *O* and \$110 in *M*.

24. *Objection*: Suppose that, conditional on choosing *O*, you are 100% sure that there's \$0 in *L* and

When I presented **IIA** above, I simply took it for granted that a decision between X and Y was a subdecision of a decision between X, Y, and Z. Now that we're being more careful, we should explicitly include this requirement. We'll then get the following principle:

IIA′ Suppose that X is an available option in the decision $\mathscr{D}$, that $\mathscr{D}$ is a subdecision of $\mathscr{D}^*$, and that X^* is the option corresponding to X in the decision $\mathscr{D}^*$. Then, if it is impermissible to choose X in $\mathscr{D}$, it is impermissible to choose X^* in $\mathscr{D}^*$.

If $\mathscr{D}$ is a subdecision of $\mathscr{D}^*$, then there will be some way of associating the options in $\mathscr{D}$ with the options in $\mathscr{D}^*$. And 'the option corresponding to X' will be whatever option in $\mathscr{D}^*$ we associate with X. (It could be that X^* is identical to X, but I think it's best if a principle like **IIA′** does not require us to say when an option in one decision is identical to an option in another decision. So I've formulated **Subdecision** in such a way as to remain neutral on questions of option-individuation.)

Similarly, we should more carefully state the principle **NEE** like this (see figure 4):

NEE′ Suppose that 1) $\mathscr{D}^*$ is a decision between the options X^*, Y^*, and Z^*; 2) both $\mathscr{D}_1$ and $\mathscr{D}_2$ are subdecisions of $\mathscr{D}^*$; 3) $\mathscr{D}_1$ is a decision between taking an option corresponding to X^*, 'X', and going on to face the decision $\mathscr{D}_2$; 4) $\mathscr{D}_2$ is a decision between options corresponding to Y^* and Z^* ('Y' and 'Z', respectively); and 5) if, in $\mathscr{D}_1$, you decide to forego X and face the decision $\mathscr{D}_2$, then you are certain to retain your beliefs, desires, and rationality while deciding between Y and Z. Then, if it is permissible to not choose X^* in $\mathscr{D}^*$, it is also permissible to not choose X in $\mathscr{D}_1$.

(Again, X^*, Y^*, and Z^* could be identical to X, Y, and Z, respectively. However, I've tried to formulate **Subdecision**, **IIA′**, and **NEE′** so that they don't require us to take a stand on questions of option-individuation.)

4 Escaping the Cycle

If we accept **Subdecision**, then a decision between boxes A and B will not count as a subdecision of a decision between all three boxes. In a decision between

O and \$10 in M. Then O will introduce a new potential post-choice perspective on $\langle L, M\rangle$, and the utility profile $\mathcal{U}_{\{L,M,O\}}(\langle L, M\rangle)$ will not be identical to $\mathcal{U}_{\{L,M\}}(\langle L, M\rangle)$. But O should still be treated as an irrelevant alternative, and you should still choose M once it is added. *Reply*: I agree that you should still choose M in this decision, and that, in some good sense of 'irrelevant', O is an irrelevant option (you certainly shouldn't choose it), but I don't take it to be an objection to **Subdecision** that it, together with **IIA**, does not tell us so. We can't expect weak principles like **IIA** to tell us *everything*; it is enough that they tell us something non-trivial.

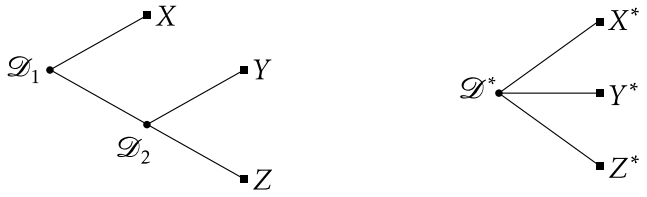


FIGURE 4: **NEE′** says that, if both $\mathscr{D}_1$ and $\mathscr{D}_2$ are subdecisions of $\mathscr{D}^*$, and it is permissible to not choose X^* in $\mathscr{D}^*$, then it is permissible to choose to leave behind the option corresponding to X^*, X, in $\mathscr{D}_1$ (so long, that is, as you are certain to retain your beliefs, desires, and rationality when choosing in $\mathscr{D}_2$).

just boxes A and B, the utility profile of $\langle A, B\rangle$ will only contains perspectives from which the utility of B exceeds the utility of A. That is: $\mathcal{U}_{\{A,B\}}(\langle A, B\rangle) = \{\langle 0, 70\rangle, \langle -70, 0\rangle\}$. Whereas, in a decision between A, B, and C, the utility profile of $\langle A, B\rangle$ will contain an additional perspective from which the utility of A exceeds the utility of B. That is: $\mathcal{U}_{\{A,B,C\}}(\langle A, B\rangle) = \{\langle 0, 70\rangle, \langle -70, 0\rangle, \langle 70, -70\rangle\}$. So, according to **Subdecision**, a decision between A and B is relevantly different when you have the additional option of taking box C. This means that, if we accept **Subdecision**, then **Minimal CDT** will not violate **IIA′** in UTILITY CYCLE.

For similar reasons, accepting **Subdecision** means that **Minimal CDT** will not violate **NEE′**. Consider the decision between taking box A and going on to choose between boxes B and C, $\{A, \{B, C\}\}$. In that decision, if you know that your future self will have the same beliefs and desires as your current self, and if you know that they will abide by **Minimal CDT**, then you know that, in a decision between B and C, they will choose C. So, at stage 1, the option of not taking box A corresponds to the option of taking box C in a decision between A, B, and C, $\{A, B, C\}$. But even so, $\{A, \{B, C\}\}$ will not count as a subdecision of $\{A, B, C\}$. For, in the decision $\{A, \{B, C\}\}$, the utility profile of $\langle A, \{B, C\}\rangle$ contains only perspectives on which the utility of A exceeds the utility of $\{B, C\}$: $\mathcal{U}_{\{A,\{B,C\}\}}(\langle A, \{B, C\}\rangle) = \{\langle 0, -70\rangle, \langle 70, 0\rangle\}$. (Since you are sure that not taking A will lead your future self to choose C, your probability distribution over states, conditional on you taking option $\{B, C\}$, is the same as your probability distribution over states, conditional on C, so $\mathcal{U}_{\{B,C\}}(A)$ will be equal to $\mathcal{U}_C(A)$.) However, in the decision $\{A, B, C\}$, the utility profile of $\langle A, C\rangle$ will contain an additional perspective from which the utility of C exceeds the utility of A: $\mathcal{U}_{\{A,B,C\}}(\langle A, C\rangle) = \{\langle 0, -70\rangle, \langle 70, 0\rangle, \langle -70, 70\rangle\}$. So, when you are asked to take box A or leave it, your decision is not a subdecision of the decision you face when you're asked to take either box A, B, or C. So we do not have a violation of **NEE′**. (For similar reasons, the decision between B and C at stage 2, $\{B, C\}$, will not count as a subdecision of the decision $\{A, B, C\}$.)

No amount of quibbling about subdecisions will change the fact that those

who abide by **Minimal CDT** will lose $60, on average, in the sequential decisions from §2.3, while those who are always indifferent between *A, B,* and *C* in a decision between any two, will break even, on average. But I think that causalists should accept and defend this consequence of their view. In the first place, they can offer a *tu quoque*: in *other* sequential decisions, evidentialists will end up predictably poorer than causalists.[25] More convincingly, they can object to using outcomes in *sequential* decisions to evaluate the rationality of agents who are incapable of binding their future selves to a certain course of action. The temporal parts of these agents are like separate agents, each facing their own, separate decisions, and incapable of coordinating their actions. The fact that such agents can be led to predictable ruin through a series of rational choices is just an intrapersonal tragedy of the commons.[26]

(We may think that intrapersonal tragedies of the commons are not possible, because we think that the rationality of later choices is importantly constrained in some way by which choices were made earlier, and for which reasons.[27] Whether that's so is an interesting debate, but it cross-cuts the debate between evidentialists and causalists. Causalists and evidentialists both have the option to affirm or deny that, at the beginning of a sequential decision, you should form the plan or the intention which is most choiceworthy, and that, *ceteris paribus*, rationality demands that you stick to that plan or follow through on that intention. If either affirms, they won't face these kinds of objections; if either denies, they will.)

5 Further Discussion

The question of when causalists should count one decision as a subdecision of another is interesting in its own right. But the discussion here bears on other, internecine causalist disputes. As I briefly mentioned in §1.3 above, there are decisions in which orthodox CDT's verdicts depend upon how likely you think you are to choose each available option. Some find CDT's verdicts about these cases objectionable,[28] and some have suggested heterodox causalist theories of rational choice to treat these cases.[29] An objection which has been raised to

25. See Wells (2019) and Ahmed (2020).

26. See Arntzenius et al. (2004) for further defense of this view, and see Meacham (2010) for a reply. See also Ahmed (2014b, §7.4.3) and Spencer (forthcominga, §5).

27. See, *e.g.*, McClennan (1990) and Bratman (1999).

28. See, *e.g.*, Richter (1984), Egan (2007), Briggs (2010), Wedgwood (2013), Ahmed (2014a), Spencer & Wells (2019), Gallow (2020, forthcoming), Spencer (forthcominga), and Podgorski (forthcoming).

29. See, e.g., Wedgwood (2013), Barnett (ms), Spencer (forthcomingb), Gallow (2020), and Podgorski (forthcoming).

$\mathcal{D}(Row\ Col)$	K_A	K_D
A	100	0
D	0	100

$\mathcal{P}(\ Row \mid Col)$	A	D
K_A	70%	25%
K_D	30%	75%

Table 3: Desires and Probabilities for Cake in Damascus. ('A' says that you go to Aleppo, 'D' says that you go to Damascus, 'K_A' says that it was predicted that you'd got to Aleppo, and 'K_D' says that it was predicted that you would go to Damascus.)

some of these heterodox theories is that they run afoul of the independence of irrelevant alternatives.[30] One important upshot of our discussion here is that this criticism is misplaced. Apparent violations of **IIA** arise in similar ways for orthodox CDT; and the solution I've proffered causalists is available to the heterodox and orthodox both. Moreover, while this solution allows the heterodox causalist theory I favor to *always* satisfy **IIA′** and **NEE′**, the same cannot be said for orthodox CDT.

Recall, in a *self-reinforcing* decision, choosing either option would give you the good news that your choice will make things better than the alternative would. That is, in a decision between X and Y, $\mathcal{U}_X(X) > \mathcal{U}_X(Y)$ and $\mathcal{U}_Y(Y) > \mathcal{U}_Y(X)$. For a concrete case like this, consider:[31]

> Cake in Damascus
> You must choose whether to go to Damascus or Aleppo. Yesterday, your fairy godmother made a prediction about which you would choose, and she left you cake in the predicted city. Her predictions are quite reliable, but she has a tendency to guess Damascus. Conditional on you going to Damascus, you're 75% sure that cake awaits in Damascus; whereas, conditional on you going to Aleppo, you're only 70% sure that cake awaits there. Getting cake is the only thing you care about.

Your desires and probabilities for Cake in Damascus are shown in table 3. As the reader may verify for themselves, in this decision, $\mathcal{U}_A(A) = 70 > 30 = \mathcal{U}_A(D)$, and $\mathcal{U}_D(D) = 75 > 25 = \mathcal{U}_D(A)$. So this is a self-reinforcing decision. Going to Damascus gives you the good news that cake likely awaits in Damascus. And going to Aleppo gives you the good news that cake likely awaits in Aleppo.

In Cake in Damascus, choosing either option would give you good news about what you are doing to make the world better. However, one of the options (going to Damascus) gives you *better* news about what you are doing to

30. See, *e.g.*, Bassett (2015) and the discussion in Wedgwood (2013) and Barnett (ms).

31. Similar cases are discussed in Hunter & Richter (1978) and Hare & Hedden (2016). ('Cake in Damascus' is a reference to Gibbard & Harper (1978)'s 'Death in Damascus'.)

make things better. Orthodox CDT says that which option you should choose depends upon your option probabilities. I disagree. I say you should choose the option which would give you the best news about what you're doing to bring about your desired ends. I won't discuss this theory here—see Gallow (2020) for details. The important point for present purposes is just this: according to this theory, to determine which of X and Y is more choiceworthy, you must look at the quantities $\mathcal{U}_X(X), \mathcal{U}_Y(X), \mathcal{U}_X(Y)$, and $\mathcal{U}_Y(Y)$—that is, you must look at exactly the values which appear in the *utility profile* of $\langle X, Y\rangle$. Moreover, if we accept **Subdecision**, the theory of rational choice I favor will always satisfy **IIA′** and **NEE′**. Thus, there is at least one theory of rational choice which allows us to reconcile **Minimal CDT** with **IIA′** and **NEE′**.

Orthodox CDT has a harder time satisfying **IIA′** and **NEE′** in general. Suppose that you always begin deliberation thinking that you are equally likely to select each of the available options. Then, in self-reinforcing decisions, orthodox CDT will violate both **IIA′** and **NEE′**, even given **Subdecision**.[32]

For illustration, return to Cake in Damascus. Suppose that you always begin deliberation by distributing your option probabilities evenly. Then, at the beginning of deliberation, you will assign A of utility of 42.5 and D a utility of 52.5: $\mathcal{U}(A) = 42.5$ and $\mathcal{U}(D) = 52.5$. So orthodox CDT will say that A is impermissible and that D is required. It will not change this verdict as you resolve to go to Damascus and raise your option probability for D to 100%. But now suppose we introduce an additional option: a new road to Aleppo has opened up. This road doesn't differ from the original road in any respect that you care about. You now face a decision between A (going to Aleppo *via* the original road), A^* (going to Aleppo *via* the new road), and D (going to Damascus). If you again begin deliberation by distributing your option probabilities evenly, then you will assign each of A and A^* a utility of 55, and D a utility of 45: $\mathcal{U}(A) = \mathcal{U}(A^*) = 55$ and $\mathcal{U}(D) = 45$. So CDT will say that A is permissible. It will continue to say this as you resolve to choose A (or A^*) and raise your option probability for A (or A^*) to 100%. So, in your decision between A and D, CDT says that A is impermissible. But, in your decision between $A, D,$ and A^*, it says that A is permissible. Since the options $\langle A, D\rangle$ have the same utility profile in both of these decisions, your choice dispositions will have violated **IIA′** (assuming **Subdecision**).

Again, we could attempt to say that your different option probabilities are enough to make A and D in the second decision importantly different options than they were in the first. But, again, this trivializes **IIA′**—if you always begin deliberation by giving positive probability to each available option, then **IIA′** will never apply. And, again, we could attempt to say that your differ-

32. Below, I discuss orthodox CDT's violation of the independence of irrelevant alternatives. For a discussion of its violation of normal-form extensive-form equivalence, see Joyce (2018).

ent (unconditional) state probabilities are enough to make these two decisions different. But, again, this would have the uncomfortable consequence that deliberating about which option to choose can end up changing the decision you face—recall the discussion in §3.2. There is also always the possibility of simply rejecting the principle **IIA′**; though, in my view, we should want to hold on to this plausible principle if we can.

Parenthetically, heterodox causalist theories like mine have been criticized for violating the independence of irrelevant alternatives.[33] It is therefore worth noting that the apparent violations of the independence of irrelevant alternatives are not unique to the heterodox; orthodox CDT also appears to violate the principle, and in similar ways. Moreover, while orthodox CDT has additional difficulty complying with **IIA′** in cases like Cake in Damascus, my theory of rational choice will never violate **IIA′**, so long as we assume **Subdecision**.

6 Conclusion

In summation, decisions like Utility Cycle afford us three arguments against **Minimal CDT**. I've presented these arguments and offered causalists three replies. The first two objections: in Utility Cycle, **Minimal CDT** appears to violate weak versions of the *independence of irrelevant alternatives* and *normal-form extensive-form equivalence*. In response to these objections, I've counseled causalists to specify when one decision is a subdecision of another partly in terms of options' *utility profiles*. This prevents the carefully-formulated principles **IIA′** and **NEE′** from being trivialized and it prevents **Minimal CDT** from violating those principles in Utility Cycle.

The final objection: in sequential decisions, those who abide by **Minimal CDT** will end up predictably poorer than those who follow EDT, even when they have exactly the same amount of money in front of them, sitting in exactly the same place. In response to this objection, I've counseled causalists to accept this consequence of their view as an unfortunate intrapersonal tragedy of the commons—avoidable by those lucky agents capable of binding their future selves. Accepting this consequence means that those of us who like diachronic Dutch book arguments will have to be much more careful about how we formulate them. Accepting this consequence also means rejecting another recent argument against EDT. Wells (2019) argues against EDT with a sequential decision in which evidentialists make predictably less money than causalists. Notice, however, that in the sequential decision Pay or A, a parrallel argument could be mounted against anyone who abides by **Minimal CDT**. Such a person will pay \$60 at stage 1 and go on to choose *B* at stage 2. They will therefore

33. See, for instance, the discussion in Wedgwood (2013), Bassett (2015), and Barnett (ms).

be certain to make $60 less than an evidentialists who doesn't pay at stage 1, rejects *A* at stage 2, and goes on to choose *B* at stage 3. They will make $60 less than the evidentialist *no matter which prediction was made*. So endorsing an argument like Wells's means abandoning **Minimal CDT**.[34]

34. See also Ahmed (2020)'s criticism of Wells (2019).

References

Ahmed, Arif. 2012. "Push the Button." *Philosophy of Science*, vol. 79 (3): 386–395. [7]

—. 2014a. "Dicing with Death." *Analysis*, vol. 74 (4): 587–592. [23]

—. 2014b. *Evidence, Decision and Causality*. Cambridge University Press, Cambridge, UK. [3], [23]

—. 2020. "Equal Opportunity in Newcomb's Problem and Elsewhere." *Mind*, vol. 129 (515): 867–886. [23], [27]

Armendt, Brad. 2019. "Causal Decision Theory and Decision Instability." *The Journal of Philosophy*, vol. 116: 263–277. [6]

Arntzenius, Frank. 2008. "No regrets, or: Edith Piaf revamps decision theory." *Erkenntnis*, vol. 68: 277–297. [4], [7]

Arntzenius, Frank, John Hawthorne & Adam Elga. 2004. "Bayesianism, Infinite Decisions, and Binding." *Mind*, vol. 113 (450): 251–283. [23]

Bales, Adam. 2018. "Richness and Rationality: Causal Decision Theory and the WAR Argument." *Synthese*, vol. 195 (259–67). [12]

Barnett, David James. ms. "Graded Ratifiability." [7], [23], [24], [26]

Bassett, Robert. 2015. "A Critique of Benchmark Theory." *Synthese*, vol. 192 (1): 241–267. [24], [26]

Bratman, Michael. 1999. "Toxin, Temptation, and the Stability of Intention." In *Faces of Intention*. Cambridge University Press. [23]

Briggs, R. A. 2010. "Decision-Theoretic Paradoxes as Voting Paradoxes." *The Philosophical Review*, vol. 119 (1): 1–30. [23]

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review*, vol. 116 (1): 93–114. [6], [23]

Gallow, J. Dmitri. 2020. "The Causal Decision Theorist's Guide to Managing the News." *The Journal of Philosophy*, vol. 117 (3): 117–149. [7], [13], [23], [25]

—. forthcoming. "Riches and Rationality." *Australasian Journal of Philosophy*. [23]

Gibbard, Allan & William L. Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, A. Hooker, J.J. Leach & E.F. McClennan, editors, 125–162. D. Reidel, Dordrecht. [6], [12], [24]

Hare, Caspar & Brian Hedden. 2016. "Self-Reinforcing and Self-Frustrating Decisions." *Noûs*, vol. 50 (3): 604–628. [6], [7], [24]

Harper, William. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis*, vol. 24: 25–36. [6], [14]

Hunter, Daniel & Reed Richter. 1978. "Counterfactuals and Newcomb's Paradox." *Synthese*, vol. 39 (2): 249–261. [24]

Jeffrey, Richard. 1965. *The Logic of Decision*. McGraw-Hill, New York. [3]

—. 2004. *Subjective Probability: the Real Thing*. Cambridge University Press, Cambridge, UK. [3]

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge. [5], [12]

—. 2012. "Regret and instability in causal decision theory." *Synthese*, vol. 187 (1): 123–145. [6], [7]

—. 2018. "Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems." In *Newcomb's Problem*, Arif Ahmed, editor. Oxford University Press, Oxford. [7], [25]

Lewis, David K. 1981a. "Causal Decision Theory." *Australasian Journal of Philosophy*, vol. 59 (1): 5–30. [2], [5]

—. 1981b. "'Why ain'cha rich?'" *Noûs*, vol. 15 (3): 377–380. [12]

McClennan, Edward. 1990. *Rationality and Dynamic Choice*. Cambridge University Press, Cambridge. [23]

Meacham, Christopher J.G. 2010. "Binding and Its Consequences." *Philosophical Studies*, vol. 149 (1): 49–71. [23]

Pettit, Philip. 2010. "Deliberation and Decision." In *A Companion to the Philosophy of Action*, Timothy O'Connor & Canstantine Sandis, editors, chap. 32. Blackwell Publishing. doi:10.1111/b.9781405187350.2010.00034.x. [17]

Podgorski, Abelard. forthcoming. "Tournament Decision Theory." *Noûs*. [7], [9], [23]

Rabinowicz, Wlodek. 2009. "Letters from Long Ago: On Causal Decision Theory and Centered Chances." In *Logic, Ethics, and All That Jazz—Essays in Honour of Jordan Howard Sobel*, L-G. Johansson, editor, vol. 56, 247–273. Uppsala Philosophical Studies. [5]

Rabinowicz, Włodzimierz. 1982. "Two Causation Decision Theories: Lewis vs Sobel." In *Philosophical Essays Dedicated to Lennart Åqvist on His Fiftieth Birthday*, Tom Pauli, editor, vol. 34, 299–321. Uppsala Philosophical Studies, Uppsala. [5]

Ray, Paramesh. 1973. "Independence of Irrelevant Alternatives." *Econometrica*, vol. 41 (5): 987–991. [9]

Richter, Reed. 1984. "Rationality Revisited." *Australasian Journal of Philosophy*, vol. 62 (4): 392–403. [6], [23]

Sen, Amartya. 1993. "Internal Consistency of Choice." *Econometrica*, vol. 61 (3): 495–521. [14]

Skyrms, Brian. 1982. "Causal Decision Theory." *Journal of Philosophy*, vol. 79 (11): 695–711. [5]

—. 1990. *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge, ma. [7]

Sobel, Jordan Howard. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge University Press, Cambridge. [5]

Spencer, Jack. forthcominga. "An Argument Against Causal Decision Theory." *Analysis*. [23]

—. forthcomingb. "Rational Monism and Rational Pluralism." *Philosophical Studies*. [7], [23]

Spencer, Jack & Ian Wells. 2019. "Why Take Both Boxes?" *Philosophy and Phenomenological Research*, vol. 99 (1): 27–48. [23]

Wedgwood, Ralph. 2013. "Gandalf's solution to the Newcomb Problem." *Synthese*, vol. 190 (14): 2643–2675. [7], [14], [23], [24], [26]

Weirich, Paul. 1985. "Decision Instability." *Australasian Journal of Philosophy*, vol. 63 (4): 465–478. [6]

Wells, Ian. 2019. "Equal Opportunity and Newcomb's Problem." *Mind*, vol. 128 (510): 429–457. [5], [12], [23], [26], [27]

Williamson, Timothy Luke. forthcoming. "Causal Decision Theory is Safe From Psychopaths." *Erkenntnis*. [6]