Decision and Foreknowledge

J. DMITRI GALLOW †

Abstract: 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 problem for causal decision theory (CDT). My thesis is that foreknowledge poses no new problems for CDT. Some of the purported problems are not problems. Others are problems, but they are not problems for CDT. Rather, they are problems for our theories of subjunctive supposition. Others are problems, but they are not new problems. They are old problems transposed into a new key. Nonetheless, decisions made with foreknowledge illustrate important lessons about the instrumental value of our choices. Once we've appreciated these lessons, we are left with a version of CDT which faces no novel threats from foreknowledge.

1 Introduction

Say that you have *foreknowledge* when you know something about future events, and, moreover, this knowledge is caused by the future events it is about. My topic is how to make decisions when you have 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, p. 18) says that they are "much more problematic for decision theory than the Newcomb problems". Price (2012) thinks that these decisions should push causalists towards a subjectivist theory of causation. Hitchcock (2016) and Stern (2021) have proposed decision theories which disagree with orthodox causal decision theory only when you have 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.

Final Draft. Forthcoming at *Noûs*.

† Thanks to Melissa Fusco, Bryan Parkhurst, Martin Thomson-Jones, Reuben Stern, audiences at the Dianoia Institute of Philosophy and Oberlin College, and three anonymous reviewers.

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.

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, decisions made with foreknowledge vividly illustrate important lessons for causalists. These lessons should be familiar from less exotic situations, but they are needed to navigate the unfamiliar and confusing terrain faced by agents with foreknowledge. Not everyone will want to draw these lessons. But if we do, 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. 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. I'll call knowledge that ϕ *fore*knowledge iff ϕ is at least partly about the future, and your belief that ϕ is caused by events in the future which ϕ is about. Even if you think that foreknowledge like this is impossible, you needn't deny that it could be rational to take seriously the possibility that you have foreknowledge.

I'll say that whatever foreknowledge you have comes from *the oracle*. The oracle could be a time traveller, fortune teller, crystal ball, angel, demon, or prophetic dream. You take the oracle's testimony about the future to be like an ordinary human's testimony about the present or past. 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's important is just that, firstly, her prophesies are in general about events which have yet to unfold, events which lie in the future of the prophesy; secondly, her prophesies are, in general, caused by the events they are about; and, thirdly, in general, her prophesies are accurate.

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. But I'll assume that facts about what will happen are just as metaphysically fixed, determinate, and unchanging as facts about what has happened. If time branches or changes (over *hyper*time, perhaps), then it is less clear in what sense you could have foreknowledge of

your future.

3 Causal Decision Theory

When you face a decision, you choose from a collection of available acts, $\mathcal{A}$. And there is a collection of relevant ways things might be. Call the ways things might be 'worlds', and denote their collection with '$\mathcal{W}$'. For expository convenience, I'll suppose that both $\mathcal{A}$ and $\mathcal{W}$ are finite. 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 write '$\mathcal{D}(w)$'.

The theory I'm going to call 'causal decision theory' will also rely upon a family of functions measuring each act's causal powers at each possibility. For the act $\mathrm{A} \in \mathcal{A}$, I'll call this function '$would_A$'. You hand $would_A$ a world, w, and it hands you back a probability distribution, $would_A(w)$. The interpretation of this probability distribution is that $would_A(w)(w^*)$ gives A's causal tendency to bring about w^*, in the world w. For counterfactualists, we can think of it as the probability, at w, that w^* would result, were you to select A. Since A would certainly bring about a world in which 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 to 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, $\mathcal{D} = [\mathcal{D}(w_1), \mathcal{D}(w_2), \ldots, \mathcal{D}(w_N)]'$.[2] 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:

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

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

(Here, ' · ' is matrix multiplication.) That is: the utility of A consists of three ingredients: your credences, your desires, and information about what A would bring about. Multiply these ingredients together, and you get A's utility.

I'm going to use 'causal decision theory' as an umbrella term for any decision theory which says that you should choose an option which maximises utility, as defined above. Given this terminology, the theories of Stalnaker (1981), Gibbard & Harper (1978), Lewis (1981), Skyrms (1982), Sobel (1994), and Joyce (1999) will all count as versions of causal decision theory. All of these theorists advise you to maximise utility, as specified above. But that doesn't mean that they all agree. For 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.[3] Others, like Sobel (1994) and Rabinowicz (1982, 2009), will want to understand the imaging function differently.[4] In §5 below, I will argue that we should impose a constraint on the imaging function which none of these authors have explicitly discussed. 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.

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.[5] 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.[6] Then, CDT tells you to evaluate an act by taking an expectation of desirability, $\mathcal{D}$, with respect

3. 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 is admissible, and it entails the chance of w^*, given that you choose A, so Lewis (1980)'s *principal principle* will imply that $C(w^* \mid AK_w)$ should be $Ch_w(w^* \mid A)$.

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

5. CDT is presented in terms of imaging your credences on A in Lewis (1981), Sobel (1994), and Joyce (1999).

6. See Gärdenfors (1982).

to these 'subjunctive credences', C_A. This is to be contrasted with *evidential* decision theory, EDT, which tells you to evaluate an act by taking an expectation of desirability with respect to the your credences *conditioned* on A, $C|A$. That is, according to EDT, you should evaluate acts in terms of their *news-value*, $\mathcal{V}$, where

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

In my view, CDT's philosophical commitments are clearer if we instead group '$would_A$' with $\mathcal{D}$. 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 choosing A would bring about at world w_i.[7] Then, CDT says that the choiceworthiness of an act, A, is given by your expectation of this quantity. That is,

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

As this formulation makes clear, CDT follows from the view that a choice's instrumental value is given by the desirability of what it would bring about, together with the assumption that you should choose an act with the greatest expected instrumental value. On this view, the only thing that makes an act worth choosing is what it would do to bring about desirable ends.

For a decision which illustrates this idea, consider

> NO DIFFERENCE
> You may either take the box on the left, 'Lefty', or the box on the right, 'Righty'. There's no difference between them. Their contents were decided yesterday on the basis of a prediction about which box you would choose. If it was predicted that you would choose Lefty, then both boxes contain $100. If it was predicted that you would choose Righty, then both boxes contain $10. You are certain that the prediction is accurate.

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 are certain that the prediction is accurate, you are certain that

7. To be clear: '$\mathcal{D}_A(w_i)$' is the ith row in the vector $\mathcal{D}_A$.

	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.

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 tells you to evaluate each box by asking how much money you *would* get, were you to take it. You won't know for sure, since you won't know for sure what's inside the boxes, but you will 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, at w_{HL} and w_{HR}, had you taken Lefty, there'd still be \$100 in both boxes; and, had you taken Righty, there'd still be \$100 in both boxes. And at w_{TL} and w_{TR}, had you taken Lefty, there would still be \$10 in both boxes; and, had you taken Righty, there'd still be \$10 in both boxes.

So choosing L and choosing R would accomplish exactly the same thing. If there's \$100 in both boxes, both options would get you \$100. And if there is \$10 in both boxes, both options would get you \$10.

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

So taking Lefty and taking Righty have the same utility. Both have a utility equal to 100 times your credence that you'll take Lefty plus 10 times your credence that you'll take Righty. CDT therefore says that you have just as much reason to take Lefty as you have reason to take Righty.

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 should go 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 should go up. In this state of indecisive waffling, your rational opinions about how much money is in the boxes will change along with your opinions about how you'll choose. Since you have control over which box to take, you have control over your rational credence that there's $100 in the boxes. But, by stipulation, you don't have any control over whether there's $100 in the boxes.

It can seem like you have reason to take Lefty. But causalists think that this is an illusion. You succumb to the illusion by conflating control over *what to believe* with control over *what is the case*. They encourage you to break the illusion of control by considering the decision from a better informed, third-personal perspective. For instance, suppose that you are watching your friend decide between Lefty and Righty, and you can see inside the boxes. No matter what you see, there won't appear to be any more instrumental reason to take Lefty than there is to take Righty. Once you are free of the illusion of control, causalists think the inclination to say that you have reason to take Lefty vanishes.

Indeed, they think that, from this better-informed, third-personal perspective, a preference for Lefty over Righty appears counter-productive. Suppose your friend is an evidentialist, who values taking Lefty $90 more than taking Righty. Then, they would be willing to pay up to $90 to take Lefty. You look into the boxes and see $100 in both. From your perspective, it seems that paying $90 to take Lefty is just throwing money away to no end. Your friend could easily get the $100 for free by taking Righty. And nothing depends upon there being $100 in the boxes. Things are the same if you see $10 in the boxes instead.

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—by stipulation—the only thing that's under your control is your own epistemic state.

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 each of the possible better informed, third-personal perspectives.

Not everyone is going to learn this lesson from NO DIFFERENCE. But it is a lesson which

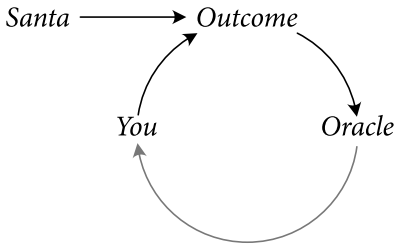

FIGURE 2: The causal structure of STICKER.

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 have a purely decorative sticker which doesn't make any difference with respect to who gets to open which gift. The oracle arrives with news from the future: the gift you'll put the 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. *Santa* says whether Santa put the toy in your gift or your sister's. *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. *You* says whether you put the sticker on your gift or your sister's. And *Outcome* tells us both which gift has the toy and which gift has the sticker.

Santa and *You* causally influence *Outcome*. And *Outcome* causally influences *Oracle*—since the oracle is likely to tell the truth about *Outcome*. It is possible that the oracle's pronouncement will influence your choice, whence the arrow from *Oracle* to *You* in figure 2. But you might instead make up your mind about what to do in a way that's insensitive to the oracle's prophesy, whence the arrow is grey, rather than black.

In this decision, there are four relevant 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}'.

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. Therefore, at the world w_{TY}, had you put the sticker on your sister's gift, your sister would have been gifted the toy. This is 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 would have to change, were you to choose differently. 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.

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 on your gift gives you evidence that you'll get the toy, and putting the sticker on your sister's gift

$$\begin{array}{c} \\ w_{TY} \\ w_{TS} \\ w_{CY} \\ w_{CS} \end{array} \begin{array}{c} \begin{array}{cccc} w_{TY} & w_{TS} & w_{CY} & w_{CS} \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right] \end{array}$$

(a) $would_Y$

$$\begin{array}{c} \\ w_{TY} \\ w_{TS} \\ w_{CY} \\ w_{CS} \end{array} \begin{array}{c} \begin{array}{cccc} w_{TY} & w_{TS} & w_{CY} & w_{CS} \end{array} \\ \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array}$$

(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.

gives you evidence that you'll get the coal, but the sticker does nothing at all to change what Santa has gifted you. So CDT says you have no more reason to put the sticker on your gift than you have to put it on your sister's gift.

Hitchcock (2016) thinks that this is the wrong advice. He proposes a decision theory which agrees with the view I'm calling 'causal decision theory' in decisions like NO DIFFERENCE, but disagrees in decisions like STICKER. To understand Hitchcock's theory, it helps to first consider how orthodox causalists 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. If you abide by this norm and E is your total evidence, then your credences will be $C_0 \mid E$. Then, CDT advises you to evaluate acts for choiceworthiness with

$$\mathcal{U}(A) \;=\; (C_0 \mid E) \cdot would_A \cdot \mathcal{D}$$

In contrast, Hitchcock advises you to evaluate acts for choiceworthiness with

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

That is: you should *first* image your probability function on the performance of A, and *then* take your evidence into account by conditioning the probability function $C_0 \cdot would_A$ on E. You should then take an expectation of desirability using the resulting probability function, $(C_0 \cdot would_A) \mid E$.[8] In most decisions, there won't be any difference between $\mathcal{U}(A)$ and $\mathcal{H}(A)$. But if E includes foreknowledge about factors causally downstream of

8. Hitchcock (2016) uses causal Bayes nets rather than imaging to determine this probability function. If you're interested, I give the details in the appendix. There, I also show that his theory may be reformulated in terms of an 'imaging' function, as I've presented it in the main text.

your choice, then $\mathcal{U}(A)$ and $\mathcal{H}(A)$ can differ. (For the interested reader, I explain why in the appendix.)

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 say that you're required to put the sticker on your gift. For concreteness, suppose that, in the absence of any evidence, you think Santa was just as likely to gift the toy to you as he was to gift it to your sister, you think that you're equally likely to put the sticker on either gift, and you take these choices to be 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, $E = \{w_{TY}, w_{CS}\}$. Let's take $\mathcal{D} = [\mathcal{D}(w_{TY}), \mathcal{D}(w_{TS}), \mathcal{D}(w_{CY}), \mathcal{D}(w_{CS})]'$ to be $[1, 1, 0, 0]'$.[9] Then,

$$\mathcal{H}(Y) = (C_0 \cdot would_Y) \mid E \cdot \mathcal{D} = 1$$
$$\text{and} \qquad \mathcal{H}(S) = (C_0 \cdot would_S) \mid E \cdot \mathcal{D} = 0$$

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 alternative. 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, by stipulation, 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. 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 Santa gifted you. **Lesson #1** warns us to be on guard: our intuitive judgements about rational choice can lead us astray in precisely these kinds of cases.

Causalists should also notice that the 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. You see that you have been gifted the coal and she has been gifted the toy. From your perspective, there doesn't seem to be any instrumental value in placing the sticker on her gift. Moreover, from your perspective, a preference for placing the sticker on her gift can appear counter-productive. Suppose your sister values the toy at $100, and she

9. I've arbitrarily chosen to make 1 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.

is only allowed to put the sticker on her gift if she pays $90. From your perspective, it seems that paying to put the sticker on her gift is just throwing $90 away to no end. She'd still have the toy if she didn't pay the $90. Nothing depends on what you see inside the gifts. If you instead see that you have been gifted the toy and her the coal, it is equally difficult to see any instrumental value in paying $90 to place the sticker on her gift.

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 it seems to me that causalists have no reason to worry about orthodox CDT's verdicts in cases like STICKER. To be sure, there is an intuition that you shouldn't put the sticker on your sister's gift. But there is *also* an intuition that you shouldn't take Righty. Causalists who have learnt **Lesson #1** are used to dismissing knee-jerk intuitions when you have control over what to believe about ϕ without having control over whether ϕ. They should not care whether the illusion of control is due to reliable prediction or reliable prescience.

5 Foreknowledge and Chance

Imagine that you refuse a bet on whether a flipped coin lands heads, we flip the coin, and it lands heads. At the time when you had to choose whether to bet, there was only a 50% chance of you winning, conditional on you betting. Nonetheless, as Sidney Morgenbesser observed, the counterfactual "If you had taken the bet, you would have won" doesn't seem to have a 50% probability. Instead, its probability seems to be 100%.[10] I follow a number of authors in drawing the lesson that, when we make subjunctive suppositions, we hold fixed factors which are *causally independent* of our supposition.[11] When we subjunctively suppose that you take the bet, we imagine a possibility in which the coin still has a 50% chance of landing heads, and still lands heads. Both the outcome of the coin toss and the chances of that outcome are causally independent of how you choose, so both are held fixed when we consider what would have happened, had you taken the bet.

Cases like these suggests the following general lesson.

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 ϕ is causally independent of

10. The observation is attributed to Morgenbesser by Slote (1978, fn 33).

11. See, for instance, Bennett (2003, ch. 15), Edgington (2004), Schaffer (2004), and Kment (2006), among many others.

your choice, then ϕ would not change its truth-value, were you to choose differently.

This is a lesson we can learn without the extravagance of foreknowledge, and it is one we can learn before turning our attention to decision theory. But the lesson is relevant for decisions made with foreknowledge.

Consider, for instance[12]

FOREKNOWN LOSS

A fair coin will be flipped. Before it is flipped, you are offered a bet which pays out \$150 if the coin lands on heads, and only costs \$50. Before you decide whether to bet, the oracle arrives with news from the future: the coin will land on tails.

In this decision, Hall (1994), Meacham (2010), and Spencer (2020) 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 thinks that, if we say this, CDT will say that you are required to bet, in spite of your foreknowledge that it's a bet you'll lose.

Does CDT say that? That depends upon what would happen, were you to bet. There are four relevant possibilities. The coin either lands heads, H, or tails, T. And you either *b*et, B, or you do *n*ot, N. Let w_{HB} be a possibility at which the coin lands heads and you bet. Let w_{HN} be a possibility at which the coin lands heads and you do not bet, and likewise for w_{TB} and w_{TN}. Then—ignoring **Lesson #2** for the nonce—let us suppose that the probability that the coin would land heads, were you to bet, is 50%. That is, let us suppose that $would_B$ and $would_N$ are as shown in figure 4. We can assume that your desires are linear in 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 betting is constant. It is worth \$25, no matter which world you're at:

$$would_B \cdot \mathcal{D} = [25, 25, 25, 25]'$$

12. 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 to happen at each world, were you to bet. In figure 4b, the matrix $would_N(row)(column)$, which describes what would likely happen at each world, were you to not bet.

Whereas not betting has a constant instrumental value of $0:

$$would_N \cdot \mathcal{D} = [0, 0, 0, 0]'$$

So it won't matter what foreknowledge you possess. CDT will say that betting is rational, and not betting irrational, no matter what the oracle has told you.

I agree with Spencer that this is terrible advice. Given your foreknowledge, you know that the bet is a loser, this knowledge is not in any way contingent upon whether you bet or not, and betting doesn't make any difference to whether the coin lands heads or tails. So you shouldn't bet. Spencer and I both lay the blame on the functions $would_B$ and $would_N$ shown in figure 4. In generating those functions, I made two assumptions. First: the probability that ϕ would result, were you to choose A, is the chance of ϕ, conditional on your choosing A. And second: the chance of the coin landing heads, both conditional on you betting and conditional on you not betting, is 50%. Spencer rejects the second assumption. In his view, your foreknowledge makes it so that the objective chance of the coin landing heads is 0%. Having learnt **Lesson #2**, I reject the first assumption. In my view, we must distinguish the probability that ϕ would result, were you to choose A, from the objective chance of ϕ, conditional on you choosing A.

Sobel and Rabinowicz notice that, in decisions like FOREKNOWN LOSS, the first assumption 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 certainly 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 coun-

	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.

terfactuals. On that semantics, if A is true at w, then so too is $A \boxrightarrow w$.[13] In the context of Lewis's semantics for the subjunctive conditional, this imposes the requirement that $A \wedge C$ entails $A \boxrightarrow C$.[14] 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.[15] 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**.[16]

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 instrumental value of N will remain unchanged (it's still certain to bring you \$0, no matter what). But the instrumental value of B is different. If you don't actually bet, betting would bring you \$25 in expectation. But, if you *do* actually bet, betting would win you \$100, if the coin lands heads, and lose you \$50, if the coin

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

14. For a compelling defence of this principle, known as 'conjunction conditionalisation', see Walters & Williams (2013).

15. 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'. 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.

16. For instance, if we do not endorse **Strong Centering**, CDT will make implausible claims about the decision CHOOSING THE CHANCES in §6 below.

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_{TN}. So the utility of betting 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 betting will be less than the utility of not betting. So, if you find yourself inclining towards betting (and therefore, give yourself evidence that you will bet), CDT will advise you to *not* bet. However, if you listen to this advice, and learn that you have, $C(B)$ will fall below $1/3$. And at that point, the utility of betting will exceed the utility of not betting, and CDT will advise you to bet.

So, if we simply impose **Strong Centering**, then FOREKNOWN LOSS turns into a case in which CDT's recommendations are sensitive to your 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 usually correct about 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. But this 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 betting would lose you money.

If we take **Lesson #2** to heart, then we should want the imaging function $would_A$ to hold fixed factors which are causally independent of how you choose. That is, we should want it to satisfy the following constraint.

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

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

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

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

(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.

$would_B$ and $would_N$ are as shown in figure 6. If we use these imaging functions, then we will say that, if the coin lands heads, then betting would certainly win you $100; and, if the coin lands tails, betting would certainly lose you $50.

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

Given your foreknowledge that the coin will land tails, the utility of betting will be a certain loss of $50, and CDT will correctly tell you to not bet.

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 won't try to persuade causalists to accept this consequence of CDT. Instead, I will try to persuade them that it is not a problem which is unique to decisions made with foreknowledge. And I will try to persuade them that there is something deeply right about CDT's prediction-sensitive advice.

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,

$$\begin{array}{c} \\ w_{\alpha A} \\ w_{\alpha D} \\ w_{\delta A} \\ w_{\delta D} \end{array} \begin{array}{c} \begin{array}{cccc} w_{\alpha A} & w_{\alpha D} & w_{\delta A} & w_{\delta D} \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right] \end{array}$$

(a) $would_A$

$$\begin{array}{c} \\ w_{\alpha A} \\ w_{\alpha D} \\ w_{\delta A} \\ w_{\delta D} \end{array} \begin{array}{c} \begin{array}{cccc} w_{\alpha A} & w_{\alpha D} & w_{\delta A} & w_{\delta D} \end{array} \\ \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array}$$

(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.

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

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 kill you. Whereas, if Death is in Damascus, then going to Aleppo would save your life.

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

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 kill you.

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

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. So the utility of going to

Aleppo and Damascus, respectively, is

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

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 more than 50% confident that you'll go to Aleppo, then CDT says that going to Aleppo is irrational and you must instead go to Damascus. On the other hand, if you are more than 50% confident that you'll go to Damascus, then CDT says that going to Damascus is irrational, and you should instead go to Aleppo. Just to have a name for this kind of phenomenon, we can say that, in DEATH IN DAMASCUS, CDT *forbids your prediction*.

For another kind 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. You take your fairy godmother's predictions to be 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 just like DEATH IN DAMASCUS, except that 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 choice you think you're most likely to make is rational, and whichever you think you're least likely to make is irrational. Just to have a name for this kind of phenomenon, say that, in CAKE IN DAMASCUS, CDT *demands your prediction*.

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 explored revisions to CDT which deal with cases like these in a prediction-insensitive way.[17] 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 (for instance), 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 (for instance), 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!

So, in my view, these decisions teach us an important lesson. This is the dual of **Lesson #1**. **Lesson #1** taught us that, if whether ϕ is not under your control, but what to think about ϕ is under your control, then you may be subject to an agential illusion of control, and your knee-jerk intuitions about rational choice may be led astray. In decisions like DEATH IN DAMASCUS and CAKE IN 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 you actually go to Aleppo and are greeted with cake or death. Then, it was in your power to go to Damascus instead. 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 fatalistic 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.

The fatalistic illusion can vanish when you think about things from a better-informed, third-personal perspective. Imagine that your friend, rather than you, is deciding whether to go to Aleppo or Damascus. And imagine that, while your friend does not know where

17. See [author].

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 choice which leads them away from death—even though you know that they won't make this choice. 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 each of the possible better-informed, third-personal perspectives.

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. I'm going to take **Strong Centering** for granted,[18] 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 $\mathcal{D} = [\mathcal{D}(w_{HB}), \mathcal{D}(w_{HW}), \mathcal{D}(w_{TB}), \mathcal{D}(w_{TW})]' = [90, 0, -90, 0]'$. Then, the instrumental value of flipping the black coin will be $90, if you flip black and it lands heads, -$90 if you flip black and it lands tails, and $30 if you flip white. On the other hand, since flipping the white coin doesn't commit you to any bet, the instrumental value of flipping white is a constant $0.

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

18. 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.

	w_{HB}	w_{HW}	w_{TB}	w_{TW}
w_{HB}	1	0	0	0
w_{HW}	2/3	0	1/3	0
w_{TB}	0	0	1	0
w_{TW}	2/3	0	1/3	0

(a) $would_B$

	w_{HB}	w_{HW}	w_{TB}	w_{TW}
w_{HB}	0	1/3	0	2/3
w_{HW}	0	1	0	0
w_{TB}	0	1/3	0	2/3
w_{TW}	0	0	0	1

(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.

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 expect black to have a 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 each of the better informed, third-personal perspectives, 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 this 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; they are the only possibilities which your friend's foreknowledge doesn't rule out. 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 right about this, whether or not it's right to forbid your prediction.

In other decisions with foreknowledge, CDT will demand your prediction. Consider:[19]

> PAUPER'S PROBLEM
>
> You are a pauper. Tomorrow, the prince 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

19. This decision (or a close variant of it) is discussed in Lewis (1986), Rabinowicz (2009), Price (2012), and Stern (2021). 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. The two versions of CDT which Bales introduces to get the decisive 'buy the armour' and the decisive 'don't buy the armour' verdicts use imaging functions which either don't satisfy **Causal Independence** or **Strong Centering**. So those theories will give bad verdicts in either FOREKNOWN LOSS or CHOOSING THE CHANCES.

	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.

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 1, the desirability of losing your life savings but nor your life can be represented with 10, and likewise, $\mathcal{D} = [\mathcal{D}(w_{SB}), \mathcal{D}(w_{SN}), \mathcal{D}(w_{DB}), \mathcal{D}(w_{DN})]' = [10, 11, 0, 1]'$. 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

$$\mathcal{U}(B) = 9 + C(B)$$
$$\text{and} \quad \mathcal{U}(N) = 11 - 9 \cdot C(B)$$

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. But I do 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 each of the possible better informed, third-personal perspectives, 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 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. They are the only possibilities which you friend's foreknowledge doesn't rule out. 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

I have an important meeting, but I don't want to get out of bed. The meeting is more important to me than sleeping in, so sleeping in is irrational. Even so, I know that I'm going to sleep in. I reassure myself with the following: "The only reason I have to wake up is the meeting. But I already *know* that I'm going to miss the meeting (since I know that I'm going to sleep in). So there's nothing irrational about sleeping in." This reasoning is specious. Even though I know that I'm going to miss the meeting, this knowledge depends upon my knowledge that I'm going to sleep in. If I were to get out of bed, I wouldn't know that I'm going to miss the meeting.

Akratic choices like this teach us that you shouldn't always hold your knowledge fixed when deliberating about what to do. Sometimes, the things you know depend upon your irrationality. Were you to be rational, the knowledge would be lost. This knowledge shouldn't be taken for granted in deliberation about how to choose.

Lesson #4 Rational deliberation about what to do can sometimes defeat your knowledge. You shouldn't take this knowledge for granted when deliberating about what to do.

This lesson applies equally well to *fore*knowledge. Sometimes, rational deliberation

can defeat your foreknowledge. You may either have a guaranteed $1 or a guaranteed $100. Before you choose, the oracle arrives with news from the future: you will take the $1. What should you do? You should take the $100. Taking the $100 was the rational choice before hearing the prophesy. The prophesy doesn't change your desires or your beliefs about what would result from each choice. So taking the the $100 is still the rational choice after hearing the prophesy. Of course, if the prophesy is accurate, you won't take the $100—but why should this change what it's *rational* to do? Known irrationality is common in *akratic* decisions. Why should known irrationality be any more problematic when the irrationality is *fore*known?

In cases of *akrasia*, even if you know that you will choose irrationally, you wouldn't retain that knowledge, were you to choose differently. And the same is true when you *fore*know your own irrationality. Suppose you actually take the $1. 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. Knowledge is transmitted through testimony, so in these circumstances, you may come to know that you'll take the $1. But you would not have retained this knowledge, had you chosen differently. Had you taken the $100, either the oracle wouldn't have prophesied that you'd take the $1, or else she would have *falsely* prophesied that you'd take the $1. Either way, you would not be in a position to know that you take the $1 (because you don't). So, even though you actually know that you will take the $1, had you taken the $100, you would not have known this.

Suppose you actually take the $1. I say that you chose irrationally. You retort: "What could I have done? I *knew* that I was going to take the $1—it was inevitable." I reply: we should distinguish between being *known* and being *inevitable*. In the case of *akrasia*, I know that I will miss the meeting. But this doesn't mean that missing the meeting is *inevitable*. To answer your question 'what could I have done?': to start, you could have opened deliberation about what to do. At that point, you could 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. Since $100 is better than $1, you could and should have formed an intention to take the $100. As you watched yourself carry this intention out, you could and should have grown more and more confident that you will succeed in taking the $100, and so you could and should have grown less and less confident that the oracle knowingly prophesied that you will take the $1. You could then have taken the $100, being nearly certain that either the oracle prophesied falsely or that you misremembered her prophesy. Since you didn't follow this advice, you were in a position to know that you were going to take the $1. But you could have followed it; and if you had, you wouldn't have known that you were going to take the $1.

In the decisions from the previous sections, 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 other decisions, we must exercise more caution. For illustration, consider the following decision, introduced in an unrelated context by Stern (2021):

> 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 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 about how you will choose. It tells you that you are quite likely to choose to smoke.

If the oracle speaks from knowledge, then you will smoke, blacken your lungs, and thereby, quite likely, give yourself lung cancer. In my view, 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. I'm strongly inclined to say that a choice which kills you has less instrumental value than a choice would would have saved your life.

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, in this decision, you shouldn't take the oracle's prophesy for

granted in your deliberation about whether to smoke.

Parenthetically, while I am interested in how to choose in cases where your foreknowledge gives you information about what choice you will make, Stern is not. He uses this decision simply to illustrate Hitchcock's theory (recall §4), which Stern takes to give the right advice—namely, that you should smoke. He also does not discuss what other potential causes of your lungs blackening there might be. So, in explicitly stipulating that you're confident that nothing besides smoking causes your lungs to blacken, I'm building in a bit more to the case than he does. But this feature of the case is irrelevant to the advice of his decision theory. It will tell you to smoke whether or not there are other potential causes of black lung. 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. My advice is to 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 provides a decision theory which advises you to smoke in both of these decisions. 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

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

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, $C_0 \,|\, EF$, 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

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

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, $C_0 \,|\, E$, 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. (For the interested

reader, I have more to say about this theory in the appendix.)

This theory rejects **Lesson #4**, since it requires you to take your foreknowledge for granted when deliberating about what to do—even when deliberation could defeat that foreknowledge. Suppose you're deciding between a guaranteed $1 and a guaranteed $100. And suppose the oracle informs you that you will take the $100. In this decision, I am inclined to say 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. By stipulation, 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:

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

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.[20]

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

20. In the formalism of causal Bayes nets, 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.

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 decision. You are given a choice 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.

A Causal Decision Theory and Causal Bayes Nets

In this appendix, I'll explain how the formalism from §3 relates to decision theories formulated in terms of causal Bayes nets, in §A.1. In §A.2, I'll say how I think we should define the imaging function in terms of a causal Bayes net. And in §A.3 I'll discuss the relationship between evidentialism, causalism, and the form of 'interventionism' advocated by Meek & Glymour (1994), Hitchcock (2016), and Stern (2021).

A.1 Intervening and Imaging

A causal Bayes net is a pair, $\langle \mathcal{G}, \mathcal{P} \rangle$, of a directed graph, $\mathcal{G}$, and a probability function, $\mathcal{P}$, which satisfies the Markov condition relative to $\mathcal{G}$. The graph gives a collection of variables, $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$, and a collection of directed edges between those variables. In this framework, a *world* is a total assignment of values to variables. To simplify notation, I'll use 'V_w' for the value which V is assigned by the world w. When it appears inside a probability function, I'll use 'V_w' for the proposition that the variable V takes on the value it is assigned by w. If $\mathcal{P}$ satisfies the Markov condition relative to $\mathcal{G}$, then, for every world w,

$$\mathcal{P}(w) = \prod_{V \in \mathcal{V}} \mathcal{P}(V_w \mid \mathbf{PA}(V)_w)$$

where '$\mathbf{PA}(V)$' are V's 'parents' in the graphs. It is the collection of variables, P, such that a directed edge $P \rightarrow V$ is included in the graph.

I'll suppose that your act is represented in the Bayes net by a variable which I'll call 'A', for *act*. Each available act corresponds to a different value of this variable. We can define a 'manipulated' or 'intervened-upon' probability distribution $\mathcal{P}_{A=a}$ by simply replacing the conditional probability $\mathcal{P}(A_w \mid \mathbf{PA}(A)_w)$ in the product above with 1, if $A_w = a$, and 0, if $A_w \neq a$.[21]

In this framework, Meek & Glymour (1994) interpret causal decision theorists as saying that you should choose a value of A which maximises the quantity

$$\mathcal{U}(A = a) \stackrel{\text{def}}{=} \sum_w \mathcal{P}_{A=a}(w) \cdot \mathcal{D}(w) \;=\; \mathcal{P}_{A=a} \cdot \mathcal{D} \tag{1}$$

(On the right, I'm using '$\mathcal{P}_{A=a}$' for a row vector giving the 'post-intervention' probability distribution over worlds, and '$\mathcal{D}$' as a column vector giving the desirability of each world.)

21. For more, see Hitchcock 2018, §3.

Given my terminology, this counts as a version of causal decision theory, because there is an 'imaging' function, $would_{A=a}$, such that $would_{A=a}(w)(w^*)$ captures $A = a$'s causal tendency to bring about w^* at the world w, and such that $\mathcal{P}_{A=a}$ corresponds to 'imaging' on $A = a$,

$$\mathcal{P}_{A=a} = \mathcal{P} \cdot would_{A=a}$$

(Here, I'm using '$\mathcal{P}$' as a row vector for the 'unmanipulated' probability distribution over worlds, and '$would_{A=a}$' is a square matrix whose entry in row *row* and column *col* is $would_{A=a}(w_{row})(w_{col})$.) I'll explain how to define $would_{A=a}$ so that this identity holds, and prove that it does, in this footnote.[22]

22. Let '$\mathbf{ND}(A)$' be the non-descendants of A. Let '$\mathbf{DE}(A)$' be the descendants of A (excluding A itself). And let $\mathbf{1}_v(w)$ be the 'indicator' function for $V = v$, which is 1 if $V_w = v$ and is 0 if $V_w \neq v$. Then, we may define $would_{A=a}(w)(w^*)$ to be the product

$$\mathbf{1}_a(w^*) \cdot \prod_{N \in \mathbf{ND}(A)} \mathbf{1}_{N_w}(w^*) \cdot \prod_{D \in \mathbf{DE}(A)} \mathcal{P}(D_{w^*} \mid \mathbf{PA}(D)_{w^*})$$

This tells us that $A = a$ has no causal tendency to change the values of variables which aren't causally downstream of A, and that its causal tendency to change the values of the variables causally downstream of it is given by the conditional probabilities from the Bayes net.

Then, note that

$$(\mathcal{P} \cdot would_{A=a})(w^*) = \sum_w \mathcal{P}(w) \cdot would_{A=a}(w)(w^*)$$

If $A_{w^*} \neq a$, then $would_{A=a}(w)(w^*)$ will be zero for every choice of w. So assume that $A_{w^*} = a$. And write that $w \sim w^*$ iff the variables in $\mathbf{ND}(A)$ take on the same values in w as they do in w^*. Then, since $would_{A=a}(w)(w^*)$ is zero for every choice of w such that $w \nsim w^*$,

$$\begin{aligned}
(\mathcal{P} \cdot would_{A=a})(w^*) &= \sum_{w\,:\,w \sim w^*} \mathcal{P}(w) \cdot would_{A=a}(w)(w^*) \\
&= \sum_{w\,:\,w \sim w^*} \mathcal{P}(w) \cdot \prod_{D \in \mathbf{DE}(A)} \mathcal{P}(D_{w^*} \mid \mathbf{PA}(D)_{w^*}) \\
&= \left(\prod_{D \in \mathbf{DE}(A)} \mathcal{P}(D_{w^*} \mid \mathbf{PA}(D)_{w^*}) \right) \cdot \left(\sum_{w\,:\,w \sim w^*} \mathcal{P}(w) \right)
\end{aligned}$$

Summing up the probability given to every world in which the non-descendants of A take on the same values as they do in w^* is the same as taking the product of the terms $\mathcal{P}(N_{w^*} \mid \mathbf{PA}(N)_{w^*})$, for each $N \in \mathbf{ND}(A)$.

$$\sum_{w\,:\,w \sim w^*} \mathcal{P}(w) = \prod_{N \in \mathbf{ND}(A)} \mathcal{P}(N_{w^*} \mid \mathbf{PA}(N)_{w^*})$$

So, whenever $A_{w^*} = a$,

$$(\mathcal{P} \cdot would_{A=a})(w^*) = \left(\prod_{D \in \mathbf{DE}(A)} \mathcal{P}(D_{w^*} \mid \mathbf{PA}(D)_{w^*}) \right) \cdot \left(\prod_{N \in \mathbf{ND}(A)} \mathcal{P}(N_{w^*} \mid \mathbf{PA}(N)_{w^*}) \right)$$

And when $A_{w^*} \neq a$, $(\mathcal{P} \cdot would_{A=a})(w^*) = 0$. And this is the same probability distribution over worlds given by $\mathcal{P}_{A=a}$. So, in general, $\mathcal{P} \cdot would_{A=a} = \mathcal{P}_{A=a}$.

This equivalence is what allowed me, in the main text, to represent the theories of Hitchcock (2016) and Stern (2021) with an 'imaging' function $would_A$. With the causal Bayes net formalism, I could more faithfully present those theories by saying that, in the relevant decisions, Hitchcock and Stern advise you to maximise the quantities $\mathcal{H}$ and $\mathcal{S}$, respectively, as defined below.

$$\mathcal{H}(A = a) \stackrel{\text{def}}{=} \sum_w \mathcal{P}_{A=a}(w \mid EF) \cdot \mathcal{D}(w)$$
$$\mathcal{S}(A = a) \stackrel{\text{def}}{=} \sum_w \mathcal{P}_{A=a,F}(w \mid E) \cdot \mathcal{D}(w)$$

(Here, I've used 'E' for your ordinary evidence and 'F' for your foreknowledge.) Given the definition of '$would_{A=a}$' I specified above, these are equivalent to the definitions from the main text, namely:

$$\mathcal{H}(A = a) \stackrel{\text{def}}{=} ((\mathcal{P} \cdot would_{A=a}) \mid EF) \cdot \mathcal{D}$$
$$\mathcal{S}(A = a) \stackrel{\text{def}}{=} ((\mathcal{P} \cdot would_{A=a,F}) \mid E) \cdot \mathcal{D}$$

I'll suppose that, in the absence of evidence, your credences should correspond to the objective probabilities represented in the Bayes net. And if you have evidence about the values of variables, E, you should condition on this evidence, so that your credence distribution over worlds will be given by $C = \mathcal{P} \mid E$. If E doesn't concern the values of variables 'downstream' of A, then 'imaging' and 'conditioning' will commute, so that $(\mathcal{P} \mid E) \cdot would_{A=a} = (\mathcal{P} \cdot would_{A=a}) \mid E$. Thus, as long as you lack any evidence about the values of variables 'downstream' of A, both $\mathcal{H}(A = a)$ and $\mathcal{S}(A = a)$ reduce to

$$C \cdot would_{A=a} \cdot \mathcal{D}$$

which is why I say in the main text that Hitchcock and Stern agree with (what I am calling) causal decision theory in cases where you lack foreknowledge.[23] As STICKER from §4 illustrates, when you have foreknowledge, 'imaging' and 'conditioning' need not commute, which is why I say that Hitchcock and Stern disagree with (what I am calling) causal

23. To appreciate that 'imaging' and 'conditioning' commute when E isn't about any variables 'downstream' of A, note that, if $E \notin \mathbf{DE}(A)$, it follows from the Markov condition that $\mathcal{P}_{A=a}(E) = \mathcal{P}(E)$. Then, for any world w which makes E true, $\mathcal{P}_{A=a}(w \mid E) = \mathcal{P}_{A=a}(w)/\mathcal{P}_{A=a}(E) = \mathcal{P}_{A=a}(w)/\mathcal{P}(E)$. And $((\mathcal{P} \mid E) \cdot would_{A=a})(w) = \sum_{w^*} \mathcal{P}(w^* \mid E) \cdot would_{A=a}(w^*)(w) = (\sum_{w^* \in E} \mathcal{P}(w^*) \cdot would_{A=a}(w^*)(w))/\mathcal{P}(E)$. So, if $\mathcal{P}_{A=a}(w) = \sum_{w^* \in E} \mathcal{P}(w^*) \cdot would_{A=a}(w^*)(w)$, then 'imaging' and 'conditioning' commute. But, when $E \notin \mathbf{DE}(A)$, if w makes E true, then $would_{A=a}(w)(w^*) = 0$ whenever w^* doesn't make E true. So $\mathcal{P}_{A=a}(w) = \sum_{w^*} \mathcal{P}(w^*) \cdot would_{A=a}(w^*)(w) = \sum_{w^* \in E} \mathcal{P}(w^*) \cdot would_{A=a}(w^*)(w)$. □

decision theory in cases where you have foreknowledge.

A.2 A Better Imaging Function

The definition of the 'imaging' function $would_{A=a}$ which yields the equivalence

$$\mathcal{P} \cdot would_{A=a} = \mathcal{P}_{A=a}$$

is not strongly centered. In my view, decisions like CHOOSING THE CHANCES from §6 show us that $would_{A=a}$ should be strongly centered. This can be achieved by saying that, if $A_w = a$, then $would_{A=a}(w)(w^*)$ is 1 if $w^* = w$ and is 0 if $w^* \neq w$. And, if $A_w \neq a$, then $would_{A=a}(w)(w^*)$ should be given by the product

$$\mathbf{1}_a(w^*) \cdot \prod_{N \in \mathbf{ND}(A)} \mathbf{1}_{N_w}(w^*) \cdot \prod_{D \in \mathbf{DE}(A)} \mathcal{P}(D_{w^*} \mid \mathbf{PA}(D)_{w^*})$$

where $\mathbf{1}_v(w)$ is the 'indicator' function for $V = v$ which is 1 if $V_w = v$ and is 0 if $V_w \neq v$, $\mathbf{ND}(A)$ are A's *non-descendants*, and $\mathbf{DE}(A)$ are A's *descendants*, excluding itself. (Given this definition of the imaging function, $\mathcal{P} \cdot would_{A=a}$ will not always be equal to $\mathcal{P}_{A=a}$.)

If we want to characterise utility with an imaging function derived from a causal Bayes net, then in my view, this is the imaging function we should favour.

A.3 Interventionism, Causalism, and Evidentialism

Meek & Glymour pair the decision theory encoded in (1) with the assumption that you have a genuinely unpredictable Will, capable of intervening upon and interrupting the world's default causal order. Just to have a name, let's call this view about the Will 'interventionism'. Hitchcock (2016) and Stern (2021) are both interventionists, in this sense. (To be clear: I am not an interventionist. I believe that free choices are causally influenced by, and predictable on the basis of, the past state of the natural world.) If we are interventionists, then we may understand $\mathcal{P}_{A=a}$ as the probability function $\mathcal{P}$ *conditioned* on the event of the Will intervening on the world's default causal structure to make it so that $A = a$.[24]

In section A.1, I showed that, so long as you lack foreknowledge, the 'intervened upon' probability distribution $\mathcal{P}_{A=a}$ is equivalent to an 'imaged' probability function

24. To understand why we can understand $\mathcal{P}_{A=a}$ in this way, see Spirtes *et al.* (2000, theorem 3.6, p. 51).

$\mathcal{P} \cdot would_{A=a}$, given a natural definition of $would_{A=a}$ which captures $A = a$'s causal tendency to bring about various worlds. In part for this reason, Meek & Glymour (1994) teach that the debate between evidential and causal decision theorists "does not turn on any difference in normative principles, but on a substantive difference about the causal processes at work in the context of decision making" (p. 1009). Hitchcock and Stern broadly agree with this characterisation. For instance, Hitchcock (2016) writes: "We start with vague questions: What should I do? Which action is rational? The causal decision theorist, according to Meek and Glymour, aims to replace these with a precise question: If I were to intervene to set the value of the action variable, A, which value of the variable would have the highest expected payoff?...I think genuine philosophical progress has been made by making it clear what question the causal decision theorist correctly answers" (pp. 1163–64). And Stern (2018) writes: "I am in agreement with [Meek & Glymour's] key insight that the nature of the disagreement between evidential decision theorists and causal decision theorists is *not* best treated as a disagreement about fundamental normative principles that govern rational choice, but rather as a disagreement about the nature of choice" (p. 203).

I agree with Meek & Glymour this far: if we accept the interventionist's views about the Will, the evidentialist's news-value can agree with the causalist's utility whenever your evidence is ordinary. That is to say: so long as you don't have evidence about the values of variables causally downstream of the 'act' variable A, conditioning on the event of an intervention to make $A = a$ will be the same as 'imaging' on $A = a$, given a natural imaging function (that is what we showed in §A.1 above). So, when your evidence is ordinary, the news-value of an intervention can be equal to the utility of that intervention. I agree with them this far—but no farther. When you have extraordinary evidence about the effects of your choice, news-value and utility can come apart—*whether or not* we possess an extraordinary Will which interrupts the world's default causal order. This is one of the lessons of decisions like STICKER, from §4. In that decision, once you condition on the oracle's prophesy, placing the sticker on your own gift has a greater news-value than placing it on your sister's gift, but placing the sticker on your own gift has exactly the same utility as placing it on your sister's gift. And that's so *whether or not* we assume that your choice constitutes an intervention on the world's default causal order.

So, in my view, decisions made with foreknowledge show us that Meek & Glymour's diagnosis of the disagreement between evidential and causal decision theory is simply incorrect. There is a genuinely normative disagreement between these theories. That disagreement cannot be understood in terms of a non-normative disagreement about interventionism. For, in decisions made with foreknowledge, the disagreement persists whether interventionism is true or false.

Decisions made with foreknowledge teach us that interventionists must answer the same normative questions about rational choice as the rest of us. They could side with causalists and advise you to maximise utility; they could side with evidentialists and advise you to maximise news-value; or they could give you different advice altogether. Here, I've made the case for siding with causalists—whether or not interventionism is true. In contrast, Hitchcock (2016) advises you to maximise news-value even when you have foreknowledge, essentially siding with evidentialists. Stern (2021) gives different advice altogether. He advises you to choose as if your choice were evidentially irrelevant to any variables about which you have evidence.[25] Interventionism on its own will not settle these normative disagreements.

References

Bales, Adam. 2016. "The pauper's problem: chance, foreknowledge and causal decision theory." In *Philosophical Studies*, **173**: 1497–1516. [4], [23]

Bennett, Jonathan. 2003. *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press. [12]

Edgington, Dorothy. 2004. "Counterfactuals and the benefit of hindsight." In *Cause and Chance: Causation in an Indeterministic World*, edited by Phil Dowe & Paul Noordhof, London: Routledge, chapter 2, 12–27. [12]

Gärdenfors, Peter. 1982. "Imaging and Conditionalization." In *Journal of Philosophy*, **79** (12): 747–760. [4]

Gibbard, Allan & Harper, William L. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, edited by A. Hooker, J.J. Leach, & E.F. McClennan, Dordrecht: D. Reidel, 125–162. [4], [17], [36]

Hall, Ned. 1994. "Correcting the Guide to Objective Chance." In *Mind*, **103** (412): 505–517. [13]

Hitchcock, Christopher. 2016. "Conditioning, Intervening, and Decision." In *Synthese*, **194** (4): 1157–1176. [1], [10], [11], [28], [31], [33], [34], [35], [36]

25. Stern believes that this advice honours "causal-decision-theoretic reasoning", and leads to "verdicts that causal decision theorists find intuitive". But the advice won't maximise utility, as 'utility' is defined by Stalnaker (1981), Gibbard & Harper (1978), Lewis (1981), Skyrms (1982), Sobel (1994), or Joyce (1999). Nor will Stern's theory fall within the umbrella definition of 'causal decision theory' I introduced in §3. So he won't count as siding with causalists, given my terminology.

Hitchcock, Christopher. 2018. "Probabilistic Causation." In *Stanford Encyclopedia of Philosophy*. [31]

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press. [4], [36]

Kment, Boris. 2006. "Counterfactuals and Explanation." In *Mind*, **115**: 261–309. [12]

Lewis, David K. 1973. *Counterfactuals*. Malden, MA: Blackwell Publishers. [14], [15]

Lewis, David K. 1980. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability*, edited by Richard C. Jeffrey, Berkeley: University of California Press, volume II, 263–293. [4]

Lewis, David K. 1981. "Causal Decision Theory." In *Australasian Journal of Philosophy*, **59** (1): 5–30. [1], [4], [8], [15], [36]

Lewis, David K. 1986. "Postscript to 'Causal Decision Theory'." In *Philosophical Papers, volume 2*, Oxford: Oxford University Press. [15], [23]

Meacham, Christopher J. G. 2010. "Two Mistakes Regarding the Principal Principle." In *British Journal for the Philosophy of Science*, **61** (2): 407–431. [13]

Meek, Christopher & Glymour, Clark. 1994. "Conditioning and Intervening." In *The British Journal for the Philosophy of Science*, **45**: 1001–1021. [31], [34], [35]

Price, Huw. 2012. "Causation, Chance, and the Rational Significance of Supernatural Evidence." In *The Philosophical Review*, **121** (4): 483–538. [1], [13], [23]

Rabinowicz, Wlodek. 2009. "Letters from Long Ago: On Causal Decision Theory and Centered Chances." In *Logic, Ethics, and All That Jazz—Essays in Honour of Jordan Howard Sobel*, edited by Lars-Göran Johansson, Uppsala: Uppsala Philosophical Studies, volume 56, 247–273. [4], [15], [23]

Rabinowicz, Włodzimierz. 1982. "Two Causation Decision Theories: Lewis vs Sobel." In *Philosophical Essays Dedicated to Lennart Åqvist on His Fiftieth Birthday*, edited by Tom Pauli, Uppsala: Uppsala Philosophical Studies, volume 34, 299–321. [4], [14]

Roberts, John T. ms. "Must a Cause be Earlier than its Effect?" [8], [13]

Schaffer, Jonathan. 2004. "Counterfactuals, Causal Independence and Conceptual Circularity." In *Analysis*, **64** (4): 299–309. [12]

Skyrms, Brian. 1982. "Causal Decision Theory." In *Journal of Philosophy*, **79** (11): 695–711. [4], [36]

Slote, Michael A. 1978. "Time in Counterfactuals." In *The Philosophical Review*, **87** (1): 3–27. [12]

Sobel, Jordan Howard. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge: Cambridge University Press. [4], [14], [15], [36]

Spencer, Jack. 2020. "No Crystal Balls." In *Noûs*, **54** (1): 105–125. [1], [13], [14]

Spirtes, Peter, Glymour, Clark, & Scheines, Richard. 2000. *Causation, Prediction, and Search*. Cambridge, MA: The MIT Press, second edition. [34]

Stalnaker, Robert C. 1981. "Letter to David Lewis." In *Ifs*, edited by William Harper, Robert Stalnaker, & Glenn Pearce, Dordrecht: D. Reidel Publishing Company, 151–152. [4], [36]

Stern, Reuben. 2018. "Diagnosing Newcomb's Problem with Causal Graphs." In *Newcomb's Problem*, edited by Arif Ahmed, Cambridge: Cambridge University Press, chapter 10, 201–220. [35]

Stern, Reuben. 2021. "An Interventionist's Guide to Exotic Choice." In *Mind*, **130** (518): 537–566. [1], [23], [27], [28], [29], [30], [31], [33], [34], [36]

Walters, Lee & Williams, J. Robert G. 2013. "An Argument for Conjunction Conditionalization." In *The Review of Symbolic Logic*, **6** (4): 573–588. doi:10.1017/S1755020313000191. [15]