Escaping the Cycle

Abstract

I present a decision problem 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 in this case by carefully specifying when two options in different decision problems are similar enough to be counted as the same.

As I'll understand it here, the *independence of irrelevant alternatives* (IIA) says that adding an additional 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. It 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 choice to either have X or go on to choose amongst the other options, it is permissible to choose to leave X behind.[1] If, given a choice between chicken, steak, and fish, it's permissible for you to order the steak, then, given a choice between the fish and a choice 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.

Here, I'll present a decision problem—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

Word count: 8,908

[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.

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 decision 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 compare two decision problems in which you are given the same options. So in order to show that CDT violates IIA or NEE, we must make some assumptions about what it takes for two options, in two different decision problems, to count as *the same*. Given a natural assumption about what makes two options the same, CDT will violate IIA and NEE. But I'll suggest an alternative approach to causalists which allows them to satisfy the principles (§§3–4).

The question of when causalists should count two options, in different decision problems, as being effectively the same or importantly different is interesting in its own right. But the discussion here bears on other, internecine causalist disputes. There is a class of decision problems 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,[2] and some have suggested heterodox causalist theories of rational choice to treat these cases.[3] An objection which has been raised to some of these heterodox theories is that they appear to run afoul of the IIA.[4] One important upshot of my discussion here is that this criticism is misplaced. Apparent violations of IIA arise in similar ways for orthodox CDT; and the solution I'll proffer causalists here 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 (§5).

1 Causal Decision Theory

1.1 Desire. I will assume that, when you face a decision, 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.[5] Exactly one of the K_i obtains, though you know not which; nor are you in any position to influence which obtains. Though you do not know which K_i obtains, you do have opinions, represented with a probability function, Pr, defined over both $\mathcal{O}$ and $\mathcal{K}$. Finally, we can represent

[2] See, *e.g.*, Richter (1984), Egan (2007), Briggs (2010), Wedgwood (2013), Ahmed (2014a), Spencer & Wells (2017), and Spencer (msa).

[3] See, e.g., Wedgwood (2013), Barnett (ms), Spencer (msb), and [author].

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

[5] 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.

$$\begin{array}{cc} \mathcal{D}(\mathit{Row}\ \mathit{Col}) & \begin{array}{cc} K_L & K_M \end{array} \\ \begin{array}{c} L \\ M \end{array} & \begin{bmatrix} 100 & 0 \\ 110 & 10 \end{bmatrix} \end{array} \qquad \begin{array}{cc} \Pr(\mathit{Row} \mid \mathit{Col}) & \begin{array}{cc} L & M \end{array} \\ \begin{array}{c} K_L \\ K_M \end{array} & \begin{bmatrix} 90\% & 10\% \\ 10\% & 90\% \end{bmatrix} \end{array}$$

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.

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 $X \in \mathcal{O}$,

$$\mathcal{D}(X) = \sum_K \Pr(K \mid X) \cdot \mathcal{D}(XK)$$

$\mathcal{D}(X)$ 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.[6] 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 problem 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 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.

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

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

$$\begin{aligned} \mathcal{D}(L) &= \Pr(K_L \mid L) \cdot \mathcal{D}(LK_L) + \Pr(K_M \mid L) \cdot \mathcal{D}(LK_M) \\ &= 90\% \cdot 100 + 10\% \cdot 0 \\ &= 90 \\ \text{while} \quad \mathcal{D}(M) &= \Pr(K_L \mid M) \cdot \mathcal{D}(MK_L) + \Pr(K_M \mid M) \cdot \mathcal{D}(MK_M) \\ &= 10\% \cdot 110 + 90\% \cdot 10 \\ &= 20 \end{aligned}$$

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.[7]

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 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

[7] See ARNTZENIUS (2008)

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.[8]

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 just choose the option which you'd be happiest to learn that you've 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

$$\mathcal{U}(X) \overset{\text{def}}{=} \sum_{K} \Pr(K) \cdot \mathcal{D}(XK)$$

tells us how desirable you expect choosing X to *make* the world.[9]

The difference between $\mathcal{D}$ and $\mathcal{U}$ is that, in $\mathcal{D}$, we conditioned the probability function Pr 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 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

[8] See WELLS (forthcoming)

[9] This is SKYRMS's formulation of causal decision theory. There are alternatives—see, *e.g.*, LEWIS (1981a) and JOYCE (1999). The differences between these version of CDT won't make a difference to anything I have to say here.

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 *immediately* after choosing Y, before learning anything else.) From this perspective, you would have learned that you've chosen Y, so your probability for each state K would be $\Pr(K \mid Y)$, and

$$\mathcal{U}_Y(X) \overset{\text{def}}{=} \sum_K \Pr(K \mid Y) \cdot \mathcal{D}(XK)$$

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.[10]

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

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

> SELF-UNDERMINING CHOICE
> 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 CHOICE
> 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 choice, 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 choice, if you disregard its advice and do what it deemed irrational, CDT will change its mind and call you rational for doing so.[11]

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 choices won't be relevant to the arguments

[10] In much of what follows, 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 $\mathcal{D}(Row\ Col)$ by the matrix $\Pr(Row|Col)$. 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 in the body can then be used to easily calculate the unconditional utilities, $\mathcal{U}(Row)$.

[11] *Cf.* GIBBARD & HARPER (1978), RICHTER (1984), WEIRICH (1985), HARPER (1986), EGAN et al. (2005), JOYCE (2012), HARE & HEDDEN (2016), ARMENDT (2019), and WILLIAMSON (forthcoming).

against CDT which I'll introduce below.[12] For those arguments, I need only appeal to the following, minimal commitment of CDT, which is also endorsed by heterodox causalists like myself:[13]

CDT In a choice 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 the boldface **CDT**. The latter is strictly weaker than the former; **CDT** only applies in choices between two options, where the choice is neither self-undermining nor self-reinforcing.) In Newcomb, **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 CDT

Consider the following decision problem:[14]

> 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 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:

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

[12] Though I'll return to these kinds of choices in §6.

[13] **CDT** is accepted by Wedgwood (2013), Barnett (ms), Spencer (msb), and [author].

[14] *Cf.* Ahmed (2012) and Hare & Hedden (2016).

$$\begin{array}{c|ccc} \mathcal{D}(Row\;Col) & K_A & K_B & K_C \\ \hline A & 0 & -100 & 100 \\ B & 100 & 0 & -100 \\ C & -100 & 100 & 0 \end{array} \qquad \begin{array}{c|ccc} \Pr(\,Row\,|Col) & A & B & C \\ \hline K_A & 80\% & 10\% & 10\% \\ K_B & 10\% & 80\% & 10\% \\ K_C & 10\% & 10\% & 80\% \end{array}$$

TABLE 2: Desires and Probabilities for UTILITY CYCLE

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 are given a choice 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:

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

No matter the value of *a*, *B* will have a higher utility than *A*. So **CDT** says that, in a choice 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 are given a choice 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:

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

Again, no matter the value of *b*, the utility of *C* will exceed the utility of *B*. So **CDT** says that, in a choice between *B* and *C*, *C* is required and *B* is impermissible. Similarly, if *B* is removed from the menu, and you are given a choice 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 **CDT** says that, in a choice 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 **CDT** appears to lead to a violation of a principle known as *the independence* *of irrelevant alternatives* (or just '**IIA**').

IIA: If, given a choice between X and Y, Y is not permissible, then, given a choice between X, Y, and Z, Y is not permissible.

According to **CDT**, *every* option in Utility Cycle is impermissible in a one-on-one choice with some alternative. So, if *some* option is permissible,[15] we will have a violation of **IIA**. For illustration: suppose that A is a permissible choice in Utility Cycle. By **CDT**, given a choice between A and B, A is impermissible. So A is not a permissible choice when you are presented with the restricted menu $\{A, B\}$, but it *is* a permissible choice when you are presented with the larger menu $\{A, B, C\}$. And this contradicts **IIA**. The same goes if we say that B or C is permissible instead. For **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 shows that **CDT** violates a weak principle of *normal-form extensive-form equivalence* (or just '**NEE**').

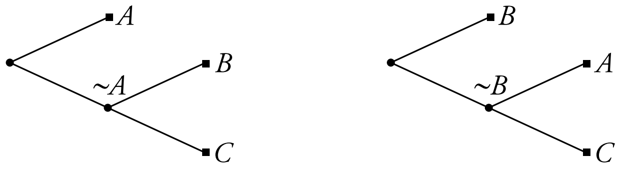
NEE: If you are certain to remain rational and your beliefs and desires are certain to not change, then, if it is permissible to not choose X when given a choice between X, Y, and Z, then, given a choice between X and going on to choose between Y and Z, it is permissible to not choose X.

The antecedent of **NEE** is important. Suppose you think that your beliefs or desires might change after choosing $\sim X$ and before choosing between Y and Z. Then, it may be rational to choose X now in order to take the decision out of the hands of your not-entirely-trustworthy future self. Likewise, if you fear that your future self will not choose rationally, this could give 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 choices:

> A or ∼A
> Money was distributed between boxes A, B, and C as in Utility Cycle. At stage 1, you are given a choice to either take box A or to not. If you take box A, then you receive its contents. If you don't take A, then at stage 2, you choose between B and C. (See figure 1a.) You are certain to retain your beliefs, desires, and rationality throughout.

[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.



(a) A or ∼A (b) B or ∼B

Figure 1

> B or ∼B
> Money was distributed between boxes A, B, and C as in Utility Cycle. At stage 1, you are given a choice to either take box B or to not. If you take box B, then you receive its contents. If you don't take B, then at stage 2, you choose between A and C. (See figure 1b.) You are certain to retain your beliefs, desires, and rationality throughout.

Assume you know that you abide **CDT**, and that you will continue to do so throughout any sequential decisions. Then, in A or ∼A, if you choose $\sim A$ at stage 1, at stage 2, you will choose C, and you know this at stage 1. So, at stage 1, you face a choice between A and C. So A is required at stage 1. In B or ∼B, if you choose $\sim B$ at stage 1, then, at stage 2, you will choose A, and you know this at stage 1. So, at stage 1, you face a choice between B and A. So B is required at stage 1.

We can now show that, assuming *some* option is permissible, **CDT** violates **NEE** in Utility Cycle. For, given the choice between A, B, and C, B is either permissible or it is not. Suppose it is. Then, **NEE** says that $\sim A$ is permissible in A or ∼A. **CDT** on the other hand, says that $\sim A$ is impermissible, contradicting **NEE**. Suppose on the other hand that B is impermissible. Then, it is permissible to not choose B. In that case, **NEE** says that $\sim B$ is permissible in B or ∼B. **CDT**, on the other hand, says that $\sim B$ is impermissible, contradicting **NEE**. Either way, **CDT** contradicts **NEE**.

2.3 Predictable Long-run Poverty.

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 choice. (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 sixty dollars.' (You are certain that they will take this decision to be final, they will take the box you decide upon, 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 staying with A and changing to B and losing sixty dollars. If a is your probability for taking A, then the utilities of the available options are:

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

In this new decision, switching to B will have a higher utility than staying with A, no matter whether you take A or switch to B. So **CDT** says to hand your assistant sixty dollars 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. No matter which box you initially select, the assistant will be able to offer you a trade for another box, at a cost of $60, which you will see as favorable, as long as you abide by **CDT**. There's no initial selection which will prevent your future self from making the trade.

Note that, if you make the trade, 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 decision over and over again, with your assistant offering the trade each time, you will lose $60 on average. You could have instead broke even on average, if only you'd refused the assistant's trade.

Causalists are used to making less money in certain decision problems. 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 when given a choice between the two would never pay to switch, and they would predictably end up making more money in the long run.

CDT will advise you to pay to have the options presented to you in a certain order—even when you're certain to retain your beliefs, desires, and rationality

[16] See, *e.g.*, GIBBARD & HARPER (1978), LEWIS (1981b), JOYCE (1999), BALES (2018), and WELLS (forthcoming).

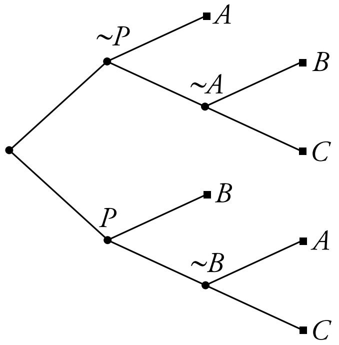


Figure 2: Pay or A

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, 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 2.)

If you know that you abide **CDT**, you will choose A in A or ~A. So, if you don't pay, you will end up choosing A. If you abide **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 choice 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, **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 decide 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 when given a choice between any two would predictably make more money when facing exactly the same choice in exactly the same circumstances. The series of choices advised by **CDT**—pay, then take B—are *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 Options

These consequences of **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 **CDT**. However, I want to urge caution. Though I reject orthodox causal decision theory, I believe that the weaker claim **CDT** is correct, and I accept what it says about

Utility Cycle.[17] Defenders of **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: taking box A is an importantly different (and worse) option when it appears on the menu $\{A, B\}$ than it is when it appears on the menu $\{A, B, C\}$.

3.1 Individuating Options. When thinking about principles like **IIA** and **NEE**, it is important to be clear about when two options, in two different decision problems, are similar enough to be counted as effectively the same option.[20] Some apparent counterexamples to the principle are not genuine counterexamples, because they conflate importantly different options. 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).[21] 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 beer and soda, and it makes them relevantly different options.

When I argued that **CDT** violated **IIA** above, I was implicitly assuming that taking box A was the same option when it appeared on the menu $\{A, B\}$ as it was when it appeared on the menu $\{A, B, C\}$. In general terms, I was assuming that X and X^*, in two different decision problems, are the same iff you desire them to the same degree, in every possible state of nature, and your subjective probability distribution over states, conditional on X, is the same as your subjective probability distribution over states, conditional on X^*. Call this 'the simple view' of option individuation.

The Simple View Options X and X^*, in two different decision problems, are *the*

[17] That is to say: I accept what CDT says about the choice between any two options in Utility Cycle. In a choice 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 [author] for details.

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

[19] *Cf.* Harper (1986).

[20] To say that options are the same is not to say that they are numerically identical. If X and X^* are options in different decision problems, then they *ipso facto* differ; by Leibniz's Law, they are distinct. Nonetheless, they may be *the same*—they may be similar in all respects which are relevant to choice. It is the latter notion of sameness which is my focus here; when I talk about 'option individuation', I am talking about when to say two options are the same, and not when to say that they are identical.

[21] *Cf.* Sen (1993).

same iff, for each state of nature K:

(a) $\mathcal{D}(XK) = \mathcal{D}(X^*K)$; and

(b) $\Pr(K \mid X) = \Pr(K \mid X^*)$.[22]

This is a natural and plausible way of saying when two options are the same. According to it, choosing beer over soda is relevantly different from choosing beer over soda and whiskey—for, when you are offered whiskey as an alternative, this changes your opinions about which signal beer will send, which changes either the degree to which you desire choosing beer in some state of nature, or your opinions about how likely some state is, or both.[23]

Some causalists may wish to say that your option probabilities also play an important role in determining when options are the same. They may wish to say that X and X^* are the same iff: a) $\mathcal{D}(XK) = \mathcal{D}(X^*K)$, for each state of nature K; b) $\Pr(K \mid X) = \Pr(K \mid X^*)$, for each K; and c) $\Pr(X) = \Pr(X^*)$. This would reconcile **CDT** with **IIA** and **NEE**, but at the price of trivializing the latter. Given a menu of options $\{X, Y\}$, you will necessarily have $\Pr(X) + \Pr(Y) = 1$. Then, given the larger menu $\{X, Y, Z\}$, the only way for X and Y to remain the same options would be for $\Pr(Z)$ to be zero. So long as you leave open that you'll select each available option, you'll never be presented with the same options on a larger menu, and principles like **IIA** and **NEE** will impose no constraint at all.

Alternatively, we may wish to say that your (unconditional) state probabilities help to determine when options are the same. That is, we may suggest: X and X^*, in two different decision problems, are the same iff, for each K: a) $\mathcal{D}(XK) = \mathcal{D}(X^*K)$; b) $\Pr(K \mid X) = \Pr(K \mid X^*)$; and c) K has the same unconditional probability, $\Pr(K)$, in both decision problems. This suggestion does not trivialize **IIA** and **NEE**, though it has other undesirable consequences. For 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, $\Pr(K) = \sum_X \Pr(K \mid X) \cdot \Pr(X)$. And note that, while your conditional probabilities $\Pr(K \mid X)$ are fixed, as you deliberate about what to do, your option probabilities, $\Pr(X)$, will change. Let us narrow our attention to the kinds of cases in which EDT and CDT disagree—cases in which $\Pr(K \mid X) \neq \Pr(K \mid Y)$, for some $X \neq Y$. Call these the 'interesting' cases. In the interesting cases, changes in your option probabilities will lead to changes in your

[22] To be clear: when I write equations like '$\mathcal{D}(XK) = \mathcal{D}(X^*K)$' and '$\Pr(K \mid X) = \Pr(K \mid X^*)$', the desire and probability functions on the left should be understood to be the desire and probability functions from the first decision problem, and those on the right should be understood to be the desire and probability functions from the second decision problem.

[23] Which we say is changed will depend upon what we say the states of nature are when we specify the decision problem. Since, in this decision problem, states of nature are probabilistically independent of acts, changes in your opinions about how likely some state is will violate condition (b) of the simple view.

(unconditional) state probabilities. So, on this suggestion, the very act of deliberating about what to choose changes the options between which you are choosing. This is odd. We normally think that the options about which you should be deliberating are the possible objects of choice for you. And, according to this proposal for individuating options, this won't be so in interesting cases. If you were to resolve to choose X, you would give yourself the evidence that you will choose X; so your option probability $\Pr(X)$ would go up, and your unconditional probability distribution over states would change. So the option you would end up choosing would be relevantly different from the one about which you were initially deliberating. Also, note that, if we individuate decision problems partly in terms of the options available to you, then this method of individuating options would mean that, in interesting cases, it is impossible to make up your mind about what to do in an interesting decision problem. Making up your mind would change the decision problem you face. (The same considerations apply to the suggestion to individuate options in terms of your option probabilities.)

So I don't think that we should appeal to your option probabilities or your state probabilities as a way of distinguishing the option of choosing box A from the menu $\{A, B\}$ and the option of choosing box A from the menu $\{A, B, C\}$. Nonetheless, I think that we *should* distinguish these two options. In the following, I'll offer the causalist a different account of when two options are the same (§3.2). I will then explain how this account allows causalists to dispute the charge that they violate **IIA** and **NEE** in Utility Cycle (§4).

3.2 Utility Profiles. If the probabilities assigned to states were constant throughout deliberation—if they did not vary with your option probabilities—then causalists could say: options X and X^* are the same iff $\mathcal{D}(XK) = \mathcal{D}(X^*K)$, for each state K, and you have the same probability distribution over states. From these quantities, we can calculate the utilities of the options X and X^*; so, if these quantities are the same, X and X^* will have the same utilities, and neither these utilities nor the utilities of the other options would change over the course of deliberation.

In many decision problems, the probabilities assigned to states are constant throughout deliberation, so that, as you make up your mind about what to do, your utilities do not change. Nonetheless, this does not hold in general—and it does not hold in the kinds of cases in which EDT and CDT disagree. We should want to individuate options in such a way that the available options remain the same throughout deliberation, so we should want to individuate them in terms of a property which does not change during deliberation. One property like this is the conditional probability of each state, K, given each option X, $\Pr(K \mid X)$. This is one reason why the simple view is so natural.

But there are other important properties of a decision problem which do not change over the course of deliberation. In particular: the utility you *would* assign to an option, *were* you to choose any of the options—that is, the values $\mathcal{U}_Y(X)$, for each Y—do not change as you deliberate. And my suggestion to causalists is that they individuate an option, X, partly in terms of these quantities. If this is done carelessly, it can end up trivializing principles like **IIA** and **NEE**. Given a choice between X and Y, there are two different potential post-choice perspectives from which to evaluate the utility of X: $\mathcal{U}_X(X)$ and $\mathcal{U}_Y(X)$. Given a choice between X, Y, and Z, there are *three*: $\mathcal{U}_X(X)$, $\mathcal{U}_Y(X)$, and $\mathcal{U}_Z(X)$. If this is enough to make X count as a different option, then it will be impossible for the same option to appear on two different menus, and **IIA** will impose no constraint at all.

Let us proceed more carefully. Suppose you face a decision problem with the menu of options **M**, and an option $X \in \mathbf{M}$. Let us define X's *utility profile*, on the menu **M**—which I'll write '$\mathcal{U}_{\mathbf{M}}(X)$'—to be the set of utilities which are assigned to X, from the perspective you'd occupy after having selected any of the options $Y \in \mathbf{M}$.

$$\mathcal{U}_{\mathbf{M}}(X) \overset{\text{def}}{=} \{\mathcal{U}_Y(X) \mid Y \in \mathbf{M}\}$$

Notice that X's utility profile does not vary with your option probabilities. So, if we individuate options in terms of their utility profiles, the available options will not change during deliberation. Nor does individuating options in terms of their utility profiles trivialize principles like **IIA**. Take a mundane example: the utility of steak does not depend upon whether you order steak, chicken, or fish. So we may say that steak is the same option whether it's on a menu with chicken or on a menu with chicken and fish. Then, **IIA** will say that, if it is not permissible to choose steak from the first menu, it's not permissible to choose it from the second menu, either.

Additionally, in the interesting 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 'L^*', 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 L^* and \$110 in M. If it was predicted that you'd take either L^* or M, then there's \$10 in M and nothing in either L or L^*. As in the original NEWCOMB, these predictions are 90% reliable.[24] This additional option will not affect the utility profiles of either L or M, so we will say that L and M are the same options after L^* is added, and **IIA** will entail that, if L is impermissible to select in the original NEWCOMB, it is also impermissible to select when L^* is included on the menu of options.[25]

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

[25] *Objection*: Suppose that, conditional on choosing L^*, you are 100% sure that there's \$0 in L and L^* and \$10 in M. Then including L^* *will* change the utility profiles of L and M. But L^* should

More carefully, I recommend the following way of individuating options:

Same Option Given two different decision problems, with the menus of options $\mathbf{M}$ and $\mathbf{M}^*$, $X \in \mathbf{M}$ and $X^* \in \mathbf{M}^*$ are *the same* iff, for each state of nature K:

(a) $\mathcal{D}(XK) = \mathcal{D}(X^*K)$;

(b) $\Pr(K \mid X) = \Pr(K \mid X^*)$; and

(c) $\mathcal{U}_{\mathbf{M}}(X) = \mathcal{U}_{\mathbf{M}^*}(X^*)$.

4 Escaping the Cycle

If **Same Option** is accepted, then A on the restricted menu $\mathbf{M} = \{A, B\}$ will be a different (and worse) option than A on the expanded menu $\mathbf{M}^* = \{A, B, C\}$. On the restricted menu, A's utility profile contains no positive values, $\mathcal{U}_{\mathbf{M}}(A) = \{-70, 0\}$. Whereas, on the expanded menu, A's utility profile contains positive values, $\mathcal{U}_{\mathbf{M}^*}(A) = \{-70, 0, 70\}$. So $\mathcal{U}_{\mathbf{M}}(A) \neq \mathcal{U}_{\mathbf{M}^*}(A)$, and the options are not the same, according to **Same Option**. This means that, if options are individuated as **Same Option** dictates, then **CDT** does not violate **IIA** in Utility Cycle. Though A is impermissible in a choice between A and B, *that very option* is not impermissible in a choice between A, B, and C. Including the additional option C changes A's utility profile, making it a different option.

For similar reasons, individuating options with **Same Option** means that **CDT** does not violate **NEE**. For, on the menu $\{A, {\sim}A\}$, A's utility profile is $\{0, 70\}$. (Since you are sure that choosing ${\sim}A$ will lead your future self to choose C, your probability distribution over states, conditional on ${\sim}A$, is the same as your probability distribution over states, conditional on C, so $\mathcal{U}_{\sim A}(A)$ is equal to $\mathcal{U}_C(A)$.) But, again, on the full menu $\{A, B, C\}$, A's utility profile is $\{-70, 0, 70\}$. So, when you are asked to take box A or leave it, you are being offered a different (and better) option than you are offered when you're given a choice between A, B, or C, and we do not have a violation of **NEE**. Likewise, if we accept **Same Option**, then we will object to my earlier claim that those who abide **CDT** will pay to have options presented to them in a certain order. On the contrary: they will pay to be presented with *different* (and better) options.

No amount of quibbling about how to individuate options will change the fact that those who abide **CDT** will lose \$60, on average, in the sequential decisions from §2.3, while those who are always indifferent between A, B, and C given

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 problem, and that, in some good sense of 'irrelevant', L^* is an irrelevant option (you certainly shouldn't choose it), but I don't take it to be an objection to **Same Option** 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.

a choice 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, evidential decision theorists will end up predictably poorer than causalists.[26] 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 various temporal parts of these agents are like separate agents, each facing their own, separate decisions, 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.[27] (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.[28] 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 problem, 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

So **Same Option** allows causalists to reconcile **CDT** with **IIA** and **NEE** and thereby escape the cycle. However, I expect some readers to find it a bit *ad hoc*. They may say: 'yes, the quantities $\mathcal{U}_Y(X)$ do not change during the course of deliberation—but why are these quantities relevant to rational choice?' I believe that there are things to be said here. For instance, the utility of an option is just a weighted average of the the quantities $\mathcal{U}_Y(X)$, with weights given by your option probabilities, $\mathcal{U}(X) = \sum_Y \mathcal{U}_Y(X) \cdot \Pr(Y)$. So, while $\mathcal{U}(X)$ corresponds to choiceworthiness, it is determined by the quantities $\mathcal{U}_Y(X)$ and your option probabilities. So X's utility profile represents the component of utility which is invariant throughout deliberation. Nonetheless, those readers who worry about **Same Option** being *ad hoc* have my sympathy. As I mentioned in §1.3 above, I do not accept CDT in full generality. In particular, I reject it in cases with the structures of Self-Undermining Choice and Self-Reinforcing Choice. And the theory of rational choice which I endorse in those cases allows us to give a different—and, from my perspective at least, less *ad hoc*—motivation for indi-

[26] See Wells (forthcoming).

[27] 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 (msa, §5).

[28] See, *e.g.*, McClennan (1990) and Bratman (1999).

$$\begin{array}{c|cc} \mathcal{D}(\mathit{Row}\ \mathit{Col}) & K_A & K_D \\ \hline A & 100 & 0 \\ D & 0 & 100 \end{array} \qquad \begin{array}{c|cc} \Pr(\ \mathit{Row} \mid \mathit{Col}) & A & D \\ \hline K_A & 70\% & 25\% \\ K_D & 30\% & 75\% \end{array}$$

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.)

viduating options with **Same Option**.

Recall, in SELF-REINFORCING CHOICE, choosing either option would give you the good news that your choice will make things better than the alternative would—$\mathcal{U}_X(X) > \mathcal{U}_X(Y)$ and $\mathcal{U}_Y(Y) > \mathcal{U}_Y(X)$. For a concrete case like this, consider:[29]

> 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 choice, $\mathcal{U}_A(A) = 70 > 30 = \mathcal{U}_A(D)$, and $\mathcal{U}_D(D) = 75 > 25 = \mathcal{U}_D(A)$. So this case has the structure of SELF-REINFORCING CHOICE. 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 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. The difference $\mathcal{U}_X(X) - \mathcal{U}_X(Y)$ says how good the news X would give you is. And the difference $\mathcal{U}_Y(Y) - \mathcal{U}_Y(X)$ says how good the news Y would give you is. So, if the former is greater than the latter—if $[\mathcal{U}_X(X) - \mathcal{U}_X(Y)] > [\mathcal{U}_Y(Y) - \mathcal{U}_Y(X)]$—then X’s news is better than Y’s, and I say that you should

[29] 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’.)

prefer X. If $[\mathcal{U}_X(X) - \mathcal{U}_X(Y)] < [\mathcal{U}_Y(Y) - \mathcal{U}_Y(X)]$, then Y's news is better than X's, and I say that you should prefer Y. If they are equal, then X's and Y's news is equally good—in that case, I say that you should be indifferent between them. (Now, don't get it twisted: when I say that choosing an option gives you 'good news', I don't mean that it tells you that the world *as a whole* is good. I agree with orthodox CDT that that kind of news is irrelevant to rational choice. I mean instead that it tells you you that your choice is making the world better than the alternative would. And I think that this kind of news is highly relevant to rational choice.) The same theory handles cases with the structure of SELF-FRUSTRATING CHOICE, as the reader may verify for themselves.[30]

More must be said to generalize this theory to choices between more than two options. I won't go into it here, but see [author] for the details. The important point for present purposes is just that, according to this theory, if you want to check whether you have reason to prefer X to Y, you should look at the quantities $\mathcal{U}_X(X), \mathcal{U}_Y(X), \mathcal{U}_X(Y)$, and $\mathcal{U}_Y(Y)$—that is, you should look at the values which appear in X and Y's *utility profiles*. On this theory, it is not *ad hoc* to want to include these quantities in our characterization of when options are the same—for these are precisely the quantities we use to evaluate acts for choiceworthiness. So, while both orthodox CDT and my preferred causalist theory of rational choice are able to escape the cycle, it appears to me that my escape route is slightly less *ad hoc* than the orthodox causalist's.

It's also worth noting that, while my theory of rational choice, together with **Same Option**, will *always* satisfy **IIA** and **NEE**—see [author] for details—the same may not be said for orthodox CDT. Suppose that you always begin deliberation thinking that you are equally likely to select each of the available options. Then, in cases with the structure of SELF-REINFORCING CHOICE, orthodox CDT will violate both **IIA** and **NEE**, even when options are individuated with **Same Option**.

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 and D the utilities $\mathcal{U}(A) = 42.5$ 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 choice between A (going to Aleppo *via* the original road), A^* (going to Aleppo *via* the

[30] This theory is also endorsed in BARNETT (ms). BARNETT and I agree about how to choose in two-option cases, though we disagree about cases like UTILITY CYCLE. BARNETT says that your preferences in UTILITY CYCLE should by cyclic—you should prefer A to B, B to C, and C to A. I disagree; I think that you should be indifferent between A, B, and C. But this disagreement isn't relevant for present purposes. See [author] for more.

new road), and D (going to Damascus). If you again begin deliberation by distributing your option probabilities evenly, then you will assign A, A^*, and D the utilities: $\mathcal{U}(A) = \mathcal{U}(A^*) = 55$ 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 choice between A and D, CDT says that A is impermissible. But, in your choice between A, D, and A^*, it says that A is permissible. Since both A and D have the same utility profiles in both of these choices, CDT violates **IIA**, given that options are individuated with **Same Option**.

Using the same decision, we can construct a counterexample to **NEE**. Again, suppose that you always distribute your option probabilities evenly. Choosing between A, A^*, and D, CDT says that it is permissible to not choose D. But, given a choice between D and going on to choose between A and A^*, CDT says that you are required to choose D. Since A, A^*, and D have the same utility profiles in each of these choices, this violates **NEE**, given that options are individuated with **Same Option**.[31]

(Again, we could attempt to say that your different option probabilities are enough to make A and D in the second choice 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 different (unconditional) state probabilities are enough to make the options A and D different. But, again, this would have the uncomfortable consequence that the options between which you are choosing change as you make up your mind about what to do—recall the discussion in §3.2. There is also always the possibility of simply rejecting the principles **IIA** and **NEE**; though, in my view, we should want to hold on to these plausible principles if we can.)

Heterodox causalist theories like mine have been criticized for violating the independence of irrelevant alternatives.[32] It is therefore worth noting that, if my conclusions here are correct, this criticism is misplaced. The apparent violations of **IIA** 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** and **NEE** in cases like Cake in Damascus, my theory of rational choice will never violate **IIA** or **NEE**.

[31] The foregoing does not apply to Arntzenius's deliberational version of causal decision theory—that theory will always tell you to (be certain that you will) choose D no matter which alternatives are on the menu. It does, however, apply to other versions of deliberational causal decision theory—Skyrms's and Joyce's theories, for instance.

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

6 Conclusion

In summation, choices like Utility Cycle afford us three arguments against **CDT**. I've presented these arguments and offered causalists three replies. The first two objections: in Utility Cycle, **CDT** appears to violate weak versions of the *independence of irrelevant alternatives* (**IIA**) and *normal-form extensive-form equivalence* (**NEE**). In response to these objections, I've counseled causalists to individuate options in part according to their *utility profiles*. Individuating options in this way both prevents the principles from being trivialized and prevents **CDT** from violating the principles in Utility Cycle.

The final objection: in sequential decision problems, those who abide **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.

The sequential decision problem from Wells (forthcoming) also affords causalists the *tu quoque* that evidentialists themselves face the very same objection. However, Utility Cycle shows that causalists should reject Wells's *why ain'cha rich?* argument against EDT. Wells contends that, if I predictably make less money than you do in a sequential decision, when we hold fixed the state of nature, then this shows that I am choosing irrationally. Notice, however, that in the sequential decision Pay or A, causalists who pay $60 and go on to choose *B* will be certain to make $60 less than evidentialists who don't pay and go on to choose *B*—no matter which prediction was made. So endorsing a sequential *why ain'cha rich?* argument like Wells's means abandoning **CDT**.

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. [2]

—. 2014b. *Evidence, Decision and Causality*. Cambridge University Press, Cambridge, UK. [3], [18]

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], [21]

Arntzenius, Frank, John Hawthorne & Adam Elga. 2004. "Bayesianism, Infinite Decisions, and Binding." *Mind*, vol. 113 (450): 251–283. [18]

Bales, Adam. 2018. "Richness and Rationality: Causal Decision Theory and the WAR Argument." *Synthese*, vol. 195 (259–67). [11]

Barnett, David James. ms. "Graded Ratifiability." [2], [7], [20], [21]

Bassett, Robert. 2015. "A Critique of Benchmark Theory." *Synthese*, vol. 192 (1): 241–267. [2], [21]

Bratman, Michael. 1999. "Toxin, Temptation, and the Stability of Intention." In *Faces of Intention*. Cambridge University Press. [18]

Briggs, R. A. 2010. "Decision-Theoretic Paradoxes as Voting Paradoxes." *The Philosophical Review*, vol. 119 (1): 1–30. [2]

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review*, vol. 116 (1): 93–114. [2]

Egan, Andy, John Hawthorne & Brian Weatherson. 2005. "Epistemic Modals in Context." In *Contextualism in Philosophy*, G. Preyer & G. Peter, editors, 131–170. Oxford University Press, Oxford. [6]

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], [11], [19]

Hare, Caspar & Brian Hedden. 2016. "Self-Reinforcing and Self-Frustrating Decisions." *Noûs*, vol. 50 (3): 604–628. [6], [7], [19]

Harper, William. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis*, vol. 24: 25–36. [6], [13]

Hunter, Daniel & Reed Richter. 1978. "Counterfactuals and Newcomb's Paradox." *Synthese*, vol. 39 (2): 249–261. [19]

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], [11]

—. 2012. "Regret and instability in causal decision theory." *Synthese*, vol. 187 (1): 123–145. [6], [21]

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

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

—. 1981b. "'Why ain'cha rich?'." *Noûs*, vol. 15 (3): 377–380. [11]

McClennan, Edward. 1990. *Rationality and Dynamic Choice*. Cambridge University Press, Cambridge. [18]

Meacham, Christopher J.G. 2010. "Binding and Its Consequences." *Philosophical Studies*, vol. 149 (1): 49–71. [18]

Richter, Reed. 1984. "Rationality Revisited." *Australasian Journal of Philosophy*, vol. 62 (4): 392–403. [2], [6]

Sen, Amartya. 1993. "Internal Consistency of Choice." *Econometrica*, vol. 61 (3): 495–521. [13]

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. [21]

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

Spencer, Jack. msa. "CDT and the Guaranteed Principle." [2], [18]

—. msb. "Rational Monism and Rational Pluralism." [2], [7]

Spencer, Jack & Ian Wells. 2017. "Why Take Both Boxes?" *Philosophy and Phenomenological Research*. [2]

Wedgwood, Ralph. 2013. "Gandalf's solution to the Newcomb Problem." *Synthese*, vol. 190 (14): 2643–2675. [2], [7], [13], [21]

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

Wells, Ian. forthcoming. "Equal Opportunity and Newcomb's Problem." *Mind*. [5], [11], [18], [22]

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