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 behaviour and long-run poverty. These consequences appear damning, but I urge caution. This decision should lead causalists to a better understanding of what it takes for a decision between some collection of options to count as a *subdecision* of a decision between a larger collection of options. And with this better understanding of subdecisions in hand, causalists will not violate the IIA or the NEE. This decision will also teach causalists that, in sequential decisions, a rational agent may be led to make a series of choices which are causally dominated by some other sequence of choices they could have made instead. I will encourage causalists to recognise this as an intrapersonal tragedy of the commons.

1 | INTRODUCTION

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 behaviour 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, 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—at least, assuming that you are certain to retain your desires and rationality and learn nothing unexpected.[1] If, in a decision between chicken, steak,

Final Draft. Forthcoming in *Mind*. DOI: 10.1093/mind/fzab047
✉: dmitri.gallow@acu.edu.au

* 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.

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.

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 (§3.1) and NEE (§3.2). These violations lead causalists to engage in exploitable behaviour 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 (§3.3). 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. For instance, IIA says that, if it is impermissible to choose X in the decision $\mathscr{D}$, 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 say what it takes for one decision to be a subdecision of another. Given a natural answer to this question, CDT will violate IIA and NEE. But I'll suggest an alternative approach to causalists which allows them to satisfy the principles (§§4–5). I'll additionally counsel causalists to defend their exploitable behaviour as an intrapersonal tragedy of the commons: agents incapable of binding themselves to a course of action can be led to predictable 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.

2 | CAUSAL DECISION THEORY

2.1 *Desire*

I will assume that, when you face a decision,[2] you have some set of available *options* $\mathcal{O} = \{X_1, X_2, \dots, X_N\}$ between which you must choose. When making this choice, there is some set of *states of nature* $\mathcal{K} = \{K_1, K_2, \dots, K_M\}$, 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, 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

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 selection you make. To avoid confusion, I will reserve 'decision' for the situation you face, and 'choice' for the selection you make. 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.

$$
\begin{array}{c|cc}
\mathcal{D}(row \wedge col) & K_L & K_M \\
\hline
L & 100 & 0 \\
M & 110 & 10
\end{array}
\qquad
\begin{array}{c|cc}
\mathcal{P}(row|col) & L & M \\
\hline
K_L & 90\% & 10\% \\
K_M & 10\% & 90\%
\end{array}
$$

FIGURE 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.

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.

2.2 *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. If it was predicted that you would take L, then \$100 was placed in L and \$110 was placed in M. If it was predicted that you would take M, then \$0 was placed in L and \$10 was placed in M. These predictions are 90% reliable—by which I mean: given that you choose X, the objective chance of it being predicted that you choose X is 90%, independent of the course your deliberation took prior to you choosing box X. Though your choice is *correlated* with how much money is in the boxes, nothing you do now will affect the boxes' contents.

We can represent this decision with the two matrices shown in figure 1. There are two relevant states of nature. Either it was predicted that you would take L, 'K_L', or it was predicted that you would take 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 is as shown in the $\mathcal{D}$-matrix on the left of figure 1. The matrix on the right says: given that you choose L, you're 90% sure that you are in the state K_L and 10% sure that you are in the state K_M. And, given that you choose M, you're 10% sure that you are in the state K_L and 90% sure that you are in the state K_M. (I make these assumptions about your subjective probabilities because I'm taking for granted that, in NEWCOMB, you know for sure that,

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

conditional on you choosing X, the objective chance that you were predicted to choose X is 90%.)

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

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

So evidential decision theorists advise you to take L. But notice that, no matter what was predicted, taking M will get you strictly more money. In each state of nature, taking 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 dramatise this violation of deontic reflection. 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 and may offer advice about which box to choose, though he cannot say what he sees. He looks inside, and says: 'Please, choose M'. (Of course he does—that's what he'd say, no matter what he sees.) The evidential decision theorist ignores Oliver's advice, and chooses L instead. They say: '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. Nonetheless, I must take L.' At this point, the producers of the game show 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 take M. But currently, I think that's worse than L. So I think it's better for me to not know.' The producers try a different tack. They say: 'If you don't let Oliver tell you what he sees, then we'll take \$60 away from him.' The evidential decision theorist knows that, if they listen to Oliver, they'll take M. They desire taking M with a strength of \$20. On the other hand, if they don't listen, they'll take 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

6. See Arntzenius (2008)

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 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 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 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 maximising *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.

2.3 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. Because the states in $\mathcal{K}$ are causally independent of your choice, for each $K \in \mathcal{K}$, $\mathcal{D}(X \wedge K)$ gives a measure of the desirability of what X would *bring about*, were you to choose it in the state K. So the quantity

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

measures 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 bring that state about (that's what it is for K to be a state *of nature*). And according to causalists, the fact that an option makes a desired state more likely doesn't speak in its favour 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 may you evaluate the utility of X from the perspective you would occupy, were you to choose another option, Y, $\mathcal{U}_Y(X)$. (I mean: the perspective you'd occupy after learning *only* that you'd chosen Y, and before learning anything else.)

7. See Wells (2019).

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

From this perspective, your probability for each state K would be $\mathcal{P}(K \mid Y)$, so

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

would be the utility of X.

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]

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. For those arguments, I need only appeal to the following, minimal commitment of CDT, which is also endorsed by heterodox causalists like myself:[10]

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

9. For further discussion, see Gibbard and Harper (1978), Richter (1984), Weirich (1985), Harper (1986), Egan (2007), Briggs (2010), Joyce (2012), Wedgwood (2013), Ahmed (2014a), Hare and Hedden (2016), Spencer and Wells (2019), Armendt (2019) Gallow (2020, 2021), Spencer (2021a), Podgorski (forthcoming), and Williamson (2021).

10. **Minimal CDT** is accepted by Wedgwood (2013), Barnett (forthcoming), Spencer (2021b), Gallow (2020), and Podgorski (forthcoming), as well as by *deliberational* causal decision theorists like Skyrms (1990), Arntzenius (2008), and Joyce (2012, 2018).

$$
\begin{array}{c} \mathcal{D}(row \wedge col) \\ A \\ B \\ C \end{array}
\begin{array}{c} \begin{array}{ccc} K_A & K_B & K_C \end{array} \\ \begin{bmatrix} 0 & -100 & 100 \\ 100 & 0 & -100 \\ -100 & 100 & 0 \end{bmatrix} \end{array}
\qquad
\begin{array}{c} \mathcal{P}(row \mid col) \\ K_A \\ K_B \\ K_C \end{array}
\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 80\% & 10\% & 10\% \\ 10\% & 80\% & 10\% \\ 10\% & 10\% & 80\% \end{bmatrix} \end{array}
$$

FIGURE 2: Desires and Probabilities for UTILITY CYCLE

decision is neither self-undermining nor self-reinforcing.) For instance, in NEWCOMB, **Minimal CDT** tells us that M is required and L is impermissible.

3 | UTILITY CYCLE, AND THREE OBJECTIONS TO MINIMAL CDT

Consider the following decision:[11]

> UTILITY CYCLE
> Before you are three boxes, labelled 'A', 'B', and 'C'. You may take one and only one of the boxes. 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—by which I mean: conditional on you selecting X, the chance that it was predicted that you would select X is 80%, independent of the course your deliberation takes prior to your choosing box X. Nothing you do now will affect how much money is in the boxes.

Your desires and probabilities are shown in figure 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 A, B, and C, respectively. Then:

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

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

11. Similar decisions are discussed in Ahmed (2012), Hare and Hedden (2016), and Gallow (2020).

probability for C, c, is constrained to be zero, and the utilities for A and B are:

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

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 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:

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

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.

3.1 *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 *the independence of irrelevant alternatives* (**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.[12]

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, 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 on the restricted menu $\{A, B\}$, but it *is* a permissible choice on 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 $\{C, A\}$.

12. We should 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 related discussion.

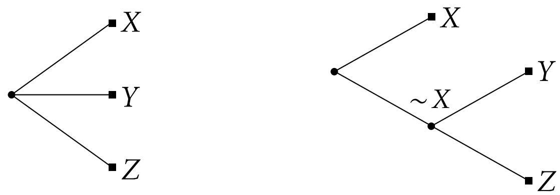


FIGURE 3: **NEE** says that, if it is permissible to not choose X in the decision between $X, Y,$ and Z on the left, then it is permissible to not choose X in the decision between X and a decision between Y and Z on the right—so long, that is, as you are certain to retain your desires and rationality, and you are certain to learn nothing unexpected.

3.2 *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 it is permissible to not choose X in a decision between $X, Y,$ and Z, then—so long as you are certain that you will remain rational, your desires will not change, and you will learn nothing unexpected—in a decision between X and a decision between Y and Z, it is permissible to not choose X. (See figure 3.)

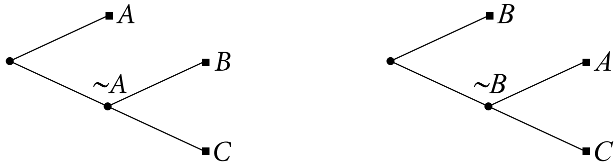
The qualification to **NEE** is important. Suppose you think that, if you were to forego X, your desires might change before you decide between Y and Z. Then, it may be rational to choose X now in order prevent your not-entirely-trustworthy future self from choosing against your current interests. Likewise,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

Consider now the following two decisions (see figure 4):

> 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.
>
> 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.

In both decisions, you are certain to learn nothing unexpected and retain your desires and rationality throughout.

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



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

FIGURE 4

are deciding 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 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 $\sim A$ is permissible in A OR ~A. **Minimal 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. **Minimal CDT**, on the other hand, says that $\sim B$ is impermissible, contradicting **NEE**. Either way, **Minimal CDT** contradicts **NEE**.

3.3 *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.) Suppose you tell your assistant to take box A. Before your assistant departs, they say: 'Are you sure? I'll give you an opportunity to change to B (but not C—I'm taking C 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 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. 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 M in NEWCOMB will predictably make less money, over the long run, than someone who takes L. The usual causalist reply is: 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 opportunities for wealth a sign of irrationality. So predictable poverty in NEWCOMB is no sign of irrationality.[13] A comparable defence 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 learn nothing unexpected and to retain your desires and rationality. For instance, consider PAY OR A:

> PAY OR A
> Money is distributed 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 $\sim$B. If you do not, then, at stage 2, you will face A OR $\sim$A. (See figure 5.)

If you know that you abide by **Minimal CDT**, you will choose A in A OR $\sim$A. So, if

13. For this response, see for instance Gibbard and Harper (1978), Lewis (1981b), Joyce (1999), Bales (2018), and Wells (2019).

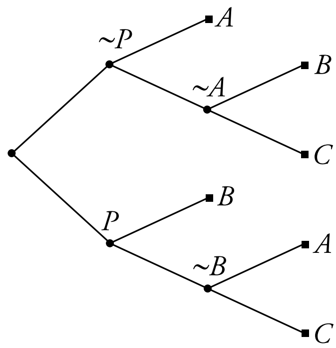


FIGURE 5: PAY OR A

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. 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, 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.

4 | SUBDECISIONS AND CYCLES

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.[14] Defenders of **Minimal CDT** could reject the principles IIA and NEE;[15] or they could insist that UTILITY CYCLE is a rational dilemma in which no option is permissible.[16] 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 under-

14. 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.

15. See Wedgwood (2013) and Barnett (forthcoming), who reject **IIA**.

16. Harper (1986) says this about similar decisions.

stood, 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.

4.1 Subdecisions

IIA says that, if you add an *irrelevant* option, 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. Some apparent counterexamples to IIA are not genuine counterexamples, because they involve new considerations which are relevant to how you should evaluate the 'original' options. 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).[17] 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 and beer is not a *subdecision* of the decision you face when presented with the options of soda, beer, and whiskey. Because it changes the way you should compare soda to beer, whiskey is not an *irrelevant* alternative.

When I argued in §3 above that **Minimal CDT** violated IIA and NEE, I was implicitly assuming that the decision between A and B was a subdecision of the decision between $A, B,$ and C. That is: I was implicitly assuming that including the option of taking box C doesn't make any difference to how you should compare A to 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. 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 *characterised by* the quadruple $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$.

Insofar as we have an arbitrary choice of how fine-grained to make the states in $\mathcal{K}$, and insofar as we have an arbitrary choice of zero and unit for the function $\mathcal{D}$, a single decision may be characterised by *multiple* such quadruples. If one decision, $\mathscr{D}$,

17. Cf. Sen (1993).

is eligible to be a *subdecision* of another, $\mathscr{D}^*$, then there must be an appropriate correspondence between a quadruple $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$ which characterises $\mathscr{D}$ and a quadruple $\langle \mathcal{O}^*, \mathcal{K}^*, \mathcal{D}^*, \mathcal{P}^* \rangle$ which characterises $\mathscr{D}^*$. A correspondence like that will tell us which options in $\mathcal{O}$ correspond to which options in $\mathcal{O}^*$, as well as which states in $\mathcal{K}$ correspond to which states in $\mathcal{K}^*$. In my discussion below, I will take for granted that we are given appropriate quadruples which characterise the decisions $\mathscr{D}$ and $\mathscr{D}^*$, along with an injection which maps each $X \in \mathcal{O}$ to a unique option in $\mathcal{O}^*$, which I will write 'X^*', and that we are given an injection which maps each $K \in \mathcal{K}$ to a unique state in $\mathcal{K}^*$, which I will write 'K^*'. I'll call those quadruples and those injections 'a mapping of $\mathscr{D}$ into $\mathscr{D}^*$'.

I was assuming in §3 that $\mathscr{D}$ is a subdecision of $\mathscr{D}^*$ iff there's a mapping of $\mathscr{D}$ into $\mathscr{D}^*$ such that (1) in $\mathscr{D}$, you desire each option in each state to the same degree that, in $\mathscr{D}^*$, you desire the 'corresponding' option in the 'corresponding' state, and (2) in $\mathscr{D}$, your probability for each state, given each option, is the same as the probability you give to each 'corresponding' state, given each 'corresponding' option in $\mathscr{D}^*$. Call this 'the simple view' of when one decision is a subdecision of another.

The Simple View $\mathscr{D}$ is a subdecision of $\mathscr{D}^*$ iff there's a mapping from $\mathscr{D}$ into $\mathscr{D}^*$ such that, for each option X and each state K in $\mathscr{D}$,

(1) $\mathcal{D}(X \wedge K) = \mathcal{D}^*(X^* \wedge K^*)$; and

(2) $\mathcal{P}(K \mid X) = \mathcal{P}^*(K^* \mid X^*)$

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 Simple 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 (1). And, according to **The Simple View**, the decision you face between 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 Simple 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 Simple 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:

(3) $\mathcal{P}(X) = \mathcal{P}^*(X^*)$.

to **The Simple View**. This would reconcile **Minimal CDT** with IIA, but at the price of trivialising IIA. 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 Simple View** the requirement that each state $K \in \mathcal{K}$ has the same unconditional probability in the decision $\mathscr{D}$ as its corresponding state K^* has in the decision $\mathscr{D}^*$. That is, we may wish to add:

(4) $\mathcal{P}(K) = \mathcal{P}^*(K^*)$

This suggestion does not trivialise IIA, though it has other undesirable consequences. To appreciate these consequences, first note 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 some state K and every pair of distinct options X and Y. In decisions like this, changes in your option probabilities automatically lead to changes in your (unconditional) state probabilities.

For this reason, if we were to add (4) to **The Simple View**, it could 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 (4) to **The Simple View**, then the decision you would face *after* deliberation (once 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. 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 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. Then, when you begin deliberation, you know that, by the time your deliberation is finished, you will no longer face your current decision. Instead, because your option 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? 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. So I'm inclined to resist adding (4) to **The Simple View**. (All of these considerations apply with equal force to the proposal to add condition (3).)

In sum: I wish to respond to **Minimal CDT**'s apparent violation of IIA in Utility Cycle by insisting that the decision between A and B is not a subdecision of the decision between $A, B,$ and C. However, I don't think that we should do so by appealing to to either your option probabilities or your state probabilities. In §4.2, I'll offer the causalist a different account of when one decision is a subdecision of another. I will then explain, in §5, how this account allows causalists to dispute the charge that they violate IIA and NEE in Utility Cycle.

4.2 *No New Cycles*

Let me begin with the notion of a *pairwise rational preference* between two options. Intuitively, a pairwise rational preference between X and Y is the preference it would be rational for you to have, were you forced to choose between just X and Y. More carefully, if we begin with a decision, $\mathscr{D}$, characterised by the quadruple $\langle \mathcal{O}, \mathcal{K}, \mathcal{D}, \mathcal{P} \rangle$, with X and Y included in $\mathcal{O}$, then we may construct a new decision $\mathscr{D} \downharpoonleft \{X, Y\}$ (the decision $\mathscr{D}$, restricted to the options X and Y) as follows: the new set of options is just $\{X, Y\}$, the states and desires remain unchanged, and the new probability function is $\mathcal{P}$ conditioned on the proposition that you will choose either X or Y. If it is rational for you to prefer X to Y in the decision $\mathscr{D} \downharpoonleft \{X, Y\}$, then I will say that, in the decision $\mathscr{D}$, you have a pairwise rational preference for X over Y. If it is rational for you to be indifferent between X and Y in the decision $\mathscr{D} \downharpoonleft \{X, Y\}$, then I will say that, in $\mathscr{D}$, you are pairwise rationally indifferent between X and Y.

If we accept **Minimal CDT**, then we will think that your pairwise rational preferences can form cycles. For, in Utility Cycle, **Minimal CDT** tells us that you have a pairwise rational preference for B over A, for A over C, and for C over B. If we use a directed arrow $Y \to X$ to indicate that you have a pairwise rational preference for X over Y, then we may diagram your pairwise rational preferences in Utility Cycle as follows:

$$\begin{tikzcd}
B \arrow[dd, bend right] & \\
 & A \arrow[ul, bend right] \\
C \arrow[ur, bend right] &
\end{tikzcd}$$

In my view, adding C to the menu of options $\{A, B\}$ changes the way that you should evaluate the options A and B precisely because adding C introduces a new chain of pairwise rational preference leading from B to A: $B \to C \to A$. For this reason, adding C to the menu of options introduces a new reason to prefer A to B: namely, A is preferable to C, and C is preferable to B (in a pairwise decision between the two). If C is not an option, then you only have reason to prefer B to A. But once C is included as an option, you have additional *pro tanto* reason to prefer A to B. For this reason, I believe that we should say that C is not an *irrelevant* alternative.

In general, I believe we should say that some new option is not an *irrelevant* alternative if it introduces a cycle in your pairwise rational preferences. That is, I think we should emend **The Simple View** by adding the condition that the new options in the 'larger' decision $\mathscr{D}^*$ do not introduce any cycles in your pairwise rational preferences between the 'original' options from the 'smaller' decision $\mathscr{D}$.[18] Call the resulting view **No New Cycles**.

No New Cycles $\mathscr{D}$ is a *subdecision* of $\mathscr{D}^*$ iff there's a mapping from $\mathscr{D}$ into $\mathscr{D}^*$ such that, for each option X and each state K in $\mathscr{D}$,

(1) $\mathcal{D}(X \wedge K) = \mathcal{D}^*(X^* \wedge K^*)$;

(2) $\mathcal{P}(K \mid X) = \mathcal{P}^*(K^* \mid X^*)$; and

(5) X^* does not enter into any cycle of pairwise rational preference with any of the options from $\mathscr{D}^*$ which do not 'correspond' to any option from $\mathscr{D}$.

My advice to causalists is this: reject **The Simple View** of when one decision is a subdecision of another, and accept in its place **No New Cycles**.

5 | ESCAPING THE CYCLE

If causalists accept my advice, they will be able to accept the plausible principles IIA and NEE.

By the way, when I stated the principle **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 about this requirement, we should explicitly include it in our statement of the principle. I will also generalise the principle to cover decisions between more than three options.

IIA′ Suppose that X is an option in the decision $\mathscr{D}$, $\mathscr{D}$ is a subdecision of $\mathscr{D}^*$, and 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 $\mathscr{D}^*$.

18. Terminology: a *cycle* in your pairwise rational preferences is any sequence of options $(X_1, X_2, \ldots, X_N)$ such that (a) for each $i \in \{1, 2, \ldots, N\}$, X_i is either pairwise rationally preferred to $X_{i+1 \bmod N}$ or you are pairwise rationally indifferent between X_i and $X_{i+1 \bmod N}$; and (b) for some $i \in \{1, 2, \ldots, N\}$, X_i is pairwise rationally preferred to $X_{i+1 \bmod N}$.

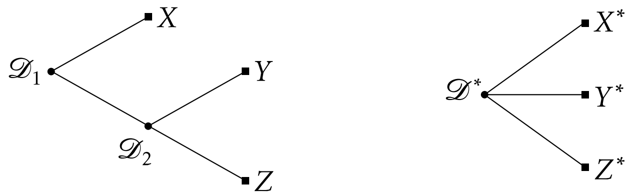


FIGURE 6: **NEE′** says that, if both $\mathscr{D}_1$ and $\mathscr{D}_2$ are subdecisions of $\mathscr{D}^*$, and it is permissible to not choose X^* in $\mathscr{D}^*$, then it is permissible to choose to leave behind the option corresponding to X^*, X, in $\mathscr{D}_1$—so long, that is, as you are certain to retain your desires and rationality and learn nothing unexpected before choosing in $\mathscr{D}_2$.

Similarly, we should more carefully state the principle **NEE** (see figure 6).

NEE′ Suppose that 1) $\mathscr{D}_1$ is a decision between X and $\mathscr{D}_2$, which is a decision between Y and Z; 2) in $\mathscr{D}_1$, you are sure that, if you forego X, you will learn nothing unexpected nor lose your desires or rationality before choosing in $\mathscr{D}_2$; 3) $\mathscr{D}^*$ is a decision between an option corresponding to X, X^*, and options corresponding to Y and Z, Y^* and Z^*; and 4) $\mathscr{D}_1$ and $\mathscr{D}_2$ are subdecisions of $\mathscr{D}^*$. Then, if it is permissible to not choose X^* in $\mathscr{D}^*$, it is also permissible to not choose X in $\mathscr{D}_1$.

According to **No New Cycles**, a decision between boxes A and B, $\{A, B\}$, will not count as a subdecision of a decision between boxes $A, B,$ and C, $\{A, B, C\}$. For the 'new' option C enters into a cycle of pairwise rational preference with the 'old' options A and B. For the same reason, neither $\{B, C\}$ nor $\{A, C\}$ will count as subdecisions of $\{A, B, C\}$. So **IIA′** will not tell us anything about UTILITY CYCLE, if we accept **No New Cycles**.

For similar reasons, if we accept **No New Cycles**, **NEE′** will not tell us anything about either A OR ∼ A or B OR ∼ B. Consider A OR ∼ A. At stage 1, you face a decision between taking box A and going on to choose between boxes B and C, $\{A, \{B, C\}\}$. If you know that your future self will neither change their desires nor learn anything unexpected, 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 A, $\{B, C\}$, corresponds to the option of taking box C in a decision between $A, B,$ and C, $\{A, B, C\}$. But $\{A, \{B, C\}\}$ will not count as a subdecision of $\{A, B, C\}$. For, in $\{A, B, C\}$, the 'new' option B will introduce a cycle in your pairwise rational preferences which involves the 'old' options A and 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 §3.3, 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.[19] More convincingly, they can object to using outcomes in *sequential* decisions

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

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.[20]

(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.[21] 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.)

6 | FURTHER DISCUSSION

No New Cycles allows us to avoid a conflict between **Minimal CDT** and the plausible principles **IIA′** and **NEE′** in the decision UTILITY CYCLE. But this does not on its own guarantee that **No New Cycles** will allow **Minimal CDT** to *always* satisfy the principles **IIA′** and **NEE′**.

In this section, I'll show that, if we accept **No New Cycles**, then there will *never* be a conflict between **Minimal CDT** and either **IIA′** or **NEE′**. To establish this, it is enough to show that there is one theory of rational choice which entails **Minimal CDT** and which does not violate **IIA′** or **NEE′**, so long as we understand subdecisions in the way proposed by **No New Cycles**. The theory I propose and defend in Gallow (2020) is one such theory—though there are others.[22] According to this theory, pairwise rational preferences are determined in the following way: you consider how much more you would expect X to improve the world than Y, were you to learn that you had chosen X: $\mathcal{U}_X(X) - \mathcal{U}_X(Y)$. This difference measures the news that X carries about how much more it is doing to make the world better than Y would—call that X's *improvement news* (vis-a-vis Y). Likewise, $\mathcal{U}_Y(Y) - \mathcal{U}_Y(X)$ measures the improvement news that Y carries (vis-a-vis X). If X's improvement news is better than Y's, then I say that you have a pairwise rational preference for X over Y. If X's improvement news is just as good as Y's, then I say that you are pairwise rationally indifferent between X and Y. Notice that this theory entails **Minimal CDT**. For, in a decision between X and Y, if $\mathcal{U}_X(X) > \mathcal{U}_X(Y)$, then X's improvement news is positive; and if $\mathcal{U}_Y(X) > \mathcal{U}_Y(Y)$,

20. See Arntzenius et al. (2004) for further defence of this view, and see Meacham (2010) for a reply. See also Ahmed (2014b, §7.4.3) and Spencer (2021a, §5).

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

22. For instance, the theories discussed in Podgorski (forthcoming, §6) will also satisfy **IIA′** and **NEE′** if we assume **No New Cycles**.

then Y's improvement news is negative. So I will say that you have a pairwise rational preference for X over Y, and that (therefore) X is required and Y is impermissible.

More must be said to generalise this theory to decisions between more than two options. I won't go deep into the details here, but the basic idea is this: you use pairwise rational preferences to construct a total preference ordering. This can lead to cyclic preferences in some cases—as in Utility Cycle. In those cases, some of the pairwise preferences are ignored and not incorporated into your total preference ordering. Importantly, a pairwise preference is only ever ignored if it leads to a cycle. So, if an additional option is *irrelevant*, in the sense of **No New Cycles**—that is to say: if the additional option does not create any new cycles with the 'old' options—then it will not affect which pairwise preferences between the 'old' options end up being incorporated into your total preference ordering, and which end up being ignored. It is for this reason that the theory will satisfy **IIA′**, given **No New Cycles**. Suppose that X was *impermissible* in the 'old' decision $\mathscr{D}$. Then, there must have been some 'old' option, Y, which you pairwise rationally preferred to X, and this pairwise rational preference must have been incorporated into your total preference ordering in $\mathscr{D}$. So long as the 'new' options in $\mathscr{D}^*$ don't introduce any new cycles involving X^* and Y^*, the 'corresponding' pairwise rational preference will be incorporated into your total preference ordering in $\mathscr{D}^*$, and so X^* will be dispreferred to Y^* in $\mathscr{D}^*$. So X^* will be impermissible in $\mathscr{D}^*$. (For similar reasons, the theory will not violate **NEE′**, so long as we accept **No New Cycles**.)

For present purposes it doesn't matter whether the theory is correct or even plausible. Whatever we think about the theory, it teaches us that there will never be a conflict between **Minimal CDT** and the principles **IIA′** and **NEE′**, so long as we accept **No New Cycles**.

As an aside, heterodox causalist theories like mine have been criticised for violating the independence of irrelevant alternatives.[23] It is therefore worth noting that apparent violations of the IIA are not unique to the heterodox. Orthodox CDT also appears to violate the principle, and in similar ways. Moreover, while the heterodox theory of rational choice I favour will always satisfy **IIA′** and **NEE′** (given **No New Cycles**), the same cannot be said for orthodox CDT. To appreciate this, consider the following decision:[24]

> Fortunate Prediction
> You must decide between three boxes, labelled 'F', 'F^*', and 'G'. If it was predicted that you'd take either F or F^*, then there is \$50 in both F and F^* and nothing in G. If it was predicted that you'd take G, then there is nothing in either F or F^* and \$60 in G.

Suppose that your desires and probabilities are as shown in figure 7. And suppose that

23. See the discussion in Wedgwood (2013), Bassett (2015), and Barnett (forthcoming).

24. Below, I focus on orthodox CDT's violation of **IIA′**. To appreciate its violation of **NEE′**, see Joyce (2018)'s discussion of Ahmed (2014a).

$$\begin{array}{cc} \mathcal{D}(row \wedge col) & \begin{array}{cc} K_F & K_G \end{array} \\ \begin{array}{c} F \\ F^* \\ G \end{array} & \begin{bmatrix} 50 & 0 \\ 50 & 0 \\ 0 & 60 \end{bmatrix} \end{array} \qquad \begin{array}{cc} \mathcal{P}(row \mid col) & \begin{array}{ccc} F & F^* & G \end{array} \\ \begin{array}{c} K_F \\ K_G \end{array} & \begin{bmatrix} 80\% & 80\% & 20\% \\ 20\% & 20\% & 80\% \end{bmatrix} \end{array}$$

FIGURE 7: Desires and probabilities for FORTUNATE PREDICTION. K_F is the state in which it was predicted that you'd take F or F^*. K_G is the state in which it was predicted that you'd take G.

your option probabilities are distributed evenly: $\mathcal{P}(F) = \mathcal{P}(F^*) = \mathcal{P}(G) = 1/3$. Then, the utility of F and F^* will both be 30, whereas the utility of G will be 24. So orthodox CDT will say that F is permissible. (And it will continue to say this as your option probability for F rises to 100%.) Abbreviate this decision with '$\mathscr{FP}$'. And consider the decision $\mathscr{FP} \downharpoonright \{F, G\}$ (the decision FORTUNATE PREDICTION, restricted to the options F and G). In this decision, your option probabilities for F and G will be even: $\mathcal{P}(F) = \mathcal{P}(G) = 1/2$. Then, the utility of F will be 25, whereas the utility of G will be 30. So orthodox CDT will say that F is impermissible in $\mathscr{FP} \downharpoonright \{F, G\}$. (And it will continue to say this as your option probability for G rises to 100%). Moreover, according to **No New Cycles**, $\mathscr{FP} \downharpoonright \{F, G\}$ is a subdecision of $\mathscr{FP}$. In that case, **IIA′** says that, if F is an impermissible choice in $\mathscr{FP} \downharpoonright \{F, G\}$, then F should be an impermissible choice in $\mathscr{FP}$. Orthodox CDT disagrees, so orthodox CDT violates **IIA′** (given **No New Cycles**).

Again, the orthodox causalist could suggest that your different option probabilities are enough to make F and G in the decision $\mathscr{FP} \downharpoonright \{F, G\}$ importantly different from the 'corresponding' options in $\mathscr{FP}$. But, again, this trivialises **IIA′**. If you always begin deliberation by giving a positive probability to each option, then **IIA′** would never apply, given this criterion for one decision being a subdecision of another. Alternatively, they could suggest that your different (unconditional) state probabilities are enough for $\mathscr{FP} \downharpoonright \{F, G\}$ to not count as a subdecision of $\mathscr{FP}$. But, again, this would have the uncomfortable consequence that deliberating about which option to choose can change the decision you face—recall the discussion from §4.1.

Of course, I haven't ruled out that there is another criterion for subdecisions which will tell us that $\mathscr{FP} \downharpoonright \{F, G\}$ is not a subdecision of $\mathscr{FP}$. I don't mean to suggest that the orthodox causalist is *forced* to reject **IIA′**. Instead, I simply want to register that, while many heterodox causalist theories, including the one I favour, will always abide by **IIA′**, given **No New Cycles**, orthodox CDT has additional difficulty complying with **IIA′** in decisions like FORTUNATE PREDICTION.

7 | 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 equiva-*

lence. In response to these objections, I've counselled causalists to say that, if a 'new' option introduces a cycle in your pairwise rational preferences between 'old' options, then the new option is not an *irrelevant* alternative, and the 'old' decision is not a subdecision of the 'new' one. This prevents the carefully-formulated principles **IIA′** and **NEE′** from being trivialised 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 counselled 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 parallel 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 be certain to make \$60 less than an evidentialist 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**.[25]

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

REFERENCES

Ahmed, Arif 2012, 'Push the Button,' *Philosophy of Science* 79, pp. 386–395 [7]

—— 2014a, 'Dicing with Death,' *Analysis* 74, pp. 587–592 [6], [20]

—— 2014b, *Evidence, Decision and Causality* (Cambridge: Cambridge University Press) [3], [19]

—— 2020, 'Equal Opportunity in Newcomb's Problem and Elsewhere,' *Mind* 129, pp. 867–886 [18], [22]

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

Arntzenius, Frank 2008, 'No regrets, or: Edith Piaf revamps decision theory,' *Erkenntnis* 68, pp. 277–297 [4], [6]

Arntzenius, Frank, Hawthorne, John, and Elga, Adam 2004, 'Bayesianism, Infinite Decisions, and Binding,' *Mind* 113, pp. 251–283 [19]

Bales, Adam 2018, 'Richness and Rationality: Causal Decision Theory and the WAR Argument,' *Synthese* 195, pp. 259–267 [11]

Barnett, David James forthcoming, 'Graded Ratifiability,' *Journal of Philosophy* [6], [12], [20]

Bassett, Robert 2015, 'A Critique of Benchmark Theory,' *Synthese* 192, pp. 241–267 [20]

Bratman, Michael 1999, 'Toxin, Temptation, and the Stability of Intention,' in 'Faces of Intention,' (Cambridge: Cambridge University Press) [19]

Briggs, R. A. 2010, 'Decision-Theoretic Paradoxes as Voting Paradoxes,' *The Philosophical Review* 119, pp. 1–30 [6]

Egan, Andy 2007, 'Some Counterexamples to Causal Decision Theory,' *Philosophical Review* 116, pp. 93–114 [6]

Gallow, J. Dmitri 2020, 'The Causal Decision Theorist's Guide to Managing the News,' *The Journal of Philosophy* 117, pp. 117–149 [6], [7], [12], [19]

—— 2021, 'Riches and Rationality,' *Australasian Journal of Philosophy* 99, pp. 114–129 [6]

Gibbard, Allan and Harper, William L. 1978, 'Counterfactuals and Two Kinds of Expected Utility,' in A. Hooker, J.J. Leach, and E.F. McClennan (eds.), 'Foundations and Applications of Decision Theory,' pp. 125–162 (Dordrecht: D. Reidel) [6], [11]

Hare, Caspar and Hedden, Brian 2016, 'Self-Reinforcing and Self-Frustrating Decisions,' *Noûs* 50, pp. 604–628 [6], [7] Harper, William 1986, 'Mixed Strategies and Ratifiability in Causal Decision Theory,' *Erkenntnis* 24, pp. 25–36 [6], [12]

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

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

Joyce, James M. 1999, *The Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press) [5], [11]

—— 2012, 'Regret and instability in causal decision theory,' *Synthese* 187, pp. 123–145 [6]

—— 2018, 'Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems,' in Arif Ahmed (ed.), 'Newcomb's Problem,' (Oxford: Oxford University Press) [6], [20]

Lewis, David K. 1981a, 'Causal Decision Theory,' *Australasian Journal of Philosophy* 59, pp. 5–30 [2], [5]

—— 1981b, "Why ain'cha rich?," *Noûs* 15, pp. 377–380 [11]

McClennan, Edward 1990, *Rationality and Dynamic Choice* (Cambridge: Cambridge University Press) [19]

Meacham, Christopher J. G. 2010, 'Binding and Its Consequences,' *Philosophical Studies* 149, pp. 49–71 [19]

Podgorski, Abelard forthcoming, 'Tournament Decision Theory,' *Noûs* [6], [8], [19]

Rabinowicz, Wlodek 2009, 'Letters from Long Ago: On Causal Decision Theory and Centered Chances,' in Lars-Göran Johansson (ed.), 'Logic, Ethics, and All That Jazz—Essays in Honour of Jordan Howard Sobel,' volume 56, pp. 247–273 (Uppsala: Uppsala Philosophical Studies) [5]

Rabinowicz, Włodzimierz 1982, 'Two Causation Decision Theories: Lewis vs Sobel,' in Tom Pauli (ed.), 'Philosophical Essays Dedicated to Lennart Åqvist on His Fiftieth Birthday,' volume 34, pp. 299–321 (Uppsala: Uppsala Philosophical Studies) [5]

Ray, Paramesh 1973, 'Independence of Irrelevant Alternatives,' *Econometrica* 41, pp. 987–991 [8]

Richter, Reed 1984, 'Rationality Revisited,' *Australasian Journal of Philosophy* 62, pp. 392–403 [6]

Sen, Amartya 1993, 'Internal Consistency of Choice,' *Econometrica* 61, pp. 495–521 [13]

Skyrms, Brian 1982, 'Causal Decision Theory,' *Journal of Philosophy* 79, pp. 695–711 [5]

—— 1990, *The Dynamics of Rational Deliberation* (Cambridge, MA: Harvard University Press) [6]

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

Spencer, Jack 2021a, 'An Argument Against Causal Decision Theory,' *Analysis* 81, pp. 52–61 [6], [19]

—— 2021b, 'Rational Monism and Rational Pluralism,' *Philosophical Studies* 178, pp. 1769–1800 [6]

Spencer, Jack and Wells, Ian 2019, 'Why Take Both Boxes?' *Philosophy and Phenomenological Research* 99, pp. 27–48 [6]

Wedgwood, Ralph 2013, 'Gandalf's solution to the Newcomb Problem,' *Synthese* 190, pp. 2643–2675 [6], [12], [20]

Weirich, Paul 1985, 'Decision Instability,' *Australasian Journal of Philosophy* 63, pp. 465–478 [6]

Wells, Ian 2019, 'Equal Opportunity and Newcomb's Problem,' *Mind* 128, pp. 429–457 [5], [11], [18], [22]

Williamson, Timothy Luke 2021, 'Causal Decision Theory is Safe From Psychopaths,' *Erkenntnis* 86, pp. 665–685 [6]