Decision and Foreknowledge

J. DMITRI GALLOW

1 | INTRODUCTION

My topic is how to make decisions when you possess foreknowledge of the consequences of your choice. Many have thought that these kinds of decisions pose a distinctive and novel threat to causal decision theory.[1] Lewis (1981) says that they are "much more problematic for decision theory than the Newcomb problems".[2] Price (2012) thinks that these decisions should push causalists towards a subjectivist theory of causation. Hitchcock (2016) and Stern (forthcoming) have proposed amendments to causal decision theory which only apply in cases of foreknowledge. And Spencer (2020) denies the possibility of certain kinds of foreknowledge specifically in order to rescue causal decision theory from cases he views as counterexamples.

My thesis is that causal decision theory does not face any new problems from decisions involving foreknowledge. Some of the purported problem cases are not problems. Others are problems, but not problems for causal decision theory. They are instead problems for our theories of subjunctive supposition. Other of the purported problem cases *are* problems for causal decision theory, but they are not *new* problems for causal decision theory. They are old problems transposed into a new key.

Nonetheless, reflection on decisions made with foreknowledge can teach and vividly illustrate important lessons for causalists. In particular, there are four lessons causalists should learn from, and bear in mind when considering, decisions made with foreknowledge. Firstly, our intuitive judgements about the instrumental value of our choices can be misled when we have control over our rational credence that ϕ without having control over whether ϕ. Secondly, we should distinguish the probability of an outcome, were you to choose an option, from that chance of that outcome, conditional on you choosing the option. Thirdly, our intuitive judgements about the instrumental value of our choices can be misled when we don't have control over our rational cre-

Draft of March 24, 2021; Word Count: 14,482
✉: dmitri.gallow@acu.edu.au

1. English doesn't distinguish between the situation you face when deciding between options, and the option you end up selecting. It allows us to use 'decision' and 'choice' for both. To avoid confusion, I'll adopt the terminological convention of always using 'decision' for the situation you face, and 'choice' for the selection you make. Thus: you *face* a decision, and *make* a choice.

2. p. 18.

dence that ϕ, but we do have control over whether ϕ. And fourthly, you should not always take your foreknowledge for granted when deliberating about what to do.

Some of these lessons require the extravagance of foreknowledge to learn. Others should be familiar from less exotic decisions involving reliable prediction. Nonetheless, decisions made with foreknowledge provide a particularly striking and evocative illustration of the lessons. And all of the lessons are needed to navigate the unfamiliar and confusing terrain faced by agents with foreknowledge. Once we've appreciated these lessons, we are left with a version of causal decision theory which faces no novel threats from foreknowledge.

2 | FOREKNOWLEDGE

Foreknowledge is not just knowledge of the future, nor knowledge of the future consequences of your choice. Most knowledge like this is unremarkable. I know that it will snow next winter in Toronto, that I'll make mapo tofu for dinner tonight, and that I'll enjoy it. As I'll use the term here, what makes your knowledge that ϕ *fore*knowledge is that ϕ is at least partly about the future, and your belief that ϕ is caused by events in the future which ϕ is about. In my view, rational choice is not a function of what you know, but instead of what credences have been rationalised by your information. So, more generally, if you have some *information* about ϕ, I will call this information 'foreknowledge' iff ϕ is at least partly about the future, and the information is caused by the future events which ϕ is about. (Thus, as I use the term, some, but not all, foreknowledge is knowledge.)

In the foregoing, I said what it was for you to *have* foreknowledge. Even if you think that foreknowledge like this is impossible, you needn't deny that a rational agent could be rational in taking seriously the possibility that they have foreknowledge. And my interest is in how you should choose in circumstances like these. Throughout, then, I will be focused on decisions in which it is a live epistemic possibility for you that you have foreknowledge. Whether or not you really *do* have foreknowledge is irrelevant.

I'll say that whatever foreknowledge you may possess comes to you from *the oracle*. This is a stipulative use of the term. 'The oracle' could be a time traveller, fortune teller, crystal ball, angel, demon, or prophetic dream. You take the oracle's knowledge of the future to be like an ordinary knower's knowledge of the past and present. Whereas ordinary humans perceive and recall only what is or has been, she perceives and recalls what will be. Perhaps her eyes have receptors for tachyons instead of photons. Perhaps she is a time traveller. Perhaps God whispers news from the future into her ear. The mechanism is unimportant. What makes someone an oracle is just that: 1) her prophesies are in general about events which have yet to unfold, events which lie in the future of the prophesy; 2) her prophesies are, in general, caused by the events they are about; and 3) in general, her prophesies are accurate. It doesn't matter for my purposes whether such an oracle really could exist (though, in fact, I think she could). All that matters for my purposes is that it could be rational for you to take seriously the possibility that she does.

By the way, I'm going to take for granted that the future is not *open* in any interesting or controversial sense of the term—and I'm going to take it for granted that you *also* take this for granted. There are interesting asymmetries between past and future.[3] But I'll assume that facts about what will happen are just as metaphysically fixed, determinate, and unchanging as facts about what has happened—and I'll also assume that you will assume this throughout any rational deliberation about how to choose. If time branches or changes (over *hyper*time, perhaps), then it is less clear in what sense you could have foreknowledge of your future. You may have knowledge of what *may be* in your future, along one open branch, or what *was* in your future at some earlier hypertime, but this needn't be knowledge of what *must be* in your future on every open branch, or what *is* in your future, at your current hypertime.

3 | CAUSAL DECISION THEORY

When you face a decision, you'll have some collection of available *acts*, $\mathcal{A}$, between which you must choose. And there will be some collection of relevant ways things might be—let's call these ways things might be 'worlds', and denote their collection with '$\mathcal{W}$'.[4] I'll suppose throughout that both $\mathcal{A}$ and $\mathcal{W}$ are finite.[5] I'll also suppose that you have a subjective probability, or *credence*, function, C, defined over every set of worlds from $\mathcal{W}$. For every act $\mathrm{A} \in \mathcal{A}$, there will be a set of worlds in which you choose A. I'll refer to that set of worlds with 'A' (in italics). Then, your credence that you'll choose A is given by $C(A)$. I will also assume that, for every world $w \in \mathcal{W}$, there is a degree to which you desire that w is actual, which I'll call the *desirability* of w, and I'll write '$\mathcal{D}(w)$'.

Causal decision theory says that, when you face a decision, you should make your choice by considering how desirable things would be, were you to choose each $\mathrm{A} \in \mathcal{A}$. In some decisions, there may be a unique possibility which you think would result, were you to choose A. In other decisions, you may not be sure which possibility would result, were you to choose A. Think about an act like flipping a coin. Suppose that, in fact, you don't flip the coin. Then, if you were to flip it, there would be a 50% chance that it would land heads and a 50% chance that it would land tails. The principle of 'conditional excluded middle' (CEM) nonetheless insists that the coin would either land heads if it were flipped, or else it would land tails if it were flipped. I accept CEM, so I think that there is a unique possibility which would have resulted, were the coin flipped.[6] But I don't think that this possibility is relevant to decision theory, since

3. See, *e.g.*, Horwich (1987) and Albert (2003).

4. More carefully, $\mathcal{W}$ contains ways for things to be which you can't know *a priori* to be (metaphysically) impossible. As the subsequent examples should make clear, worlds should be mutually exclusive and jointly exhaustive, but they need not settle every matter of fact. It is enough for them to settle every *relevant* matter of fact.

5. If $\mathcal{W}$ is infinite, then the expectations I write out below should be exchanged for integrals in the usual way.

6. For more on CEM, see Lewis (1973), Stalnaker (1975), and Hájek (ms).

you don't know whether the coin would have landed heads or tails, had you flipped it. (Indeed, in my view, you *can't* know this, since it's indeterminate whether it would have landed heads or tails.) So, when you're thinking about what would have happened, had you flipped the coin, you'll have to consider *both* a world in which the coin lands heads *and* a world in which it lands tails. And you'll have to take them into account in proportion to your probability that they would result, were you to flip the coin.

So let us introduce a function, $would_A$, which tells us which worlds you think would result, were you to choose A at each possibility. You hand $would_A$ a world, w, and it hands you back a probability distribution over $\wp(\mathcal{W})$, $would_A(w)$. The interpretation of this probability distribution is that $would_A(w)(w^*)$ is how likely you think it is that w^* would result, were you to chose A at the world w. Since A would certainly result, were you to choose A, we can stipulate that $would_A(w)(A) = 1$. A function like this is standardly called an *imaging* function.

I'm going to present CDT in a slightly non-standard way that turns out to be a bit easier to work with, and which in my opinion makes its commitments about instrumental value easier to appreciate. It involves a bit of linear algebra, but just the tiniest bit. All you need to know in order to check my math is how the multiply matrices together, and even if you can't follow the math, I hope you will be able to follow the philosophical discussion. Fix some enumeration of the worlds in $\mathcal{W}$, $w_1, w_2, \ldots, w_N$. Then, let 'C' be a $1 \times N$ vector whose ith column is your credence in the world w_i. That is: $C = [C(w_1), C(w_2), \ldots, C(w_N)]$. Let '$\mathcal{D}$' be an $N \times 1$ vector whose ith row is the desirability of the world w_i, $\mathcal{D} = [\mathcal{D}(w_1), \mathcal{D}(w_2), \ldots, \mathcal{D}(w_N)]'$.[7] And let '$would_A$' be an $N \times N$ matrix whose entry in the ith row and the jth column is $would_A(w_i)(w_j)$. Then, CDT says that the choiceworthiness of an act, A, is measured by its *utility*, $\mathcal{U}(A)$, where:

$$\mathcal{U}(A) \stackrel{\text{def}}{=} C \cdot would_A \cdot \mathcal{D}$$

(Here, '$\cdot$' is matrix multiplication.) That is: the utility of A consists of three ingredients: your credences, your desires, and information about what would happen, were you to choose A.[8] Multiply these ingredients together, and you get A's utility.

CDT says that you should choose an option which maximises utility. As will become clear below, this definition of utility is somewhat under specified. The reason is that there are a variety of ways of specifying the imaging function $would_A$, and different choices lead to different definitions of utility. For instance, Lewis (1981) effectively understands $would_A(w)(w^*)$ to be $Ch_w(w^* \mid A)$, where Ch_w is the objective chance function at w at the time of choice.[9] Others, like Sobel (1994) and Rabinow-

7. V' is the transpose of the vector V.

8. In order to count as a version of *causal* decision theory, the imaging function $would_A$ must be understood as providing us with information about what likely *would* happen, *were* you to choose A—and not just information about what likely *is* the case, *if* you choose A. For instance, if we set $would_A(w)(w^*) = C(w^* \mid A)$, then utility reduces to 'news-value' and we get back *evidential* decision theory (introduced below).

9. In fact, Lewis says that $would_A(w)(w^*)$ is $C(w^* \mid AK_w)$, where K_w is the causal dependency hypothesis true at world w. K_w will *entail* information about the chance of w^*, given that you choose A, so Lewis (1980)'s

ICZ (1982, 2009), will want to understand the imaging function differently.[10] In §5 below, I will argue that causalists should understand the imaging function in a way that none of these authors have understood it. This point is important for understanding my thesis. My thesis is that *causal decision theory* does not face any new problems from decisions involving foreknowledge. This doesn't mean that, e.g., Lewis's version of causal decision theory doesn't face any new problems from these decisions. Indeed, I think that decisions involving foreknowledge have an important lesson to teach us about how to understand the imaging function, and that this lessons teaches us that most extant versions of CDT should be rejected.

Notice that, by the associativity of matrix multiplication, it doesn't matter whether we group the imaging function $would_A$ with your credences, C, or with your desirabilities, $\mathcal{D}$. A standard presentation of CDT takes the former route, saying that

$$\mathcal{U}(A) = C_A \cdot \mathcal{D}$$

where $C_A \stackrel{\text{def}}{=} C \cdot would_A$ is your credence function *imaged on* the performance of A.[11] One way of thinking about the imaging function $would_A$, then, is as an important ingredient in a subjunctive analogue to conditioning. While your credences conditioned on A—which we can write '$C\,|\,A$'—tell you how likely each possibility is on the *indicative* supposition that you've performed A, your credences imaged on A, C_A, tell you how likely each possibility is on the *subjunctive* supposition that you've performed A. Then, CDT tells you to evaluate an act by looking at the expectation of desirability, $\mathcal{D}$, where the expectation is taken with respect to these 'subjunctive credences', C_A. This is to be contrasted with *evidential* decision theory, EDT, which tells you to evaluate an act by looking at the expectation of desirability, $\mathcal{D}$, when the expectation is taken with respect to the your credences *conditioned* on A, $C\,|\,A$. That is, according to EDT, you should evaluate act in terms of their *news-value*, $\mathcal{V}$, where

$$\mathcal{V}(A) \stackrel{\text{def}}{=} C\,|\,A \cdot \mathcal{D}$$

I think this standard presentation has unfortunately led to some misunderstandings. In particular, it has led some to confuse these 'subjunctive credences' with your actual credences, and to think that, according to CDT, whether you choose A cannot give you any evidence about whether you were predicted to choose A.[12] I think that CDT's philosophical commitments are clearer if we group '$would_A$' with your desirabilities, $\mathcal{D}$. So let us define $\mathcal{D}_A \stackrel{\text{def}}{=} would_A \cdot \mathcal{D}$, which gives us the desirability of what A's performance would bring about. That is, $\mathcal{D}_A(w_i)$ is the desirability of what choos-

principal principle will imply that $C(w^* \mid AK_w)$ should be $Ch_w(w^* \mid A)$.

10. See BALES (2016) for a nice overview of the differences between Lewis, Sobel, and Rabinowicz.

11. CDT is presented in terms of imaging your credences on A in LEWIS (1981), SOBEL (1994), and JOYCE (1999).

12. This misunderstanding of CDT is most pronounced in MEEK & GLYMOUR (1994).

ing A would bring about at world w_i.[13] Then, CDT says that the choiceworthiness of an act, A, is given by your expectation of this quantity. That is,

$$\mathcal{U}(A) = C \cdot \mathcal{D}_A$$

This makes it clearer that, according to CDT, your credences should never be, at any point in rational deliberation, updated by *imaging* them on the performance of an act. Of course, as you deliberate, you may learn that you're more likely to select the act A. In that case, you should change your credences by responding this new information and becoming more confident in A. However, you should do that by conditioning (or Jeffrey conditioning) on this new information, not by updating your degrees of beliefs in any other way.

As this formulation makes clear, CDT's distinctive feature is a certain theory of instrumental value. According to CDT, the only thing which gives an act instrumental value is what it would do to bring about desirable ends. A nice illustration of this commitment is given by NO DIFFERENCE:

> NO DIFFERENCE
> Before you are two boxes. You may either take the box on the left, 'Lefty', or the box on the right, 'Righty'. There is no difference between the boxes. They both contain exactly the same amount of money. Their contents were decided yesterday on the basis of a prediction about what you would freely choose in this decision. If it was predicted that you would freely choose Lefty, then both boxes contain \$100. If, on the other hand, it was predicted that you would freely choose Righty, then both boxes contain \$10. You take these predictions to be *very* reliable.

In NO DIFFERENCE, EDT says that you have decisive reason to take Lefty. There are four relevant possibilities to consider: the world where there are one *h*undred dollars in the boxes and you take *L*efty, w_{HL}, the world where there are a hundred dollars in the boxes and you take *R*ighty, w_{HR}, the world where there are *t*en dollars in the boxes and you take Lefty, w_{TL}, and the world where there are ten dollars in the boxes and you take Righty, w_{TR}. Because you take these predictions to be so accurate, let's just assume that you are certain that you're either at w_{HL} or you're at w_{TR}. Then, your credence in w_{HL} is just your credence that you'll take Lefty, $C(L)$, and your credence in w_{TR} is just your credence that you'll take Righty, $C(R)$. Assuming that your desires are linear in dollars, the news value of taking Lefty is 100, and the news value of taking Righty is 10. Learning that you've taken Lefty is better news that learning that you've taken Righty. So EDT requires you to take Lefty.

In contrast, CDT says that you have no more reason to take Lefty than you have reason to take Righty. It says that, when you evaluate your options for choicewor-

13. To be clear: '$\mathcal{D}_A(w_i)$' is the ith row in the vector $\mathcal{D}_A$. To be slightly more precise, $\mathcal{D}_A(w_i)$ gives the *expected* desirability of what A's performance would bring about at world w_i.

	w_{HL}	w_{HR}	w_{TL}	w_{TR}
w_{HL}	1	0	0	0
w_{HR}	1	0	0	0
w_{TL}	0	0	1	0
w_{TR}	0	0	1	0

(a) $would_L$

	w_{HL}	w_{HR}	w_{TL}	w_{TR}
w_{HL}	0	1	0	0
w_{HR}	0	1	0	0
w_{TL}	0	0	0	1
w_{TR}	0	0	0	1

(b) $would_R$

FIGURE 1: In table 1a, the matrix $would_L(row)(column)$, which describes what would happen at each world, were you to select Lefty. In table 1b, the matrix $would_R(row)(column)$, which describes what would happen at each world, were you to select Righty.

thiness, you should ask yourself: what would each of the options bring about? You don't know for sure, since you don't know for sure what's inside the boxes, but you know that, if there's $100 in the boxes, then taking either box would get you $100, and if there's $10 in the boxes, then taking either box would get you $10. Changing your choice of box wouldn't change the boxes' contents. Formally, $would_L$ and $would_R$ are the matrices shown in figures 1a and 1b, respectively. Those matrices tell us that, had you taken Lefty at world w_{HL} or w_{HR}, there'd still be $100 in both boxes, so you would be at w_{HL}. And had you taken Lefty at world w_{TL} or w_{TR}, there would still be $10 in both boxes, so you would be at w_{TL}. Likewise, had you taken Righty at world w_{HL} or w_{HR}, there'd still be $100 in both boxes, so you would be at w_{HR}. And, had you taken Righty at world w_{TL} or w_{TR}, there'd still be $10 in both boxes, so you would be at w_{TR}.

This means that choosing L and choosing R would accomplish exactly the same thing. In any world in which there is $100 in both boxes, both options would get you $100. And in any world in which there is $10 in both boxes, options would get you $10.

$$would_L \cdot \mathcal{D} \;=\; would_R \cdot \mathcal{D} \;=\; [100, 100, 10, 10]'$$

So the utility of taking Lefty is exactly the same as the utility of taking Righty. They are both equal to 100 times your credence that you'll take Lefty plus 10 times your credence that you'll take Righty.

$$\begin{aligned} & \mathcal{U}(L) = 100 \cdot C(L) + 10 \cdot C(R) \\ \text{and} \qquad & \mathcal{U}(R) = 100 \cdot C(L) + 10 \cdot C(R) \end{aligned}$$

So CDT says that you have just as much instrumental reason to take Lefty as you have instrumental reason to take Righty. Importantly, CDT does *not* say that you should be *equally glad* to find yourself choosing either box. If you learn that you are taking Lefty, this is fantastic news: it tells you that there is very likely $100 in the boxes. And if you learn that you are taking Righty, this is terrible news: it tells you that there is very likely only $10 in the boxes. Learning that you are taking Lefty is reason to be glad. And learning that you are taking Righty is reason to be sad. If you care about whether you're glad or sad before you open your chosen box, then causalists will say that you have decisive reason to take Lefty. However, if we stipulate that the *only* thing

you care about is how much money you get, then causalists insist you have no more instrumental reason to choose Lefty than you have to choose Righty. You know for sure that choosing Lefty would get you no more money than choosing Righty would. So, according to causalists, choosing Righty is not the least bit instrumentally irrational.

Suppose you find yourself waffling back and forth between Lefty and Righty. As you incline towards Lefty, you give yourself evidence that you will eventually choose Lefty, so your credence that there's $100 in the boxes goes up. As you incline towards Righty, you give yourself evidence that you will eventually choose Righty, so your credence that there's $10 in the boxes goes up. In this state of indecisive waffling, your rational *opinions* about the money in the boxes will change, depending upon which choice you're most inclined to make. This can make it feel like how much money is in the boxes is under your control—but this is an illusion. *Ex hypothesi*, you have *no* control over how much money is in the boxes. At this point, the *only* thing you have control over is your own rational credences. Taking Lefty would raise your rational credence that there's $100, and taking Righty would raise your rational credence that there's only $10. But you don't care at all about your rational credence that there's $100 in front of you. You only care about the $100 being there or not. And since this won't change, depending upon which box you select, the causalist says that you've no reason to select either Lefty or Righty over the other. For causalists, the only thing that speaks in favour of or against a choice is what that choice would bring about.

As a deliberative agent making this decision, you have control over your rational credence that there's $100 in the boxes, but you do not have control over whether there is $100 in the boxes. This can lead to the agential illusion that you have control over the box's contents. And this agential illusion of control leads to the illusion that taking Lefty is more instrumentally valuable than taking Righty. One way of appreciating that this instrumental value is illusory is by considering the decision from a better informed, third-personal perspective. For instance, suppose that you are watching your friend make a choice between Lefty and Righty. You are allowed to look inside of the boxes, though you're not allowed to communicate with your friend. Notice that, no matter what you see, you will see no more instrumental value in your friend taking Lefty than you see in them taking Righty. So, once you are free of the agential illusion of control, the inclination to say that there's more instrumental value in taking Lefty vanishes.

In fact, from this better-informed, third-personal perspective, a preference for Lefty over Righty appears irrational. Suppose that your friend is an evidentialist, so they think that taking Lefty has an instrumental value $90 greater than taking Righty. Then, your friend would be willing to *pay* up to $90 in order to take Lefty. Suppose, then, that your friend faces a different decision, between paying $80 and taking Lefty or paying nothing and taking Righty. From your perspective, it appears clear that there is no instrumental value in paying this money. You can see into the boxes, and you recognise that they both contain the $100. Paying the $80 is just throwing money away—your friend could easily get $100 by just taking Righty for free. Nonetheless, if your friend is an evidentialist, then they will throw away $80 in order to take the

$100 in Lefty, rather than taking the free $100 in Righty. From your better-informed, third-personal perspective, it is difficult to see this as a rational choice.

There is a general lesson here. It is, without a doubt, *intuitive* that you should take Lefty in NO DIFFERENCE. But causalists diagnose this intuition as arising from the illusion that the amount of money in the boxes is under your control, when in fact—*ex hypothesi*—the only thing that's under your control is your own epistemic state. In the vast majority of cases, if your rational credence that ϕ is under your control, then you have reason to think that whether ϕ may be under your control. So it's understandable that our knee-jerk intuitions would not be sensitive to the distinction between giving yourself evidence that the world is good and *making* the world good. However, cases like NO DIFFERENCE show us that giving yourself evidence that the world is good can come apart from *making* the world good. The case also teaches us that, when your rational credence that ϕ is under your control, but you know for sure that whether ϕ is not, your knee-jerk intuitions about rational action are not to be entirely trusted. You must be on your guard to clearly distinguish the way the world is from the way you have reason to *think* the world is. And one way of being on your guard is by considering the question of what instrumental value a choice has from a better-informed, third-personal point of view.

Lesson #1 When you have control over your rational credence that ϕ, but you know for sure that you do not have control over whether ϕ, your intuitive judgements about rational choice can lead you astray by conflating control over your *epistemic state* with control over *the world*. In these cases, you should consider what instrumental value a choice has when viewed from a better informed, third-personal perspective.

Not everyone is going to learn this lesson from NO DIFFERENCE. But it is a lesson which should be learnt by any causalist deserving the name. Those who refuse to learn this lesson think that you should choose to give yourself good news about the way the world is, even when this has no effect on the world whatsoever. As LEWIS (1981, p. 5) puts it, they "counsel an irrational policy of managing the news so as to get good news about matters which you have no control over".

4 | MANAGING THE NEWS FROM THE FUTURE

Consider the following decision:

> STICKER
> It is Christmas eve. Under the tree are two gifts from Santa: one for you, one for your sister. You know for sure that one of them contains this year's hottest toy, and the other contains a lump of coal, though you don't know which is which. You are absent-mindedly putting stickers on the gifts. These stickers are purely decorateive; they don't make any difference with respect to who gets to open which gift. Before you decide where to place

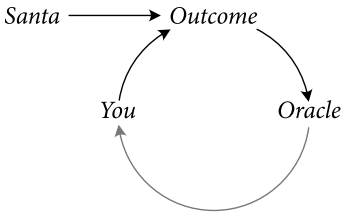

FIGURE 2: The causal structure of STICKER.

> the reindeer sticker, the oracle informs you: the gift which you will put the reindeer sticker on contains the toy.

I first learnt about decisions with this structure from ROBERTS (ms). ROBERTS argues that putting the sticker on your own gift is a rational means of getting the toy; and from this, he concludes that, if you put the sticker on your gift, this *causes* Santa to have gifted you the toy in the past. I'm going to take for granted that this conclusion about the causal structure of the case is incorrect. I'll suppose that the causal structure of STICKER is as shown in figure 2. In that figure, think of *Santa*, *Outcome*, *Oracle*, and *You* as variables which can take on certain values, depending upon what Santa, you, and the oracle say and do, and depending upon which gift has the sticker and which gift has the toy. So the variable *Santa* says whether Santa put the toy in your gift or your sister's (we can suppose that Santa has only one toy to give; the other sibling is getting coal). We can suppose that the oracle will tell you about the contents of the gift with the sticker on it. So the variable *Oracle* says whether the Oracle tells you that the stickered gift has the toy in it, or whether she tells you that the stickered gift has the coal in it. The variable *You* says whether you put the sticker on your gift or your sister's. Finally, *Outcome* says both which gift has the toy and which gift has the sticker.

Santa and *You* causally determine *Outcome*. And *Outcome* causally determines *Oracle*—that is, the oracle is likely to tell the truth about *Outcome*, so whether the gift with the sticker has the toy in it or not causally determines what the oracle tells you. It is possible that the oracle's pronouncement has a causal effect on your choice. For instance, you might have the following dispositions: if the oracle tells you that the gift with the sticker has a toy, then you are disposed to place the sticker on your own gift; and if the oracle tells you that the gift with the sticker has coal in it, then you are disposed to place the sticker on your sister's gift. If you have these dispositions, the oracle's pronouncement will have a causal effect on your choice, whence the arrow from *Oracle* to *You* in figure 2. (Of course, you might also choose in a way which is insensitive to the oracle's pronouncement, whence the arrow is grey, rather than black.)

In this decision, there are four relevant kinds of possibilities. Either Santa gave you a toy, T, or he gave you a lump of coal, C. And either you will put the sticker on your gift, Y, or you will put it on your sister's gift, S. Let 'w_{TY}' be a world in which Santa gave you a toy and you put the sticker on your own gift. Let 'w_{TS}' be a world in which Santa gave you a toy and you put the sticker on your sister's gift. And likewise for 'w_{CY}' and 'w_{CS}'.

	w_{TY}	w_{TS}	w_{CY}	w_{CS}
w_{TY}	1	0	0	0
w_{TS}	1	0	0	0
w_{CY}	0	0	1	0
w_{CS}	0	0	1	0

(a) $would_Y$

	w_{TY}	w_{TS}	w_{CY}	w_{CS}
w_{TY}	0	1	0	0
w_{TS}	0	1	0	0
w_{CY}	0	0	0	1
w_{CS}	0	0	0	1

(b) $would_S$

FIGURE 3: In figure 3a, the matrix $would_Y(row)(column)$, which describes what would happen at each world, were you to affix the sticker to your gift. In figure 3b, the matrix $would_S(row)(column)$, which describes what would happen at each world, were you to affix the sticker to your sister's gift.

There are interesting questions to be raised about what *would* happen in STICKER, were you to choose differently. For instance, in the world w_{TY}, you put the sticker on your gift. At this world, what would have happened, were you to put the sticker on your sister's gift instead? There's a temptation to answer: were you to put the sticker on your sister's gift, the oracle would still have told you that the gift with the sticker has the toy. And since the oracle is making every effort to speak truly, this means that it would have to be the case that the gift with the sticker *did* contain the toy. Since your sister's gift would be the one with the sticker on it, this means that Santa must have given the toy to your sister, and given the coal to you. Therefore, at the world w_{TY}, had you put the sticker on your sister's gift, you'd have been in a world at which Santa gave you coal: w_{CS}. This can seen like a natural way of reasoning, in part because ordinarily, when we think about what would happen, were we to choose differently, we hold fixed our causal past—and, in this case, the oracle's prognostication lies in your causal past. But we plainly cannot hold fixed *all* of your causal past, since, in this case, your choice *also* lies in your causal past. In contrast, we could figure out what would have happened, had you chosen differently, by considering a scenario in which your choice does not *depend* upon its causal past, and then thinking through how the rest of the world might change, were you to choose differently.[14] I'm going to take it for granted here that this second way of thinking about what would happen, were you to choose differently, is the one which is relevant to rational choice. So, at the world w_{TY}, were you to put the sticker on your sister's gift, this decision would not have been influenced by the oracle's pronouncement. Since your sister's gift contains the coal at w_{TY}, this would be a world at which the gift with the sticker contains the coal—and therefore, it would very likely be a world at which the oracle *tells you* that the gift with the sticker has the coal.

More generally, I'll suppose that $would_Y$ and $would_S$ are as shown in figure 3. And I'll suppose that you only care about who gets the toy and who gets the coal—you don't intrinsically desire the sticker being on your gift or your sister's gift. With these assumptions in place, CDT treats STICKER just like NO DIFFERENCE. Putting the sticker

14. In the lingo: we could consider a possibility in which you choose differently as the result of an *intervention*.

on your gift gives you evidence that you'll get the toy, and putting the sticker on your sister's gift gives you evidence that you'll get the coal, but the sticker does nothing at all to change what Santa has gifted you. No matter what gift you have, putting the sticker on your gift and putting it on your sister's gift would bring about exactly the same results. That is, $would_Y \cdot \mathcal{D} = would_S \cdot \mathcal{D} = \mathcal{D}$.

Hitchcock (2016) thinks that CDT gives the wrong verdict about sticker. He thinks that, when you have information about the consequences of your choices, this information needs to be taken into account in a different way than other kinds of information. And so he endorses a variant of CDT for decisions with foreknowledge. To understand Hitchcock's variation of CDT, it helps to first consider how orthodox CDT tells you to take foreknowledge into account. Let's use 'C_0' for your *ur-prior* credence function—the credence function you are disposed to hold in the absence of any evidence. The norm of ur-prior conditionalisation says that your credences at any time should be C_0 conditioned on your total evidence at that time.[15] So, suppose that your total evidence consists of the ordinary evidence E and the foreknowledge F. Then, whenever you're making a choice, your credences will be $C_0 \mid EF$, where 'EF' is the conjunction of E and F. Then, CDT advises you to evaluate acts for choiceworthiness with the expectation

$$\mathcal{U}(A) \;=\; (C_0 | EF) \cdot would_A \cdot \mathcal{D}$$

Hitchcock thinks that this is a mistake. According to him, foreknowledge of the consequences of your choices should not be incorporated into decision-making in the same was as ordinary evidence. Instead, he advises you to evaluate acts for choiceworthiness with:

$$\mathcal{H}(A) \;=\; ((C_0 | E) \cdot would_A) | F \cdot \mathcal{D}$$

That is: you should *first* take into account your ordinary evidence by conditioning C_0 on E, *next* image your probability function on the performance of A, and only *then* take into account your foreknowledge by conditioning the probability function $(C_0 | E) \cdot would_A$ on F. You should then take an expectation of desirability using the resulting probability function, $((C_0 | E) \cdot would_A) | F$.[16]

In sticker, if you most want the toy for yourself and you don't care about the placement of the sticker, then Hitchcock's theory will tell you it is rationally obligatory to put the sticker on your gift. For concreteness, suppose that, in the absence of any evidence, you think Santa was equally likely to gift the toy to you as he was to gift it to your sister, that you're equally likely to put the sticker on either gift, and that

15. I am focusing on conditionalisation for the sake of simplicity, but similar points could be made about the generalisations of conditionalisation proposed by Jeffrey (1965), Schoenfield (2017), or Gallow (forthcomingb).

16. Hitchcock (2016) uses causal Bayes nets rather than imaging to determine this probability function, but this won't make any difference to my discussion here. You can determine a joint probability distribution over the variables *Santa*, *You*, *Outcome*, and *Oracle* using the causal Markov condition in the natural way, and, when it comes to the possibilities described by w_{TY}, w_{TS}, w_{CY}, and w_{CS}, the probability function which results from intervening on *You* in the Bayes net is exactly the same as the probability function which results from imaging on your action.

these choices are independent. Then, $C_0 = [C_0(w_{TY}), C_0(w_{TS}), C_0(w_{CY}), C_0(w_{CS})] = [1/4, 1/4, 1/4, 1/4]$. And your only evidence is the foreknowledge that you put the sticker on the gift with the toy—that is, $F = \{w_{TY}, w_{CS}\}$. Let's take $\mathcal{D} = [\mathcal{D}(w_{TY}), \mathcal{D}(w_{TS}), \mathcal{D}(w_{CY}), \mathcal{D}(w_{CS})]'$ to be $[100, 100, 0, 0]'$.[17] Then,

$$\begin{aligned} \mathcal{H}(Y) &= (C_0 \cdot would_Y) \mid F \cdot \mathcal{D} \\ &= 100 \\ \text{and} \qquad \mathcal{H}(S) &= (C_0 \cdot would_S) \mid F \cdot \mathcal{D} \\ &= 0 \end{aligned}$$

So HITCHCOCK says that it is irrational to place the sticker on your sister's gift—even though the sticker itself does not affect the contents of you or your sister's gifts.

Evidentialists should be pleased with this verdict, but if we are causalists, and we've taken **Lesson #1** to heart, then I think we should side with orthodox CDT over HITCHCOCK's proposed amendment. There is, to be sure, an intuition in this case that you should decide where to put the sticker by considering who you'd rather get the toy. But this is, *ex hypothesi*, a case in which you have no control over who gets the toy. What you *do* have control over is your rational credence that you got the toy. For, by putting the sticker on your own gift, you give yourself very strong evidence that Santa has decided to give you the toy. And, by putting the sticker on your sister's gift, you give yourself very strong evidence that Santa has decided to give your sister the toy. But the sticker is just a sticker. It doesn't change what's on the inside of the packages. **Lesson #1** warns us to be on guard: our intuitive judgements about rational choice can lead us astray in precisely these kinds of cases. While *evidentialists* should be comfortable advising you to manage the oracle's news from the future—giving yourself news that the world is good, even what your choice does absolutely nothing to change the world for the better—causalists should not.

Causalists should also notice that the strong intuition that placing the sticker on your gift has instrumental value vanishes with additional information. Suppose that it is not you making this decision, but instead your sister. While your sister cannot look into the gifts, you can (though you cannot communicate with her). You see that you have been gifted the toy and your sister has been gifted the coal. From this perspective, it does not appear that there is *any* instrumental value in placing the sticker on her package. Moreover, from this better-informed, third-personal point of view, a preference for placing the sticker on either gift can appear irrational. For illustration, suppose that your sister wants you to have the toy (she's just a better person than you are—Santa really should have given the toy to her), and suppose that she subscribes to HITCHCOCK's decision theory. Then, she will think that placing the sticker on your gift has an instrumental value equal to the value of you getting the toy. Suppose that you

17. I've arbitrarily chosen to make 100 the desirability of getting the toy and 0 the desirability of getting the coal. Since desirability is measured on an interval scale, any other pair of numbers where the first is higher than the second would be an equivalent representation of your desirabilities.

getting the toy is worth $100 to your sister, and that she is only allowed to place the sticker on your gift if she pays $90. From your perspective, it appears clear that there is no instrumental value in paying the $90 to place the sticker on your gift. You can look into the gifts and see that you have the toy and your sister has the coal. Paying to put the sticker on your gift is just throwing $90 away for no reason. You'd still have the toy if your sister didn't pay the $90. Nonetheless, HITCHCOCK advises your sister to throw away $90 on a decorative sticker which does nothing to affect whether you get the toy or not. From your better-informed, third-personal point-of-view, it is difficult to see this as a rational choice.

So it seems to me that causalists should say precisely the same thing about STICKER that they say about NO DIFFERENCE: you've no reason to choose either of the options over the other, since neither of the options would make any difference to anything you care about. So causalists have no reason to worry about orthodox CDT's verdicts in cases like STICKER. On the contrary, they should worry about any decision theory, like HITCHCOCK's, which counsels an irrational policy of managing the news from the future so as to get good news about matters over which you have absolutely no control. They should worry about ceding to an agential illusion of control in STICKER after having fought against it in cases like NO DIFFERENCE. In both STICKER and NO DIFFERENCE, one option gives evidence that your desires are satisfied without doing anything to satisfy those desires, and another option gives evidence that your desires are frustrated without doing anything to frustrate those desires. So I have a hard time seeing why a causalist should treat these decisions differently.[18] I agree that there's a strong intuition that it's irrational to put the sticker on your sister's gift. But I *also* think that there's a strong intuition that it's irrational to take Righty in NO DIFFERENCE. Why should we dismiss the verdicts of intuition when an agential illusion of control is due to reliable *prediction*, but not when it is due to reliable *prescience*?

5 | FOREKNOWLEDGE AND CHANCE

Some believe that foreknowledge makes a difference to the chances. For illustration: suppose we are about to perform a quantum coin flip which has an objective, tychistic 50% chance of landing heads and an objective, tychistic 50% chance of landing tails.[19] Before the coin is flipped, the oracle informs you that the coin will land tails. Hearing all this, it's natural to say: you should think that the chance that the coin lands heads is 50%, but your credence that the coin lands heads should be less than 50%. HALL, MEACHAM, and SPENCER object to this description of the case. They say: if the oracle is actually able to foresee the outcome of the coin flip, then the fact that she sincerely

18. As I'm using the term, HITCHCOCK is not a causalist. Rather, he is an evidentialist who accepts a libertarian metaphysics according to which free choices are the consequence of an uncaused Will. For this reason, HITCHCOCK will not want to draw **Lesson #1** from NO DIFFERENCE.

19. More realistically: we're going to use a Stern-Gerlach apparatus to measure the x-spin of an electron in the quantum state $\sqrt{1/2}\,|\,x\text{-spin-up}\rangle + \sqrt{1/2}\,|\,x\text{-spin-down}\rangle$.

prophesies that the coin will land tails must raise the objective chance that the coin will land tails.[20]

HALL and MEACHAM take this position in order to argue that a principle of chance deference need not include an 'admissibility' clause, telling you to only defer to the objective chances when you lack foreknowledge. SPENCER is additionally interested in defending orthodox CDT from some putative counterexamples involving foreknowledge. For instance, consider:[21]

> FOREKNOWN LOSS
> A fair coin will be flipped. Before it is flipped, you are offered a ticket which pays out $150 if the coin lands on heads, and only costs $50. Before you decide whether to buy the ticket, the oracle informs you that the coin will land on tails.

In this decision, HALL, MEACHAM, and SPENCER say that, insofar as you think that the oracle's prophesy is known, you should think that the objective chance of tails is greater than 50%. I disagree. In my view, the objective chance of tails is still 50%, but you should be more confident in tails than heads. SPENCER (2020) thinks that, if we say this, CDT will say that you are rationally required to buy the ticket, in spite of your foreknowledge that the ticket is a loser.

Does CDT say that? That depends upon what would happen, were you to buy the ticket. There are four relevant possibilities. The coin either lands heads, H, or tails, T. And you either *b*uy the ticket, B, or you do *n*ot, N. Let w_{HB} be a possibility at which the coin lands heads and you buy the ticket. Let w_{HN} be a possibility at which the coin lands heads and you do not buy the ticket, and likewise for w_{TB} and w_{TN}. Then, let us suppose that, were you to buy, the coin would be just as likely to land heads as tails, and were you to not buy, the coin would be just as likely to land heads as tails. Then, would_B and would_N are as shown in figure 4. We can assume that your desirabilities are linear with dollars, so that $\mathcal{D} = [\mathcal{D}(w_{HB}), D(w_{HN}), \mathcal{D}(w_{TB}), \mathcal{D}(w_{TN})]' = [100, 0, -50, 0]'$. With these assumptions, CDT says that the instrumental value of buying the ticket is constant—it is worth $25, no matter which world you're at:

$$\mathit{would}_B \cdot \mathcal{D} = [25, 25, 25, 25]'$$

Whereas not buying the ticket has a constant instrumental value of $0:

$$\mathit{would}_N \cdot \mathcal{D} = [0, 0, 0, 0]'$$

20. HALL (1994) and MEACHAM (2010) only consider the case where the fact that the oracle prophesies ϕ *entails* that ϕ. SPENCER (2020) considers the more general case in which the oracle's prophesy that ϕ simply makes it more likely that ϕ.

21. Decisions like these are discussed by PRICE (2012), who uses them to argue for a subjectivism about causation, according to which you have causal control over past events which are correlated with your choice. (For instance, PRICE (2012) agrees with ROBERTS (ms) that, by putting the sticker on your gift, you cause Santa to have given you the toy.)

$$\begin{array}{c} \\ w_{HB} \\ w_{HN} \\ w_{TB} \\ w_{TN} \end{array} \begin{array}{c} \begin{array}{cccc} w_{HB} & w_{HN} & w_{TB} & w_{TN} \end{array} \\ \begin{bmatrix} 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix} \end{array}$$

(a) $would_B$

$$\begin{array}{c} \\ w_{HB} \\ w_{HN} \\ w_{TB} \\ w_{TN} \end{array} \begin{array}{c} \begin{array}{cccc} w_{HB} & w_{HN} & w_{TB} & w_{TN} \end{array} \\ \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 \end{bmatrix} \end{array}$$

(b) $would_N$

FIGURE 4: In figure 4a, the matrix $would_B(row)(column)$, which describes what would be likely happen at each world, were you to buy the ticket. In figure 4b, the matrix $would_N(row)(column)$, which describes what would likely happen at each world, were you to not buy the ticket.

So it won't matter what your credences are, or what foreknowledge you possess. CDT will say that taking the bet is required, and refusing it is impermissible, no matter what the oracle has told you.

I agree with SPENCER that this is terrible advice. Given your foreknowledge, you know that the ticket is a loser, this knowledge is not in any way contingent upon whether you buy the ticket or not, and buying the ticket doesn't make any difference to whether the coin lands heads or tails. So you should not buy the ticket. SPENCER lays the blame on my assumption that, in FOREKNOWN LOSS, the objective chance of the coin landing heads is 50%. In my view, the oracle's prophesy has not altered the chances, so I think that there is a problem here. But the problem is not a problem for CDT. It is instead a problem for our theories of subjunctive supposition. I follow RABINOWICZ (1982, 2009) in laying the blame on the assumption, encoded in $would_B$ and $would_N$, that the coin would be just as likely to land heads as tails, were you to take the bet.

Recall, I defined $would_B(w)(w^*)$ as how likely it is that w^* would result, were you to choose B at the world w. In my treatment of FOREKNOWN LOSS above, I implicitly assumed that, for each option A and each pair of worlds w, w^*, we should set $would_A(w)(w^*)$ to the objective chance of w^*, at world w, conditional on you choosing A. That is, I implicitly assumed that $would_A(w)(w^*) = Ch_w(w^* \mid A)$. SOBEL and RABINOWICZ notice that, in decisions like FOREKNOWN LOSS, this leads to a violation of **Strong Centering**.

Strong Centering If w is a world at which you choose A, then were you to choose A at w, w is the world which would result.

$$\text{if } A \text{ is true at } w, \text{ then } would_A(w)(w) = 100\%$$

The analogue of **Strong Centering** is a consequence of LEWIS's 1973 semantics for counterfactuals. On that semantics, if A is true at w, then so too is $A \mathrel{\square\!\!\rightarrow} w$.[22] In the context of Lewis's semantics for the subjunctive conditional, this imposes the re-

22. I'm using 'w' for the proposition $\{w\}$, which is true at the world w and false at all other worlds.

	w_{HB}	w_{HN}	w_{TB}	w_{TN}
w_{HB}	1	0	0	0
w_{HN}	1/2	0	1/2	0
w_{TB}	0	0	1	0
w_{TN}	1/2	0	1/2	0

(a) $would_B$

	w_{HB}	w_{HN}	w_{TB}	w_{TN}
w_{HB}	0	1/2	0	1/2
w_{HN}	0	1	0	0
w_{TB}	0	1/2	0	1/2
w_{TN}	0	0	0	1

(b) $would_N$

FIGURE 5: If we impose **Strong Centering** on the matrices $would_B$ and $would_N$ from figures 4a and 4b, and make no other changes, we get the matrices above.

quirement that $A \wedge C$ entails $A \boxrightarrow C$.[23] If we accept LEWIS's semantics, and we are thinking of $would_A(w)$ as telling us how likely it is that each world would result, were you to perform A at world w, then it is natural to expect that $would_A$ will satisfy **Strong Centering**. Nonetheless, LEWIS (1986) rejects **Strong Centering** in his formulation of causal decision theory. His motivation for this rejection is not clear to me, and not explicitly explained to the reader.[24] In any case, whatever LEWIS's reasons for disagreeing may have been, I agree with RABINOWICZ that there is good reason to endorse **Strong Centering**.

However, **Strong Centering** on its own will not afford us a satisfactory treatment of FOREKNOWN LOSS. To see why, notice that, if we impose **Strong Centering** on the imaging functions $would_B$ and $would_N$ from figure 4 and make no other changes, then we will get the imaging functions which are shown in figure 5. With these assumptions about $would_B$ and $would_N$, the desirability of what N would bring about will remain unchanged (it's still certain to bring you \$0, no matter what). But the desirability of what B would bring about at each world is now different. At the worlds where you don't take the bet, taking the bet would still bring you \$25 in expectation, but at the worlds where you *do* take the bet, taking it would win you \$100, if the coin lands heads, and lose you \$50, if the coin lands tails. That is:

$$would_B \cdot \mathcal{D} = [\mathcal{D}_B(w_{HB}), \mathcal{D}_B(w_{HN}), \mathcal{D}_B(w_{TB}), \mathcal{D}_B(w_{TN})]' = [100, 25, -50, 25]'$$

You foreknow that you are either at the world w_{TB} or the world w_{HB}. So, if $C(B)$ is

23. For a compelling defence of this principle, known as 'conjunction conditionalisation', see WALTERS & WILLIAMS (2013).

24. He does explain to the reader that, were he to accept **Strong Centering**, he could not accept his formulation of CDT in terms of 'dependency hypotheses'. For RABINOWICZ (1982) proved that, so long as Lewis allows that $would_A(w)$ does not always assign all of its probability to a single world, were he to additionally accept **Strong Cenering**, LEWIS's dependency hypotheses would not always be compatible with your choosing every available option. And LEWIS's formulation of CDT relies upon this always being the case. If there were some antecedent reason to favour LEWIS's formulation of CDT over an alternative formulation like SOBEL's which takes the imaging function $would_A$ as primitive, then this would give us a reason to reject **Strong Centering**. However, if there is a reason like that, then LEWIS (1981) does not provide it. In fact, he spends much of his article trying to persuade the reader that the differences between his formulation and its rivals are inconsequential.

your credence that you will take the bet, then the utility of taking the bet will be

$$\begin{aligned}\mathcal{U}(B) &= -50 \cdot C(B) + 25 \cdot (1 - C(B)) \\ &= 25 - 75 \cdot C(B)\end{aligned}$$

If $C(B) > 1/3$, then the utility of taking the bet will be less than the utility of not taking the bet. So, if you find yourself inclining towards taking the bet (and therefore, you give yourself evidence that you will likely take the bet), then CDT will advise you to *not* take the bet. However, if you listen to this advice, and learn that you have, then your credence that you will take the bet, $C(B)$, will fall below 1/3. And at that point, the utility of taking the bet will *exceed* the utility of not taking the bet, and CDT will advise you to take it.

So, if we simply impose **Strong Centering**, then FOREKNOWN LOSS turns into a case in which CDT's recommendations are sensitive to your own predictions about what you will choose. I think that cases like these pose a problem for CDT (I'll have more to say about them in §6 below). But usually, when CDT exhibits this kind of predictive sensitivity, there is *something* deeply correct about its advice. What's correct in the advice is that, from the perspective you'll occupy when choosing either of the options, you *should* expect that taking the other option would make things better. That's how things usually go in cases where CDT's advice is prediction-sensitive in this way. But that doesn't seem to be the case in FOREKNOWN LOSS. In this decision, when you refuse the bet, you know that the coin will land tails, so you know that taking the bet would lose you money. So, when rejecting the bet, you shouldn't expect that taking the bet would make anything better.

In my view, the problem posed by FOREKNOWN LOSS is a decision-theoretic variant of the problem posed by the so-called 'Morgenbesser conditional', attributed to Sidney Morgenbesser by SLOTE (1978, fn 33). Morgenbesser imagines a situation in which I offer you a bet on whether a flipped coin will land heads, and you decline my offer. I then flip the coin, and it lands heads. At that point,

(MC) If you had taken the bet, you would have won.

appears true, even though, at the time when you had to choose whether to take the bet or not, there was only a 50% chance that the coin would land heads. In my view, conditionals like (MC) show us that, when we make a subjunctive supposition that A, we hold fixed things which are causally independent of whether A, even if those things would have been a matter of chance at the time when A would have happened. Even though it was a matter of chance whether the coin would land heads or tails when you had to make your choice, we hold fixed how the coin landed when we suppose that you had taken the bet. And the reason is that whether you take the bet or not has no effect on how the coin lands.

So I think that, in addition to **Strong Centering**, we should accept the following principle:

Causal Independence If whether ϕ is causally independent of your choice, then ϕ

$$\begin{array}{c} \\ w_{HB} \\ w_{HN} \\ w_{TB} \\ w_{TN} \end{array} \begin{array}{c} \begin{array}{cccc} w_{HB} & w_{HN} & w_{TB} & w_{TN} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

(a) $would_B$

$$\begin{array}{c} \\ w_{HB} \\ w_{HN} \\ w_{TB} \\ w_{TN} \end{array} \begin{array}{c} \begin{array}{cccc} w_{HB} & w_{HN} & w_{TB} & w_{TN} \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

(b) $would_N$

FIGURE 6: If we impose **Causal Independence** on the matrices $would_B$ and $would_N$ from figures 4a and 4b, we get the matrices above.

would not change its truth-value, were you to choose any of your options. That is, if whether ϕ is causally independent of your choice, then for any option A,

$$would_A(w)(\phi) = \begin{cases} 1 & \text{if } \phi \text{ is true at } w \\ 0 & \text{if } \phi \text{ is false at } w \end{cases}$$

In the case of FOREKNOWN LOSS, whether you buy the ticket is causally independent of whether the coin lands heads or tails, so **Causal Independence** tells us that the way the coin lands wouldn't change, were you to buy the ticket. That is, it tells us that the matrices $would_B$ and $would_N$ are as shown in figure 6.

With *these* assumptions about what would happen, were you to buy or not buy the ticket, CDT will advise you against buying the ticket, no matter how confident you are that you will buy it. Again, not buying the ticket, N, would bring you \$0 at every world. But now we will say what is intuitively true: buying the ticket, B, would bring you \$100 at the worlds where the coin lands heads, and would lose you \$50 at worlds where the coin lands tails. That is:

$$would_B \cdot \mathcal{D} = [\mathcal{D}_B(w_{HB}), \mathcal{D}_B(w_{HN}), \mathcal{D}_B(w_{TB}), \mathcal{D}_B(w_{TN})]' = [100, 100, -50, -50]'$$

If you lack foreknowledge, then you'll think that the coin is just as likely to land heads as tails, the utility of buying the ticket will be an expected \$25, and CDT will advise you to buy the ticket. If, however, you have foreknowledge that the coin will land tails, then the utility of buying the ticket will be a certain loss of \$50, and CDT will correctly tell you to not buy.

In sum: decisions like FOREKNOWN LOSS pose a *prima facie* problem for CDT. But we need not respond to this problem by insisting that the oracle's prophesy makes a difference to the objective chances, as SPENCER does. We could instead see the problem as an instance of a well-known problem with our theories of subjunctive supposition. Then, the lesson of FOREKNOWN LOSS will be the same as the lesson of the Morgenbesser conditional. They both teach us to distinguish the probability that ϕ would result, were you to choose A, from the chance of ϕ, conditional on your choosing A.

Lesson #2 The probability that ϕ would result, were you to choose A, is not always just the chance of ϕ, conditional on your choosing A. If you choose A and ϕ is true,

then ϕ would be true, were you to choose A. And, if ϕ is causally independent of your choice, then ϕ would not change its truth-value, were you to choose differently.

6 | FOREKNOWLEDGE AND PREDICTION SENSITIVITY

As we briefly saw in §5 above, CDT's advice is sometimes sensitive to your predictions about your own choices. That is: the choices which CDT says are rational can depend upon which choice you think you will make. In my view, this *is* a problem with CDT. I will not attempt to persuade you to accept this consequence of CDT. I will try to persuade you that it is not a problem which is unique to decisions made with foreknowledge. And I will try to persuade you that there is something deeply right about CDT's advice in these cases, whether they involve foreknowledge or not.

A classic case of prediction sensitivity comes from Gibbard & Harper (1978):

> DEATH IN DAMASCUS
> You must choose whether to go to Aleppo or Damascus. And you know that Death has an appointment with you in one of these cities. Death does not watch over you, so your decision about where to go does not affect where Death awaits. But Death has made a prediction about which city you will choose, and he awaits in the predicted city. You take Death's predictions to be incredibly reliable.

In this decision, there are four relevant possibilities. Either Death awaits in Aleppo, α, or Death awaits in Damascus, δ. And either you go to Aleppo, A, or you go to Damascus, D. Let $w_{\alpha A}$ be a possibility at which Death awaits in Aleppo and you go to Aleppo. Let $w_{\alpha D}$ be a possibility at which Death awaits in Aleppo and you go to Damascus. And likewise for $w_{\delta A}$ and $w_{\delta D}$. Then, we can suppose that you prefer avoiding Death to meeting Death, and otherwise, you do not care which city you visit, so that $\mathcal{D} = [\mathcal{D}(w_{\alpha A}), \mathcal{D}(w_{\alpha D}), \mathcal{D}(w_{\delta A}), \mathcal{D}(w_{\delta D})]' = [0, 1, 1, 0]'$.[25]

By stipulation, whether you go to Aleppo or Damascus makes no difference with respect to where Death awaits. So, if Death is in Aleppo, then if you were to go to Damascus, Death would still be in Aleppo. And, if Death is in Damascus, then if you were to go to Aleppo, then Death would still be in Damascus. More generally, $would_A$ and $would_D$ are as shown in figure 7. Then, the instrumental value of going to Aleppo will depend upon whether Death awaits in Aleppo or Damascus. If Death is in Aleppo, then going to Aleppo would bring about your death. Whereas, if Death is in Damascus, then going to Aleppo would save your life.

$$would_A \cdot \mathcal{D} = [\mathcal{D}_A(w_{\alpha A}), \mathcal{D}_A(w_{\alpha D}), \mathcal{D}_A(w_{\delta A}), \mathcal{D}_A(w_{\delta D})]' = [0, 0, 1, 1]'$$

25. I've arbitrarily chosen to make 1 the desirability of living and 0 the desirability of dying. Since desirability is measured on an interval scale, any other pair of numbers where the first is higher than the second would be an equivalent representation of your desirabilities. But the choice of 1 and 0 makes the math a bit simpler.

$$\begin{array}{c} \\ w_{\alpha A} \\ w_{\alpha D} \\ w_{\delta A} \\ w_{\delta D} \end{array} \begin{array}{c} \begin{array}{cccc} w_{\alpha A} & w_{\alpha D} & w_{\delta A} & w_{\delta D} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

(a) $would_A$

$$\begin{array}{c} \\ w_{\alpha A} \\ w_{\alpha D} \\ w_{\delta A} \\ w_{\delta D} \end{array} \begin{array}{c} \begin{array}{cccc} w_{\alpha A} & w_{\alpha D} & w_{\delta A} & w_{\delta D} \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

(b) $would_D$

FIGURE 7: In figure 7a, the matrix $would_A(row)(column)$, which describes what would happen at each world, were you to go to Aleppo. In figure 7b, the matrix $would_D(row)(column)$, which describes what would happen at each world, were you to go to Damascus.

Likewise, the instrumental value of going to Damascus will depend upon where Death awaits. If Death is in Aleppo, then going to Damascus would save your life. Whereas, if Death is in Damascus, then going to Damascus would bring about your death.

$$would_D \cdot \mathcal{D} = [\mathcal{D}_D(w_{\alpha A}), \mathcal{D}_D(w_{\alpha D}), \mathcal{D}_D(w_{\delta A}), \mathcal{D}_D(w_{\delta D})]' = [1, 1, 0, 0]'$$

Death's predictions are very reliable, so we might as well suppose that they are perfect (it will simplify the math without making any substantive difference to our treatment of the case). So you are certain that you are either at $w_{\alpha A}$ or $w_{\delta D}$. Therefore, your credence that you are at $w_{\alpha A}$ is just your credence that you'll go to Aleppo, and your credence that you are at $w_{\delta D}$ is just your credence that you'll go to Damascus. Therefore, the utilities of going to Aleppo and Damascus are

$$\begin{aligned} & \mathcal{U}(A) = C(D) \\ \text{and} \quad & \mathcal{U}(D) = C(A) \end{aligned}$$

As your credence that you will go to Aleppo rises, so too does the utility of going to Damascus. And as your credence that you will go to Damascus rises, so too does the utility of going to Aleppo. If you are less than 50% confident that you'll go to Aleppo, then CDT says that going to Aleppo is rationally required. On the other hand, if you are more than 50% confident that you'll go to Aleppo, then CDT says that going to Aleppo is impermissible, and that you should instead go to Damascus.

In this case, CDT's prediction-sensitivity means that, if you both incline towards following its advice and learn that you are so inclined, then CDT will reverse that advice. Suppose you initially thought that you were going to go to Damascus. Then, CDT tells you that you must go to Aleppo. If you listen to CDT and learn that you're listening to it, then you'll get evidence that you'll likely go to Damascus. At that point, your rational credence that you'll go to Aleppo, $C(A)$, will fall below 50%, and CDT will reverse course and tell you to go to Aleppo. If you listen to it again, then it will change its mind again. In this decision, there is no option which CDT will continue to endorse after you've learnt that you intend to choose it. That is, whichever option you actually choose, CDT will say that this choice was irrational, and that you were rationally required to choose the other. Just to have a name for this kind of phenomenon,

we can say that, in DEATH IN DAMASCUS, CDT *forbids your prediction*.

By forbidding your prediction, CDT violates the following schematic principle: if A is a rational option, then, if you choose A, you will choose a rational option. For suppose that you're confident that you'll choose Aleppo. Then, CDT says that going to Damascus is a rational option. But if you then choose Damascus, you'll give yourself evidence that Death awaits in Damascus, and going to Damascus will no longer be a rational option.

For another, importantly different, case of prediction-sensitivity, consider CAKE IN DAMASCUS.

> CAKE IN DAMASCUS
>
> You must choose whether to go to Aleppo or Damascus. And you know that your fairy godmother has left cake for you in one of these cities. Your fairy godmother does not watch over you, so your decision about where to go does not affect where the cake awaits. But she has made a prediction about which city you will choose, and she left the cake in the predicted city. Your fairy godmother's predictions are incredibly reliable.

CDT's advice in CAKE IN DAMASCUS again depends upon how likely you are to go to Aleppo or Damascus. The decision is structurally exactly the same as DEATH IN DAMASCUS, except that, in this decision, you *want* your choice to match the prediction. If we now use 'α' and 'δ' for your fairy godmother leaving you cake in Aleppo and Damascus, respectively, then $would_A$ and $would_D$ are still as they are shown in figure 7. If we then suppose that $\mathcal{D} = [\mathcal{D}(w_{\alpha A}), \mathcal{D}(w_{\alpha D}), \mathcal{D}(w_{\delta A}), \mathcal{D}(w_{\delta D})]' = [1, 0, 0, 1]'$, while C is still $[C(A), 0, 0, C(D)]$, then

$$\begin{aligned} \mathcal{U}(A) &= C(A) \\ \text{and} \quad \mathcal{U}(D) &= C(D) \end{aligned}$$

So, as your credence that you will go to Aleppo rises, so too does the utility of going to Aleppo. And as your credence that you will go to Damascus rises, so too does the utility of going to Damascus. According to CDT, whichever city you think you're most likely to go to is the city that you're rationally required to go to. Suppose that you start out thinking that you're just as likely to go to Aleppo as you are to go to Damascus. Then, CDT will tell you that both of your options are rational. However, if you start to incline towards going to Aleppo, and you learn that you have, then CDT will reverse its verdict, telling you that, in fact, going to Damascus, which it previously called rational, is in fact an irrational choice. In this decision, there is no option which CDT will continue to endorse after you've learnt that you don't intend to choose it. That is, CDT will say that you are rationally forbidden from choosing whichever option you don't actually choose. Just to have a name for this kind of phenomenon, say that, in CAKE IN DAMASCUS, CDT *demands your prediction*.

By demanding your prediction, CDT violates the following schematic principle: if A is an irrational option, then, if you choose A, you will choose an irrational option.

For suppose that you're confident that you'll choose Aleppo. Then CDT says that going to Damascus is an irrational option. If you then choose to go to Damascus, you'll give yourself evidence that cake awaits in Damascus, and going to Damascus will no longer be an irrational option.[26]

I think that both of these kinds of prediction-sensitivity are a problem. In my view, rational permission is not the kind of thing which is retracted simply because the permission is exercised. Likewise, rational prohibition is not the kind of thing which is retracted simply because the prohibition is violated. Elsewhere, I've suggested a revision of CDT which deals with cases like this in a prediction-insensitive way.[27] Nonetheless, I think that there is *something* deeply right about the way that CDT treats these cases. What's deeply right about CDT's treatment of CAKE IN DAMASCUS is that, no matter which city you end up selecting, you *should* believe that your choice of destination has more instrumental value than the alternative. After all, if you find yourself going to Aleppo (e.g.), then you should think that your choice is taking you to the cake, and that the alternative would lead you away from cake. And surely a choice which leads you towards your desired ends has more instrumental value than a choice which leads you *away* from those desired ends. Likewise, what is deeply right about CDT's treatment of DEATH IN DAMASCUS is that, no matter which city you find yourself travelling towards, you *should* believe that the alternative has more instrumental value than your choice. After all, if you find yourself going to Aleppo (e.g.), then you should think that your choice is *killing you*, and that the alternative *would save your life*. Surely a choice which saves your life has more instrumental value than a choice which kills you!

In my view, then, decisions like these teach us three lessons. Firstly, it can be rational to choose an option even when you're very confident that it has less instrumental value than an alternative. To put the point more pithily: it can be rational to choose what you know to be the worst. Secondly, it can be irrational to choose an option, even when you are very confident that it has more instrumental value than an alternative. With more pith: it can be irrational to choose what you know to be the best. These are lessons which *I* draw from these cases, but I won't insist upon them here.

However, there is a third lesson which these cases teach us, and which I *do* want to insist upon. This is the dual of **Lesson #1**. **Lesson #1** taught us that, if whether ϕ is not under your control, but your rational *credence* that ϕ *is* under your control, then you may be subject to an agential illusion of control, and your intuitions about rational choice may be led astray. In decisions like DEATH IN DAMASCUS and CAKE IN

26. *Deliberational* causal decision theorists like SKYRMS (1990), ARNTZENIUS (2008), and JOYCE (2012, 2018) tell you to pursue actions in proportion to their utility until your predictions about what you will do are in an *equilibrium*. An equilibrium is a deliberational perspective from which every live option (every option you give positive credence to selecting) has an equal utility. Once you're in an equilibrium, JOYCE advises you to select between the live options in a way which is not sensitive to differences in utility. Insofar as deliberational CDT does not say anything about the rationality of choice outside of equilibrium, it will not forbid your prediction. However, it will sometimes demand your prediction; it will therefore violate this schematic principle along with orthodox CDT.

27. See GALLOW (2020) and GALLOW (forthcominga). See also BARNETT (ms) and PODGORSKI (forthcoming) for similar ideas.

DAMASCUS, the reverse is true: your rational credence that you'll get cake or death, respectively, is *not* under your control. No matter what you predict about what you'll do, you will be certain that cake or death awaits. However, *whether* you get cake or death *is* under your control. Suppose that you actually go to Aleppo and are greeted with cake or death. Then, it was in your control to go to Damascus. And, had you gone to Damascus, you wouldn't have found cake or death. In these kinds of decisions, causalists should recognise that a *lack* of control over your rational credence that ϕ can lead to the agential illusion that you have *no* control over whether ϕ. And in cases like these, too, we should be wary of our knee-jerk intuitions about rational choice.

Again, this agential illusion can vanish when you think about things from a better-informed, third-personal perspective. To illustrate, we can imagine that your friend, rather than you, is making the decision about whether to go to Aleppo or Damascus. And we can imagine that, while your friend does not know where Death is, you do. As you watch your friend deliberate about where to go, it will appear that there is *much more* instrumental value in the destination which takes them away from Death. From this point-of-view, there's no inclination towards the fatalistic verdict "just pick whichever city—it doesn't matter".

Lesson #3 When you have no control over your rational credence that ϕ, but you know for sure that you *do* have control over whether ϕ, your intuitive judgements about instrumental value can lead you astray by conflating a lack of control over your *epistemic state* with a lack of control over *the world*. In these cases, you should consider what instrumental value a choice has when viewed from a better-informed, third-personal perspective.

With this lesson appreciated, consider

CHOOSING THE CHANCES

There are two coins in front of you: a black one and a white one. You must choose which coin to flip. The black coin has a 2/3rds bias towards heads, and the white coin has a 2/3rds bias towards tails. If you flip the black coin, then you are betting on the outcome of the flip. If the black coin lands heads, then you will get $90; whereas, if the black coin lands tails, you will lose $90. Before you make your choice, the oracle informs you that the coin you flip will land on tails.

You will either flip the black coin, B, or the white one, W, and the flip will either land heads, H, or tails, T. So there are four relevant possibilities, w_{HB}, w_{HW}, w_{TB}, and w_{TW}, with the natural interpretation. Having learnt **Lesson #2**, I'm going to take **Strong Centering** for granted,[28] so I'm going to suppose that the imaging functions $would_B$ and $would_W$ are as shown in figure 8. I'll suppose that your desires are linear in dollars, so that $\mathcal{D} = [\mathcal{D}(w_{HB}), \mathcal{D}(w_{HW}), \mathcal{D}(w_{TB}), \mathcal{D}(w_{TW})]' = [90, 0, -90, 0]'$.

28. As the reader may verify for themselves, if we don't impose **Strong Centering**, then CDT will require you to flip the black coin, and this requirement won't be prediction-sensitive.

$$\begin{array}{c} \\ w_{HB} \\ w_{HW} \\ w_{TB} \\ w_{TW} \end{array} \begin{array}{c} \begin{array}{cccc} w_{HB} & w_{HW} & w_{TB} & w_{TW} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1 & 0 \\ 2/3 & 0 & 1/3 & 0 \end{bmatrix} \end{array}$$

(a) $would_B$

$$\begin{array}{c} \\ w_{HB} \\ w_{HW} \\ w_{TB} \\ w_{TW} \end{array} \begin{array}{c} \begin{array}{cccc} w_{HB} & w_{HW} & w_{TB} & w_{TW} \end{array} \\ \begin{bmatrix} 0 & 1/3 & 0 & 2/3 \\ 0 & 1 & 0 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

(b) $would_W$

FIGURE 8: In figure 8a, the matrix $would_B(row)(column)$, which describes what would happen at each world, were you to flip the black coin. In figure 8b, the matrix $would_W(row)(column)$, which describes what would happen at each world, were you to flip the white coin.

Then, the instrumental value of flipping the black coin at a will be $90, if you actually flip black and it lands heads, -$90 if you actually flip black and it lands tails, and $30 if you actually flip white. On the other hand, the instrumental value of flipping white is a constant $0.

$$would_B \cdot \mathcal{D} = [90, 30, -90, 30]'$$
$$\text{and} \quad would_W \cdot \mathcal{D} = [0, 0, 0, 0]'$$

Your foreknowledge tells you that the coin lands tails, so your credence that you're at world w_{TB} is just your credence that you flip the black coin. And your credence that you're at w_{TW} is just your credence that you flip the white coin. Therefore, the *utility* of flipping black is $30 minus $120 times your credence that you'll take black.

$$\mathcal{U}(B) = 30 - 120 \cdot C(B)$$

And the utility of flipping white is just a guaranteed $0, $\mathcal{U}(W) = 0$.

This means that, if your credence that you'll choose black is anywhere above 25%, then black will have a lower utility than white. That is: so long as you're more than 1/4th sure that you'll flip the black coin, you'll think that black has lower instrumental value than white, and CDT will advise you to flip white. *However*, if your credence that you will flip black drops below 25%, then the utility of flipping black will rise above the utility of flipping white, and CDT will change its mind, advising you to flip black instead. So, in CHOOSING THE CHANCES, CDT's advice is prediction-sensitive, and it forbids your prediction.

There is a persistent inclination to be fatalistic about this decision and insist: *Of course you shouldn't take the bet—the coin's going to land tails no matter what you do!* If we've learnt **Lesson #3**, we must guard ourselves against this inclination. For, even though this is a decision in which you know that the coin will land tails, it is not a decision in which you know that the coin will land tails *no matter what you do*. While your rational credence that the coin lands tails is not under your control, *whether* the coin lands tails is under your control. **Lesson #3** teaches us that our knee-jerk judgements of instrumental value can be led astray in precisely these kinds of decisions. So let us consider the matter from a better informed, third-personal perspective, imagining

that it is your friend making this decision, and not you. There are two better informed perspectives to consider, depending upon whether your friend flips the white or the black coin.

Suppose first that they flip the black coin. Then, the oracle's prophesy was about the black coin, so the black coin lands tails, and your friend's choice cost them $90. Losing $90 is a bad outcome. There's more instrumental value in a guaranteed $0 than there is in a $90 loss. So, in the possibility, your friend has chosen the option with the least instrumental value. Suppose, on the other hand, that your friend flips the white coin. Then, the oracle's prophesy was about the white coin, and not the black. That is, the reason why the oracle told your friend that the coin landed tails is that the *white* coin lands tails—the coin which is, after all, biased towards tails. So, if your friend flips the white coin, then the oracle's prophesy told them *nothing at all* about how the *black* coin would have landed, were they to flip it. And, were they to have flipped black instead, they would have had a 2/3rds chance of getting $90, and a 1/3rd chance of losing $90. On average, flipping the black coin with these odds would get them $30. That's a good bet. So your friend has turned down a good bet with an instrumental value of $30 in exchange for a guaranteed $0. So, in this possibility too, your friend has chosen the option with the least instrumental value.

These are not the only possibilities, of course. The oracle could be wrong. However, these are the only two possibilities that your friend is taking seriously in their deliberation. And in both of them, their choice has a lower instrumental value than the alternative; that is, in both of them, they choose the worst option. CDT is absolutely correct about this, whether or not it's correct to forbid your prediction.

In other decisions with foreknowledge, CDT will demand your prediction. Consider:[29]

> PAUPER'S PROBLEM
>
> You are a pauper. Tomorrow, the lords will send you into battle. You do not have any armour, but you could spend your life's savings to purchase some. The chance of surviving battle without armour is 10%. The chance of surviving with armour is 90%. Before you decide whether to purchase the armour, the oracle informs you that you will survive.

In this decision, either you will *s*urvive, S, or you will *d*ie, D. And either you will *b*uy the armour, B, or you will *n*ot, N. So there are four relevant possibilities: w_{SB}, w_{SN}, w_{DB}, and w_{DN}, with the natural interpretation. I'll suppose that your desire to survive is ten times stronger than your desire to not lose your life savings, so that the desirability of losing your life, but not your life savings, can be represented with 10, the desir-

29. This decision (or a close variant of it) is discussed in Lewis (1986), Rabinowicz (2009), Price (2012), and Stern (forthcoming). Bales (2016) argues that the case does not pose a problem for CDT by showing that different specifications of the imaging functions can secure whatever verdict you like. However, the two versions of CDT which yield the decisive 'buy the armour' and the decisive 'don't buy the armour' verdicts have both failed to learn **Lesson #2** from §5, so they will say that you should buy the ticket in FOREKNOWN LOSS. Any version of CDT which has learnt **Lesson #2** will demand your prediction in PAUPER'S PROBLEM.

	w_{SB}	w_{SN}	w_{DB}	w_{DN}
w_{SB}	1	0	0	0
w_{SN}	9/10	0	1/10	0
w_{DB}	0	0	1	0
w_{DN}	9/10	0	1/10	0

(a) $would_B$

	w_{SB}	w_{SN}	w_{DB}	w_{DN}
w_{SB}	0	1/10	0	9/10
w_{SN}	0	1	0	0
w_{DB}	0	1/10	0	9/10
w_{DN}	0	0	0	1

(b) $would_N$

FIGURE 9: In figure 9a, the matrix $would_B(row)(column)$, which describes what would happen at each world, were you to buy the armour. In figure 9b, the matrix $would_N(row)(column)$, which describes what would happen at each world, were you to not buy the armour.

ability of losing your life savings but nor your life can be represented with 100, and likewise, $\mathcal{D} = [\mathcal{D}(w_{SB}), \mathcal{D}(w_{SN}), \mathcal{D}(w_{DB}), \mathcal{D}(w_{DN})]' = [100, 110, 0, 10]'$. Because I have learnt **Lesson #2**, I will assume that $would_B$ and $would_N$ both satisfy **Strong Centering**, and therefore are as shown in figure 9. Finally, because the oracle has informed you that you will survive, your credence that you are in w_{SB} is just your credence that you'll buy the armour, and your credence that you are in w_{SN} is just your credence that you will not. It then follows that

$$\mathcal{U}(B) = 90 + 10 \cdot C(B)$$
$$\text{and} \quad \mathcal{U}(N) = 110 - 90 \cdot C(B)$$

If your credence that you will buy the armour is greater than 1/5th, then the utility of buying will exceed the utility of refraining. And, if your credence that you will buy the armour is less than 1/5th, then the utility of not buying will exceed the utility of buying. So, in this case, CDT demands your prediction.

I won't defend demanding your prediction. On my view, you are both permitted and required to buy the armour, no matter how likely you think you are to buy it. But I nonetheless think that we should accept everything CDT has to say about instrumental value in PAUPER'S PROBLEM. In this decision, too, there is an inclination to be fatalistic: *You shouldn't buy the armour—you're going to survive no matter whether you buy it or not!* Again, **Lesson #3** warns us to resist this fatalistic impulse. While you have no control over your rational credence that you survive, you *do* have control over *whether* you survive. Let us consider the matter from a better informed, third-personal perspective, and imagine that it is your friend making this decision, and not you. Again, there are two better informed perspectives to consider, depending upon whether your friend ends up buying the armour or not.

Suppose first that they refrain from buying the armour. Then, so long as the oracle's prophesy is accurate, they *do* survive, even without the armour, and their choice has kept them from losing their life savings. In this possibility, purchasing the armour would have accomplished nothing other than leaving them penniless and exposing them to a 10% risk of losing their life. Sparing your life savings is more instrumentally valuable than wasting it on armour that you don't need, and which exposes you to a 10% chance of death. So, in this possibility, your friend has chosen the option with

the most instrumental value. On the other hand, suppose that your friend buys the armour. In this possibility, they survive (per the oracle's prophesy), and moreover, the decision to purchase the armour very likely saved their life. Had they not purchased the armour, they would have exposed themselves to a 90% chance of death. Since they value their life 10 times more than their life savings, keeping their life savings is not worth a 90% chance of death. So, in this possibility, too, your friend has chosen the option with the most instrumental value.

These are not the only possibilities. The oracle's prophesy could be false. But these are the only possibilities which your friend takes seriously in their deliberation. And in both of them, they choose the option with the highest instrumental value. CDT is right about this, whether or not it's right to demand your prediction.

7 | FOREKNOWN IRRATIONALITY

Distinguish two kinds of decision under foreknowledge. You could have foreknowledge which does not, even conjoined with your other background information, tell you anything non-trivial about what you will in fact choose. Call decisions like this *exotic*. Alternatively, you could have foreknowledge which, conjoined with your other background information, tells you something non-trivial about what you will in fact choose. Call decisions like this *wild*.

For an example of a wild decision: you may either choose a guaranteed $1 or a guaranteed $100. Before you decide, the oracle tells you that you will actually take the $100. What should you do? Clearly, you should take the $100. Taking the $100 was the rational choice before receiving the foreknowledge, the foreknowledge doesn't appear to change your values or your views about what would result from each possible choice, so taking the $100 should still be the rational choice afterwards.

For another wild decision: suppose again that you may choose either a guaranteed $1 or a guaranteed $100. Before you decide, the oracle tells you that you will actually take the $1. Then, what should you do? Again, you should take the $100. If the oracle speaks the truth, then you won't; but it's not clear why the oracle's prophesy should change what it is *rational* for you to do. In general, it's possible for you to choose irrationally while knowing that you are doing so. In mundane cases of *akrasia*, you choose irrationally while knowing that you are doing so. Why should knowingly choosing irrationally be more problematic, just because the knowledge of your irrationality is *fore*knowledge?

In cases of *akrasia*, even if you know that you will choose irrationally, you wouldn't retain that knowledge, were you to choose differently. Because your own deliberation is the source of your knowledge that you will choose irrationally, the knowledge depends upon that very deliberation. Were you to deliberate differently, you would lose your knowledge that you are choosing irrationality. And the same is true in wild decisions. In fact, you take the $1, and you knew that you were going to, on the basis of the oracle's testimony. However, were you to have chosen the $100, you would not have known that you were going to take the $1. Either the oracle wouldn't have prophesied that you were

going to take the $1, or else this prophesy would not have been *known*.

Importantly, in wild decisions, what is known and what is merely believed may depend upon your choice. Suppose that you actually take the $1. Then, you are in a 'good case'—at least, the case is epistemically good. The oracle speaks from knowledge of your choice, and you form the belief that you will take the $1 on the basis of her known testimony. In these circumstances, you may come to know that you'll take the $1. If you had taken the $100, then either the oracle wouldn't have prophesied that you'd take the $1 (so that, at the moment of choice, you misremember her prophesy) or else she would have *falsely* prophesied that you'd take the $1. Either way, you would be in a 'bad case'—at least, the case would be epistemically bad. You would not be in a position to know that you take the $1 (because you don't). So, even though you are actually in a good case, and you know that you will take the $1, had you taken the $100, you would have been in a bad case, and you would not have known this.[30]

In actuality, you both take the $1 and *know* that you will take the $1. I say that you chose irrationally. What *should* you have done? Well, to start, you should have opened deliberation about what to do. At that point, you should have begun to take seriously the possibility that you take the $100, which is a possibility in which you don't know that you'll take the $1, and in which the oracle's prognostication is false or misremembered. Since $100 is better than $1, you should have inclined towards taking the $100. These inclinations give you evidence that you actually will take the $100, so through the course of deliberation, you should have become more confident that you will take the $100. And that means becoming less confident that the oracle speaks from knowledge, and so becoming less confident that you know that you are going to take the $1. You should have then taken the $100, and been confident that either the oracle prophesied falsely or that you misremembered her prophesy. This is my advice, and I give it to you in both the 'good case' and in the 'bad cases'. In the 'bad cases', you follow my advice and the oracle's prophesy is not known. In the 'good case', you don't follow my advice, and you know that you won't follow it.

There's a general lesson to be learnt here:

Lesson #4 In decisions made with foreknowledge, your own rational deliberation can provide you with evidence that the oracle's prophesy is false, misleading, or misremembered. So you shouldn't always take your foreknowledge for granted when deliberating about what to do.

In the previous decisions we've considered, your deliberation about what to do did not furnish you with evidence that the oracle's prophesy was false, misleading, or misremembered. So, in those decisions, there was no harm in taking the prophesy for granted throughout deliberation. However, in wild decisions, we must exercise more caution.

STERN (forthcoming) considers the following decision:

30. Thanks to Melissa Fusco for helpful conversation on this point.

FUTURE MEDICAL TEST

Smoking causes lung cancer by causing your lungs to blacken. The effects of smoking on lung cancer are entirely mediated by its effects on whether your lungs blacken. You would enjoy smoking, but you would hate to contract lung cancer. Before you decide whether to smoke, the oracle tells you about the results of a future medical test: your lungs will blacken.

To fill out the decision, we may suppose that you're very confident that nothing besides smoking causes lungs to blacken, so that you are very confident that you smoke, conditional on your lungs blackening. Then, the oracle's prophesy has provided you with foreknowledge which, conjoined with your other background information, tells you something non-trivial about what you will in fact choose: it tells you that you are quite likely to choose to smoke. So this decision is wild.

In my view, FUTURE MEDICAL TEST is not importantly different from the decision in which the oracle tells you that you will choose the $1 over the $100. If she speaks from knowledge, then you will smoke, blacken your lungs, and thereby, quite likely, give yourself lung cancer. This is an irrational choice. If you do this, then you've chosen to expose yourself to a significant chance of death. The fact that this irrational choice was foretold does nothing to alter that fact. STERN disagrees. He writes: "it seems clear (at least to this author) that you should go ahead and smoke. After all, you already know that your lungs will blacken no matter what you do. Why not savor the pleasures of the cigarette?" If we've appreciated **Lesson #3**, then we should be cautious about this kind of fatalistic reasoning. If you follow STERN's advice and smoke, then you do know that your lungs will blacken. But you emphatically do not know that your lungs will blacken *no matter what you do*. Let us consider the matter from a better informed, third-personal perspective. Suppose that your friend is deciding whether to smoke. The oracle tells them that their lungs will blacken, they decide to smoke, their lungs blacken, and they die of lung cancer. From your point of view, it appears clear that your friend's choice *killed them*. Had they not smoked, they would have lived. Your friend had control over whether or not to die of lung cancer, and they chose to die. It seems clear (at least to this author) that a choice which kills you has less instrumental value than a choice would would have saved your life.

STERN assumes that your knowledge that your lungs blacken is not under your control: whether you choose to smoke or not, you will still know that your lungs blacken, on the basis of the oracle's prophesy. However, I think that **Lesson #4** should lead us to question the assumption. Given that there are no other likely causes of black lung, learning that you have refrained from smoking gives you evidence that the oracle's prophesy is false, misleading, or misremembered. So it seems to me that you shouldn't take the oracle's prophesy for granted in your deliberation about whether to smoke.

STERN considers another version of this decision, in which there is another potential cause of black lung: a rare genetic condition which causes black lung, whether or not you smoke. In that version of the decision, given that you find yourself not smoking, you should think it most likely that you have the genetic condition, and that

smoking wouldn't do any harm. On the other hand, given that you find yourself smoking, you should think it's most likely that you don't have the genetic condition, and your choice to smoke is likely killing you. This decision is one in which CDT forbids your prediction. In my opinion, it's not importantly different from CHOOSING THE CHANCES. We should heed **Lesson #3** and guard against the fatalistic reasoning which conflates a lack of control over your rational credence that your lungs blacken with a lack of control over *whether* your lungs blacken.

STERN disagrees, and proposes a modification of CDT which will advise you to smoke in this decision.[31] To appreciate STERN's proposal, let's use 'C_0' for your ur-prior credence function, and suppose that your total evidence consists of the ordinary evidence E and the foreknowledge F. Then, as we saw in §4, CDT evaluates acts for choiceworthiness with

$$\mathcal{U}(A) \;=\; (C_0 \mid EF) \cdot \mathit{would}_A \cdot \mathcal{D}$$

That is: CDT tells you to take all of your evidence into account by updating your ur-prior credences on it, and then use the resulting probability function, $C_0 \mid EF$, to take an expectation of how desirable things would be, were you to choose A. STERN thinks that, instead, you should evaluate acts for choiceworthiness with

$$\mathcal{S}(A) \;=\; (C_0 \mid E) \cdot \mathit{would}_{AF} \cdot \mathcal{D}$$

That is: STERN tells you to take your *ordinary* evidence into account by updating your ur-prior credences on it. Then, you should use the resulting probability function, $C_0 \mid E$, to take an expectation of how desirable things would be, were you to choose A *while your foreknowledge is held fixed*. For instance, when you think about what would happen, were you to not smoke, STERN tells you to hold fixed that your lungs blacken.

This theory rejects **Lesson #4**, since it requires you to take your foreknowledge for granted when deliberating about what to do. Consider again the wild decision in which you are given a choice between a guaranteed \$1 and a guaranteed \$100, and the oracle informs you that you will take the \$100. In this decision, I think it is clear that it is irrational to take the \$1. But STERN has a hard time agreeing. The reason is that, even though the $\mathcal{S}$-value of taking the \$100 is well-defined, it's unclear how we should define the $\mathcal{S}$-value of taking the \$1. *Ex hypothesi*, you have only two available options: taking the \$1 and leaving the \$100 behind—call that option 'O', for *o*ne—and taking the \$100 and leaving the \$1 behind—call that option 'H', for *h*undred. Choosing both O and H is impossible. If you leave the \$100 behind, then you cannot take it, too. In this decision, you have foreknowledge that you take the hundred, H. So the $\mathcal{S}$-value of taking the \$1 is:

$$\mathcal{S}(O) \;=\; C_0 \cdot \mathit{would}_{OH} \cdot \mathcal{D}$$

31. As I'm using the term, STERN is not a causalist. He is instead, like HITCHCOCK (2016), an evidential decision theorist who accepts a libertarian metaphysics according to which free choices are the consequence of an uncaused Will. See footnote 18.

The issue is that it's unclear how we should think about $would_{OH}$. That is, it's unclear how we should think about what would happen, were you to—*per impossibile*—take only the $1 and take only the $100. And if we can't assign an $\mathcal{S}$-value to taking the $1, then we won't be able to say that it's rational to take the $100, nor that it's irrational to take the $1.

One suggestion which looks natural in STERN's favoured formalism is that you should make *nested* counterfactual suppositions. For instance, perhaps you should *first* imagine a possibility which is just like the actual world, except that your foreknowledge does not depend upon its causal past, and *then* imagine a possibility which is just like *that* world, except that you choose A in a way which doesn't depend upon your causal past.[32]

What this theory tells us depends upon how widely we construe 'foreknowledge'. One thing you know for sure is that you will be $100 richer in ten minutes iff you choose the $100. So, when the oracle provides you with the foreknowledge that you will take the $100, you are in a position to know that you'll be $100 richer in ten minutes. If this information counts as foreknowledge which is to be held fixed, then the $\mathcal{S}$-value of taking the $1 will equal the $\mathcal{S}$-value of taking the $100. For, in calculating the $\mathcal{S}$-value of taking the $1, we should first counterfactually suppose that you take the $100 *and* that you're $100 richer in ten minutes—and that neither of these facts depends upon their causal past. And then, we should further counterfactually suppose that you take the $1. This second supposition undoes some, but not all, of the first one. We're left with a possibility in which you take the $1, but still end up $100 richer in ten minutes. And this is just as desirable as a possibility in which you take the $100 and end up $100 richer in ten minutes. So, if we construe 'foreknowledge' broadly, then STERN's theory will tell you that it's rationally permissible to take a guaranteed $1 over a guaranteed $100. This looks like a bad permission to give, and the badness of the permission is not mitigated by the fact that you don't act on it.

We could try to construe 'foreknowledge' more narrowly, so that your foreknowledge only includes the information which the oracle explicitly provides, and not the information which you can readily deduce from her prophesy. Then, the theory would say—correctly, I think—that it is irrational to take the $1. For if foreknowledge is narrow, we should calculate the $\mathcal{S}$-value of taking the $1 by first counterfactually supposing that you take the $100, and then counterfactually supposing that you take the $1. The second supposition undoes the first, and we say that the $\mathcal{S}$-value of taking the $1 is $1. Similarly, we should calculate the $\mathcal{S}$-value of taking the $100 by first counterfactually supposing that you take the $100 and next supposing that you take the $100. The second supposition adds nothing to the first, and we say that the $\mathcal{S}$-value of taking the $100 is $100.

But consider the following variant of our wild decision: you are given a choice

32. STERN's favoured formalism utilises causal Bayes nets. In that formalism, the suggestion is that we *first* intervene so as to bring about your foreknowledge, and *then* intervene so as to bring it about that you choose A.

between $1 and $100. Before you make your choice, the oracle tells you that, in ten minutes, you'll be $100 richer. Now, to calculate the $\mathcal{S}$-value of taking the $1, we first counterfactually suppose that you are $100 richer in ten minutes (and that this fact does not depend upon its causal past) and then counterfactually suppose that you take the $1. This second supposition does not undo the first, so we say that, in this variant of the decision, the $\mathcal{S}$-value of taking $1 is the same as the $\mathcal{S}$-value of taking $100. So, again, the theory will give you permission to take the $1.

REFERENCES

Ahmed, Arif. 2014. "Dicing with Death." *Analysis*, vol. 74 (4): 587–592.

Albert, David Z. 2003. *Time and Chance*. Harvard University Press, Cambridge, MA. [3]

Arntzenius, Frank. 2008. "No regrets, or: Edith Piaf revamps decision theory." *Erkenntnis*, vol. 68: 277–297. [23]

Bales, Adam. 2016. "The pauper's problem: chance, foreknowledge and causal decision theory." *Philosophical Studies*, vol. 173: 1497–1516. [5], [26]

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

Berkovitz, Joseph. 2001. "On Chance in Causal Loops." *Mind*, vol. 110 (437): 1–23.

Cusbert, John. 2017. "Backwards Causation and the Chancy Past." *Mind*, vol. 127 (505): 1–33.

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

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

—. forthcomingb. "Updating for Externalists." *Noûs*. [12]

Gibbard, Allan & William L. Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, A. Hooker, J.J. Leach & E.F. McClennan, editors, 125–162. D. Reidel, Dordrecht. [20]

Hájek, Alan. ms. "Most Counterfactuals are False." Available at `http://philrsss.anu.edu.au/people-defaults/alanh/papers/MCF.pdf`. [3]

Hall, Ned. 1994. "Correcting the Guide to Objective Chance." *Mind*, vol. 103 (412): 505–517. [14], [15]

Hitchcock, Christopher. 2016. "Conditioning, Intervening, and Decision." *Synthese*, vol. 194 (4): 1157–1176. [1], [12], [13], [14], [31]

Horwich, Paul. 1987. *Asymmetries in Time*. The M.I.T. Press, Cambridge, MA. [3]

Jeffrey, Richard. 1965. *The Logic of Decision*. McGraw-Hill, New York. [12]

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

—. 2012. "Regret and instability in causal decision theory." *Synthese*, vol. 187 (1): 123–145. [23]

—. 2018. "Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems." In *Newcomb's Problem*, Arif Ahmed, editor. Oxford University Press, Oxford. [23]

Lewis, David K. 1973. *Counterfactuals*. Blackwell Publishers, Malden, MA. [3], [16], [17]

—. 1980. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability*, Richard C. Jeffrey, editor, vol. II, 263–293. University of California Press, Berkeley. [4]

—. 1981. "Causal Decision Theory." *Australasian Journal of Philosophy*, vol. 59 (1): 5–30. [1], [4], [5], [9], [17]

—. 1986. "Postscript to 'Causal Decision Theory'." In *Philosophical Papers, volume 2*. Oxford University Press, Oxford. [17], [26]

Maudlin, Tim. 2019. *Philosophy of Physics: Quantum Theory*. Princeton University Press, Princeton.

Meacham, Christopher J. G. 2010. "Two Mistakes Regarding the Principal Principle." *British Journal for the Philosophy of Science*, vol. 61 (2): 407–431. [14], [15]

Meek, Christopher & Clark Glymour. 1994. "Conditioning and Intervening." *The British Journal for the Philosophy of Science*, vol. 45: 1001–1021. [5]

Mellor, Hugh. 1998. *Real Time II*. Routledge, London.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel*, Nicholas Rescher, editor, 114–146. D. Reidel, Dordrecht.

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

Price, Huw. 2012. "Causation, Chance, and the Rational Significance of Supernatural Evidence." *Philosophical Review*, vol. 121 (4): 483–538. [1], [15], [26]

Rabinowicz, Wlodek. 2009. "Letters from Long Ago: On Causal Decision Theory and Centered Chances." In *Logic, Ethics, and All That Jazz—Essays in Honour of Jordan Howard Sobel*, L-G. Johansson, editor, vol. 56, 247–273. Uppsala Philosophical Studies. [5], [16], [17], [26]

Rabinowicz, Włodzimierz. 1982. "Two Causation Decision Theories: Lewis vs Sobel." In *Philosophical Essays Dedicated to Lennart Åqvist on His Fiftieth Birthday*, Tom Pauli, editor, vol. 34, 299–321. Uppsala Philosophical Studies, Uppsala. [4], [16], [17]

Roberts, John T. ms. "Must a Cause be Earlier than its Effect?" URL `https://philosophy.unc.edu/wp-content/uploads/sites/122/2013/10/BackCause-GargnanoConf.pdf`. [10], [15]

Schoenfield, Miriam. 2017. "Conditionalization does not (in general) Maximize Expected Accuracy." *Mind*, vol. 126 (504): 1155–1187. [12]

Seidenfeld, Teddy. 1984. "Comments on Causal Decision Theory." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 201–212.

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

—. 1990. *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge, ma. [23]

Slote, Michael A. 1978. "Time in Counterfactuals." *The Philosophical Review*, vol. 87 (1): 3–27. [18]

Sobel, Jordan Howard. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge University Press, Cambridge. [4], [5], [16], [17]

Spencer, Jack. 2020. "No Crystal Balls." *Noûs*, vol. 54 (1): 105–125. [1], [14], [15], [16], [19]

Stalnaker, Robert C. 1975. "Indicative Conditionals." *Philosophia*, vol. 5 (3): 269–286. [3]

Stern, Reuben. 2017. "Interventionist Decision Theory." *Synthese*, vol. 194: 4133–4153.

—. forthcoming. "An Interventionist's Guide to Exotic Choice." *Mind*. [1], [26], [29], [30], [31], [32]

Walters, Lee & J. Robert G. Williams. 2013. "An argument for Conjunction Conditionalization." *The Review of Symbolic Logic*, vol. 6 (4): 573–588. doi: 10.1017/S1755020313000191. [17]