Escaping the Cycle

J. Dmitri Gallow†

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

The principle I’ll call normal-form extensive-form equivalence (NEE) says that, so long as you’re certain to not change your beliefs or desires, and you’re certain to remain rational, if it’s permissible to choose an option other than X, then, if you’re given the choice to either have X or go on to choose amongst the other options, it is permissible to choose to leave X behind.† 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,

Draft of August 20, 2020. Word Count: 8,928
dmitri.gallow@gmail.com

† Thanks to Kevin Dorst, James Shaw, Daniel Drucker, 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.
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 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 options in different decision problems to count as the same. Given a natural assumption about what makes 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). I’ll additionally counsel causalists to defend their exploitable behavior as an intrapersonal tragedy of the commons: agents incapable of binding themselves to a course of action can be led to predictable financial ruin through a series of rational actions, in just the same way that society may be predictably led to collective tragedy through a series of individually rational actions (§4).

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.\(^2\) 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 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 \( X \in \mathcal{O} \),

\[
\mathcal{D}(X) = \sum_K \Pr(K | X) \cdot \mathcal{D}(XK)
\]

\( \mathcal{D}(X) \) 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 \( X \).

\(^2\) 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.
§1. Causal Decision Theory

<table>
<thead>
<tr>
<th>$D(\text{Row Col})$</th>
<th>$K_L$</th>
<th>$K_M$</th>
<th>$\text{Pr(Row Col)}$</th>
<th>$L$</th>
<th>$M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>100</td>
<td>0</td>
<td>$K_L$</td>
<td>90%</td>
<td>10%</td>
</tr>
<tr>
<td>$M$</td>
<td>110</td>
<td>10</td>
<td>$K_M$</td>
<td>10%</td>
<td>90%</td>
</tr>
</tbody>
</table>

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.

1.2 Newcomb. Some—known as evidential decision theorists—think that $D(X)$ provides a measure of the choiceworthiness of an option $X$. 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 $D$-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 $K_M$.

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

$$D(L) = \Pr(K_L \mid L) \cdot D(LK_L) + \Pr(K_M \mid L) \cdot D(LK_M)$$

$$= 90\% \cdot 100 + 10\% \cdot 0$$

$$= 90$$

while

$$D(M) = \Pr(K_L \mid M) \cdot D(MK_L) + \Pr(K_M \mid M) \cdot D(MK_M)$$

$$= 10\% \cdot 110 + 90\% \cdot 10$$

3. For defenses of evidential decision theory, see Jeffrey (1965, 2004) and Ahmed (2014b).
Escaping the Cycle

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.\(^4\)

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 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.\(^5\)

Imagine yourself as Oliver, pleading with the evidential decision theorist

\(^4\) See Arntzenius (2008)

\(^5\) See Wells (2019).
1. Causal Decision Theory

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

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

$$U(X) \equiv \sum_K \Pr(K) \cdot \mathcal{D}(XK)$$

tells us how desirable you expect choosing $X$ to make the world.6

The difference between $\mathcal{D}$ and $U$ is that, in $\mathcal{D}$, we conditioned the probability function $\Pr$ on the proposition that you choose $X$. In $U$, 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 currently occupy, so too may you evaluate the utility of $X$ from the perspective you would occupy, were you to choose another option, $Y$. (I mean: the perspective you would occupy after learning only that you’d chosen $Y$, and before learning anything else.) From this perspective, your probability for each

---

6. This is Skyrms’s formulation of causal decision theory. There are alternatives—see, e.g., Sobel (1994), Lewis (1981a), Joyce (1999), and especially Rabinowicz (1982) and Rabinowicz (2009). The differences between these version of CDT won’t make a difference to anything I have to say here.
state $K$ would be $\Pr(K \mid Y)$, and

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

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

$$U(X) = \sum_Y U_Y(X) \cdot \Pr(Y)$$

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

$$U_X(Y) > U_X(X) \quad \text{and} \quad U_Y(X) > U_Y(Y)$$

**Self-Reinforcing Choice**

Once chosen, each option would have a higher utility than the alternative

$$U_X(X) > U_X(Y) \quad \text{and} \quad U_Y(Y) > U_Y(X)$$

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.\(^7\)

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 against CDT which I’ll introduce below.\(^8\) For those arguments, I need only appeal to the following, minimal commitment of CDT, which is also endorsed by heterodox causalists like myself.\(^9\)

---


\(^8\) Though I’ll return to these kinds of choices in §5.

\(^9\) Minimal CDT is accepted by Wedgwood (2013), Barnett (ms), Spencer (forthcomingb), Gallow (2020), and Podgorski (forthcoming), as well as by deliberational causal decision theory.\(^6\)
§2. Utility Cycle, and Three Objections to CDT

Consider the following decision problem:¹⁰

**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

<table>
<thead>
<tr>
<th></th>
<th>Row</th>
<th>Col</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>−100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>100</td>
<td>0</td>
<td>−100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>−100</td>
<td>100</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr( Row</td>
<td>Col)</td>
<td>80%</td>
<td>10%</td>
</tr>
<tr>
<td>K_A</td>
<td>K_B</td>
<td>K_C</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>10%</td>
<td>80%</td>
<td>10%</td>
</tr>
<tr>
<td>10%</td>
<td>10%</td>
<td>80%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Desires and Probabilities for Utility Cycle

**Minimal CDT** In a choice between two options, X and Y, if X’s utility would exceed Y’s, whichever you chose,

\[ U_X(X) > U_X(Y) \quad \text{and} \quad U_Y(X) > U_Y(Y) \]

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 choices between two options, where the choice is neither self-undermining nor self-reinforcing.) In **NEWCOMB**, **Minimal CDT** tells us that M is required and L is impermissible. You know that, no matter what was predicted, M will get you $10 more than L does. So, if you choose L, then the utility of M will exceed the utility of L by 10 (\( U_L(M) = 100 \) and \( U_L(L) = 90 \)). And, if you choose M, then the utility of M will exceed the utility of L by 10 (\( U_M(M) = 20 \) and \( U_M(L) = 10 \)). 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

**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

orists like Skyrms (1990), Arntzenius (2008), and Joyce (2012, 2018).

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:

\[ U(A) = 70(c - b) \]
\[ U(B) = 70(a - c) \]
\[ U(C) = 70(b - a) \]

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 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:

\[ U(A) = 70a - 70 \]
\[ U(B) = 70a \]

No matter the value of \(a\), B will have a higher utility than A. So Minimal 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:

\[ U(B) = 70b - 70 \]
\[ U(C) = 70b \]

Again, no matter the value of \(b\), the utility of C will exceed the utility of B. So Minimal CDT says that, in a 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:

\[ U(C) = 70c - 70 \]
\[ U(A) = 70c \]

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

---

11. 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 \(D(\text{Row Col})\) by the matrix \(Pr(\text{Row Col})\). This gives the matrix \(U_{\text{Col}}(\text{Row})\), of the utility of the row option, from the perspective you'd occupy immediately after choosing the column option. The identity \(U(\text{Row}) = \sum_{\text{Col}}U_{\text{Col}}(\text{Row}) \cdot Pr(\text{Col})\) can then be used to easily calculate the unconditional utilities, \(U(\text{Row})\).
2. Utility Cycle, and Three Objections to CDT

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

IIA: If, given a choice between $X$ and $Y$, $Y$ is not permissible, then, given a choice between $X, Y,$ and $Z$, $Y$ is not permissible.12

According to Minimal CDT, every option in Utility Cycle is impermissible in a one-on-one choice with some alternative. So, if some option is permissible,13 we will have a violation of IIA. For illustration: suppose that $A$ is a permissible choice in Utility Cycle. By Minimal 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 Minimal CDT says that $B$ is impermissible on the restricted menu $\{B, C\}$, and $C$ is impermissible on the restricted menu $\{C, A\}$.

2.2 Normal-Form Extensive-Form Equivalence. Utility Cycle also shows that Minimal CDT violates a weak principle of normal-form extensive-form equivalence (or just ‘NEE’).

NEE: If you are certain to remain rational and your beliefs and desires are certain to not change, then, if it is permissible to not choose $X$ 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:

12. We should sharply distinguish IIA from other principles that go by that name. For instance, Podgorski (forthcoming) calls the following principle ‘the independence of irrelevant alternatives’: Your preference between $X$ and $Y$ is a function of $U_X(X), U_X(Y), U_Y(X),$ and $U_Y(Y)$ alone. This principle is logically independent from the one I’m calling ‘IIA’. See Ray (1973) for more on conflicting uses of ‘independence of irrelevant alternatives’ in social choice theory.

13. 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.
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.

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 Minimal CDT, and that you will continue to do so throughout any sequential decisions. Then, in $A \lor \neg A$, if you choose $\neg A$ 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 $B \lor \neg B$, if you choose $\neg B$ 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, Minimal 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 $\neg A$ is permissible in $A \lor \neg A$. Minimal CDT on the other hand, says that $\neg A$ 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 $\neg B$ is permissible in $B \lor \neg B$. Minimal CDT, on the other hand, says that $\neg B$ is impermissible, contradicting NEE. Either way, Minimal CDT contradicts NEE.
2.3 Predictable Long-run Poverty. Minimal CDT’s advice in Utility Cycle may be exploited to lose you money in the long run. Suppose that, instead of taking a box yourself, you select a box with the aid of an assistant. You tell the assistant which box to take, but it is the assistant who makes the final 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 $60.’ (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 sticking with A and switching to B and losing $60. If \( a \) is your probability for taking A, then the utilities of the available options are:

\[
\text{U}(A) = 70a - 70 \quad \text{U}(B) = 70a - 60
\]

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 offer you a choice of whether to stick with A or pay $60 to switch to B. If your initial selection at stage 1 is box B, then at stage 2, your assistant will offer you a choice of whether to stick with B or pay $60 to switch to C. If your initial selection at stage 1 is box C, then at stage 2, they’ll offer you a choice of whether to stick with C or pay $60 to switch to A. In any 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 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 decision over and over again, with your assistant offering the trade each time, you will lose $60 on average. In contrast, someone who refuses the assistant’s offer
to switch will break even, on average. Note also that every series of choices permitted by Minimal CDT is causally dominated by another series of choices. Whatever box you end up with after paying $60 to switch, you could have had that box’s contents for free by simply making it your initial selection and refusing to switch.

Causalists are used to making less money in certain 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 opportunities for wealth a sign of irrationality. So predictable poverty in Newcomb is no sign of irrationality. 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.

Minimal 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 throughout. 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, ~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 Minimal CDT, you will choose A in A or ~A. So,

---

if you don’t pay, you will end up choosing $A$. If you abide **Minimal CDT**, you will choose $B$ in $B$ or $\sim B$. 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, **Minimal CDT** tells you to pay the $60$.

Again, paying likely leads to you losing money overall. Whether you play $A$ or $\sim A$ or $B$ or $\sim B$, the expected return is $0$. So in the long run in which you decide to pay in $\text{PAY or A}$ 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 **Minimal CDT**—pay, then take $B$—is causally dominated. No matter what was predicted, another series of choices—don’t pay, then refuse $A$, then take $B$—makes $60$ more. So this predictable poverty does not appear to be a consequence of poor opportunities.

3 Options

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**.15 Defenders of **Minimal CDT** could reject the principles **IIA** and **NEE**;16 or they could insist that **Utility Cycle** is a rational dilemma in which no option is permissible.17 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 $\{A, B, C\}$.

3.1 Individuating Options and Irrelevant Alternatives. **IIA** says that, if you add an *irrelevant* option, $Z$, to the menu $\{X, Y\}$, this shouldn’t transform an impermissible option into a permissible one. However, not every additional option is truly *irrelevant* to your choice between the original options $X$ and $Y$. Some apparent counterexamples to **IIA** are not genuine counterexamples, because they involve new options which are relevant to how you should evaluate

15. 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.


your choice between $X$ and $Y$. For instance, consider the following putative counterexample to IIAS: 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 in-temperate). Do these choice dispositions violate IIAS? 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 them relevantly different options.

When I presented IIAS above, I simply took it for granted that the options $X$ and $Y$ were the same whether they appeared on the menu $\{X, Y\}$ or on the menu $\{X, Y, Z\}$. However, the sense in which $X$ and $Y$ could be the same is a subtle one. By Leibniz’s Law, options appearing in different decision problems are distinct. So, if $X$ and $Y$ appear on a menu by themselves, then options identical to $X$ and $Y$ could not appear on a larger menu with $Z$. Nonetheless, options suitably similar to $X$ and $Y$—call them ‘$X^*$’ and ‘$Y^*$’—could appear on a menu with $Z$. Suppose that the options $\langle X, Y \rangle$ on a menu $\{X, Y\}$ are similar to the options $\langle X^*, Y^* \rangle$ on a larger menu $\{X^*, Y^*, Z\}$ in all respects which are relevant to rational choice. In that case, I’ll say that the options $\langle X, Y \rangle$ are the same as the options $\langle X^*, Y^* \rangle$. (Thus, I distinguish between options being the same and options being identical.) It’s only when the options $\langle X^*, Y^* \rangle$ on the menu $\{X^*, Y^*, Z\}$ are the same as the options $\langle X, Y \rangle$ on the restricted menu $\{X, Y\}$ that $Z$ is a truly irrelevant alternative to $X$ and $Y$. So we can more carefully state the principle IIAS like this:

**IIAS** If it is impermissible to choose $X$ from the menu $\{X, Y\}$ and the options $\langle X^*, Y^* \rangle$ are the same as the options $\langle X, Y \rangle$, then it is impermissible to choose $X^*$ from the larger menu $\{X^*, Y^*, Z\}$.

Similarly, we may more carefully state the principle NEE like this:

**NEE** If you are certain to remain rational and your beliefs and desires are certain to not change, it is permissible to not choose $X$ from the menu $\{X, Y, Z\}$, $X^*$ is the same as the option $X$, and $\langle Y, Z \rangle$ are the same as $\langle Y^*, Z^* \rangle$, then, given a choice between $X^*$ and going on to choose between $Y^*$ and $Z^*$, it is permissible to not choose $X^*$.

When I argued that Minimal CDT violated IIAS and NEE above, I was implicitly assuming that taking box $A$ and taking box $B$ were the same options whether they were the only available options or it was also an option

---

§3. Options

to take box C. In general terms, I was assuming that two tuples of options, \( \mathbf{X} = \langle X_1, X_2, \ldots, X_N \rangle \) and \( \mathbf{X}^* = \langle X_1^*, X_2^*, \ldots, X_N^* \rangle \), in two different decision problems, are the same iff (a) you desire each of the options \( X \in \mathbf{X} \) to the same degree as its corresponding option \( X^* \in \mathbf{X}^* \), in every possible state of nature, and (b) your subjective probability distribution over states, conditional on each \( X \in \mathbf{X} \), is the same as your subjective probability distribution over states, conditional on the corresponding \( X^* \in \mathbf{X}^* \).\(^{19,20}\) Call this ‘the simple view’ of option individuation.

**The Simple View** Given two different decisions, with menus of options \( \mathbb{M} \) and \( \mathbb{M}^* \), and given any \( n \)-tuple of options, \( \mathbf{X} \), from the menu \( \mathbb{M} \), an \( n \)-tuple of options \( \mathbf{X}^* \) from the menu \( \mathbb{M}^* \) are the same as \( \mathbf{X} \) iff, for every state of nature \( K \),

\[
\begin{align*}
(a) & \ D(K, \mathbf{X}) = D(K, \mathbf{X}^*) ; \\
(b) & \ Pr(K | \mathbf{X}) = Pr(K | \mathbf{X}^*) .
\end{align*}
\]

This is a natural and plausible way of saying when two collections of options are the same. According to it, beer and soda are relevantly different options when they are the only options 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.

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 \( \mathbf{X} \) and \( \mathbf{X}^* \) are the same iff, for each \( X \in \mathbf{X} \) and its corresponding \( X^* \in \mathbf{X}^* \):

\[
\begin{align*}
(a) & \ D(K, \mathbf{X}) = D(K, \mathbf{X}^*) ; \\
(b) & \ Pr(K | \mathbf{X}) = Pr(K | \mathbf{X}^*) ; \\
(c) & \ Pr(X) = Pr(X^*) .
\end{align*}
\]

This would reconcile Minimal CDT with IIA, 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 a larger menu \( \{X^*, Y^*, Z\} \), the only way for \( \langle X^*, Y^* \rangle \) to be the same as \( \langle X, Y \rangle \) would be for \( Pr(Z) \) to be zero. On this proposal, 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.

\[19.\] If \( X \in \mathbf{X} \) is the \( k \)th option in \( \mathbf{X} \), then the corresponding option in \( \mathbf{X}^* \) is the \( k \)th option in \( \mathbf{X}^* \).

\[20.\] Notation: I am abusing set-theoretic notation, applying it to tuples instead of sets. So ‘\( X \in \mathbf{X} \)’ says that \( X \) is one of the options in the tuple \( \mathbf{X} \).

\[21.\] To be clear: when I write equations like ‘\( D(K, \mathbf{X}) = D(K, \mathbf{X}^*) \)’ and ‘\( Pr(K | \mathbf{X}) = Pr(K | \mathbf{X}^*) \)’, the desire and probability functions on the left should be understood to be the desire and probability functions from the first decision problem, \( d \), and those on the right should be understood to be the desire and probability functions from the second decision problem, \( d^* \).
Alternatively, we may wish to say that your (unconditional) state probabilities help to determine when some collection of options are the same. That is, we may suggest: $X$ and $X^*$, in two different decision problems, are the same options iff, for each $X \in X$ and its corresponding $X^* \in X^*$: a) $\mathcal{D}(XK) = \mathcal{D}(X^*K)$ for each $K$; b) $\Pr(K \mid X) = \Pr(K \mid X^*)$ for each $K$; and c) each $K$ has the same unconditional probability, $\Pr(K)$. 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 \textit{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 (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 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 options $\langle A, B \rangle$ on the menu $\{A, B\}$ from the options $\langle A, B \rangle$ on the larger menu $\{A, B, C\}$. Nonetheless, I think that we \textit{should} distinguish the options $\langle A, B \rangle$ on the first menu from the options $\langle A, B \rangle$ on the second.\footnote{Though I wish to distinguish the option of taking box A when box B is the only alternative from the option of taking box A when C is the only alternative, I’ll continue to use the notation ‘A’ for the option of taking box A, whatever the alternatives happen to be. Context should make it clear which options I am talking about.} In §3.2, I’ll offer the causalist a different account of when two collections of options are the same (§3.2). I will then explain how this account allows causalists to dispute the charge that they violate IIA and
§3. Options

NEE in Utility Cycle (§4).

3.2 Utility Profiles. 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 options which do not change over the course of deliberation. In particular: the utilities you would assign to each option, were you to learn that you’d chosen any of the options—that is, the values $U_Y(X)$, for each $X$ and $Y$—do not change as you deliberate. And my suggestion to causalists is that they individuate a collection of options, $X$, 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 potential post-choice perspectives from which to evaluate the utilities of $X$ and $Y$—these are the perspectives encoded in $U_X$ and $U_Y$. Given a choice between $X$, $Y$, and $Z$, there are three: $U_X$, $U_Y$, and $U_Z$. If this is enough to make $(X, Y)$ count as different options, then it will be impossible for the same options to appear on two different menus, and IIA will impose no constraint at all.

So let us proceed carefully. Suppose you face a decision problem with the menu of options $M$, and the options in the tuple $X = \langle X_1, X_2, \ldots, X_N \rangle$ are all on the menu $M$. Let us define $X$’s utility profile, on the menu $M$—which I’ll write $U_M(X)$—to be the set of the utilities which are assigned to the options $X_i$ from the perspective you’d occupy after having selected any of the options $Y \in M$.

$$U_M(\langle X_1, X_2, \ldots, X_N \rangle) \overset{\text{def}}{=} \{U_Y(X_1), U_Y(X_2), \ldots, U_Y(X_N) \mid Y \in M\}$$

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 utilities of steak and chicken do not depend upon whether you order steak, chicken, or fish. So we may say that steak and chicken are the same options whether they are the only options on the menu or whether fish is also an option. 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 ‘O’, which is guaranteed to contain the same amount of money as $L$. If it was predicted that you’d take $L$,
then there is $100 in both \( L \) and \( O \) and $110 in \( M \). If it was predicted that you’d take either \( O \) or \( M \), then there’s $10 in \( M \) and nothing in either \( L \) or \( O \). As in the original \textsc{Newcomb}, these predictions are 90% reliable.\textsuperscript{23} This additional option will not affect the utility profile of the options \( \langle L, M \rangle \). For the perspective on \( L \) and \( M \) which you’d have after learning that you’d chosen \( O \), then there’s $10 in \( M \) and nothing in either \( L \) or \( O \). As in the original \textsc{Newcomb}, these predictions are 90% reliable.\textsuperscript{23}

This additional option will not affect the utility profile of the options \( \langle L;M \rangle \). For the perspective on \( L \) and \( M \) which you’d have after learning that you’d chosen \( O \) is precisely the same as the perspective on \( L \) and \( M \) you’d have after choosing \( M \). So \( U_{\langle L;M \rangle}(\langle L,M \rangle) = U_{\langle L,M;O \rangle}(\langle L,M \rangle) = \{(90,100),(10,20)\}. And we will be able to say that taking box \( L \) and taking box \( M \) are the same options before and after box \( O \) is added. Thus, \( O \) will be an irrelevant alternative, and \textit{IIA} will tell us that, if \( L \) is impermissible to select in the original \textsc{Newcomb}, it is also impermissible to select when \( O \) is included on the menu of options.\textsuperscript{24}

More carefully, the suggestion is:\textsuperscript{25}

\textbf{Same Option} Given two different decisions, with menus of options \( M \) and \( M^* \), and given any \( n \)-tuples of options \( X \) from the menu \( M \), an \( n \)-tuple of options \( X^* \) from the menu \( M^* \) are the same as \( X \) iff, for every state of nature \( K \),

\begin{enumerate}
\item \( D(XK) = D(X^*K) \); \vspace{1mm}
\item \( \Pr(K \mid X) = \Pr(K \mid X^*) \); and \vspace{1mm}
\item \( U_M(X) = U_{M^*}(X^*) \).
\end{enumerate}

4 Escaping the Cycle

If \textbf{Same Option} is accepted, then the options \( \langle A, B \rangle \) on the restricted menu \( \{A, B\} \) will be different from the options \( \langle A, B \rangle \) on the expanded menu \( \{A, B, C\} \).

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

\textsuperscript{24} \textit{Objection}: Suppose that, conditional on choosing \( O \), you are 100% sure that there’s $0 in \( L \) and \( O \) and $10 in \( M \). Then \( O \) will introduce a new potential post-choice perspective on \( \langle L, M \rangle \). But \( O \) should still be treated as an irrelevant alternative, and you should still choose \( M \) once it is added. \textit{Reply}: I agree that you should still choose \( M \) in this decision problem, and that, in some good sense of ‘irrelevant’, \( O \) is an irrelevant option (you certainly shouldn't choose it), but I don't take it to be an objection to \textbf{Same Option} that it, together with \textit{IIA}, does not tell us so. We can't expect weak principles like \textit{IIA} to tell us \textit{everything}; it is enough that they tell us something non-trivial.

\textsuperscript{25} Notice that the following kind of situation is possible: it could be that (a) the 1-tuples \( \langle X \rangle \) and \( \langle X^* \rangle \) share a utility profile, (b) the 1-tuples \( \langle Y \rangle \) and \( \langle Y^* \rangle \) share a utility profile, yet (c) the pairs \( \langle X, Y \rangle \) and \( \langle X^*, Y^* \rangle \) do not share a utility profile. So, given \textbf{Same Option}, \( X \) could be the same option as \( X^* \), and \( Y \) could be the same option as \( Y^* \) without \( \langle X, Y \rangle \) being the same options as \( \langle X^*, Y^* \rangle \). It's therefore important that we formulated \textit{IIA} as we did, requiring that the \textit{tuple} \( \langle X, Y \rangle \) be the same as the \textit{tuple} \( \langle X^*, Y^* \rangle \), and not just that \( X \) and \( Y \) be the same as \( X^* \) and \( Y^* \) individually.
On the restricted menu, the utility profile of \( \langle A, B \rangle \) only contains perspectives from which the utility of \( B \) exceeds the utility of \( A \), 
\[
U_{\langle A, B \rangle}(\langle A, B \rangle) = \{\langle 0, 70 \rangle, \langle -70, 0 \rangle\}.
\]
Whereas, on the expanded menu, the utility profile of \( \langle A, B \rangle \) contains an additional perspective from which the utility of \( A \) exceeds the utility of \( B \), 
\[
U_{\langle A, B, C \rangle}(\langle A, B \rangle) = \{\langle 0, 70 \rangle, \langle -70, 0 \rangle, \langle 70, -70 \rangle\}.
\]
So, according to Same Option, the options \( \langle A, B \rangle \) are relevantly different when they appear on the expanded menu alongside \( C \). This means that, if options are individuated according to Same Option, then Minimal CDT does not violate IIA in Utility Cycle.

For similar reasons, individuating options with Same Option means that Minimal CDT does not violate NEE. For, on the menu \( \{A, \sim A\} = \{A, \{B, C\}\} \), \( A \)’s utility profile contains no negative values, \( \{\langle 0 \rangle, \langle 70 \rangle\} \).26 (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 \( U_{\sim A}(A) \) is equal to \( U_{C}(A) \).) But, on the full menu \( \{A, B, C\} \), \( A \)’s utility profile is \( \{\langle -70 \rangle, \langle 0 \rangle, \langle 70 \rangle\} \). 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 should object to my earlier claim that those who abide Minimal 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 Minimal CDT will lose $60, on average, in the sequential decisions from §2.3, while those who are always indifferent between \( A, B, \) and \( C \) given 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, evidentialists will end up predictably poorer than causalists.27 More convincingly, they can object to using outcomes in sequential decisions to evaluate the rationality of agents who are incapable of binding their future selves to a certain course of action. The temporal parts of these agents are like separate agents, each facing their own, separate decisions, and incapable of coordinating their actions. The fact that such agents can be led to predictable ruin through a series of rational choices is just an intrapersonal tragedy of the commons.28 (We may think that intrapersonal tragedies of the commons are not possible, because we

26. I’m sloppily conflating options and 1-tuples containing those options.
27. See Wells (2019) and Ahmed (2020).
28. See Arntzenius et al. (2004) for further defense of this view, and see Meacham (2010) for a reply. See also Ahmed (2014b, §7.4.3) and Spencer (forthcominga, §5).
think that the rationality of later choices is importantly constrained in some way by which choices were made earlier, and for which reasons. 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

The question of when causalists should count 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. As I briefly mentioned in §1.3 above, there are 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, and some have suggested heterodox causalist theories of rational choice to treat these cases. An objection which has been raised to some of these heterodox theories is that they run afoul of the _IIA_. One important upshot of our discussion here is that this criticism is misplaced. Apparent violations of _IIA_ arise in similar ways for orthodox CDT; and the solution I’ve proffered causalists is available to the heterodox and orthodox both. Moreover, while this solution allows the heterodox causalist theory I favor to _always_ satisfy _IIA_ and _NEE_, the same cannot be said for orthodox CDT.

Recall, in _Self-Reinforcing Choice_, choosing either option would give you the good news that your choice will make things better than the alternative would—\( U_X(X) > U_X(Y) \) and \( U_Y(Y) > U_Y(X) \). For a concrete case like this, consider:

**Cake in Damascus**

You must choose whether to go to Damascus or Aleppo. Yester-

---


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

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

33. 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'.)
§5. Further Discussion

<table>
<thead>
<tr>
<th>Row</th>
<th>Col</th>
<th>Pr(</th>
<th>Row</th>
<th>Col)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100</td>
<td>0</td>
<td>K_A</td>
<td>70%</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>100</td>
<td>K_D</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>75%</td>
</tr>
</tbody>
</table>

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.)

day, 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, $U_A(A) = 70 > 30 = U_A(D)$, and $U_D(D) = 75 > 25 = U_D(A)$. 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. I won’t discuss this theory here—see [author] for details. The important point for present purposes is just this: according to this theory, to determine which of $X$ and $Y$ is more choiceworthy, you must look at the quantities $U_X(X), U_Y(X), U_X(Y)$, and $U_Y(Y)$—that is, you must look at exactly the values which appear in the utility profile of $(X, Y)$. Moreover, for this reason, the theory of rational choice I favor will always satisfy IIA and NEE.

Orthodox CDT has a harder time satisfying IIA and NEE in general. Suppose that you always begin deliberation thinking that you are equally likely to select each of the available options. Then, in cases with the structure of Self-Reinforcing Choice, orthodox CDT will violate both IIA and NEE, even when options are individuated with Same Option.34

34. Below, I discuss orthodox CDT’s violation of IIA. For a discussion of its violation of NEE, see Joyce (2018).
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 $U(A) = 42.5$ and $U(D) = 52.5$. 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 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: $U(A) = U(A^*) = 55$ and $U(D) = 45$. 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 the options $\langle A, D \rangle$ have the same utility profile in both of these choices, CDT violates IIA, given that options are individuated with Same Option. (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; though, in my view, we should want to hold on to this plausible principle if we can.)

Heterodox causalist theories like mine have been criticized for violating the independence of irrelevant alternatives. It is therefore worth noting that 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 in cases like Cake in Damascus, my theory of rational choice will never violate IIA, once options are individuated with Same Option.

\[35. \text{See, for instance, the discussion in Wedgwood (2013), Bassett (2015), and Barnett (ms).}\]
6 Conclusion

In summation, choices like Utility Cycle afford us three arguments against 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 (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 prevents the principles from being trivialized and prevents Minimal CDT from violating the principles in Utility Cycle.

The final objection: in sequential decision problems, those who abide Minimal CDT will end up predictably poorer than those who follow EDT, even when they have exactly the same amount of money in front of them, sitting in exactly the same place. In response to this objection, I’ve counseled causalists to accept this consequence of their view as an unfortunate intrapersonal tragedy of the commons—avoidable by those lucky agents capable of binding their future selves. Accepting this consequence means that those of us who like diachronic Dutch book arguments will have to be much more careful about how we formulate them. Accepting this consequence also means rejecting another recent argument against EDT. Wells (2019) argues against EDT by contending that, if I predictably make less money than you do in a sequential decision when we hold fixed the state of nature, 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 an argument like Wells’s means abandoning Minimal CDT.36

36. See also Ahmed (2020)’s criticism of Wells (2019).
References


BARNETT, DAVID JAMES. ms. “Graded Ratifiability.” [6], [20], [22]


WEDGWOOD, Ralph. 2013. “Gandalf’s solution to the Newcomb Problem.” *Synthese*, vol. 190 (14): 2643–2675. [6], [13], [20], [22]

