Causal Decision Theory, Context, and Determinism

Calum McNamara*

Forthcoming in *Philosophy and Phenomenological Research*

Abstract

The classic formulation of causal decision theory (CDT) appeals to counterfactuals. It says that you should aim to choose an option that *would* have a good outcome, *were* you to choose it. However, this version of CDT faces trouble if the laws of nature are deterministic. After all, the standard theory of counterfactuals says that, if the laws are deterministic, then if anything—including the choice you make—were different in the present, either the laws would be violated or the distant past would be changed. And as several authors have shown, it's easy to transform this upshot of the standard theory of counterfactuals into full-blown counterexamples to CDT. In response to these counterexamples, I argue here that the problem lies, not so much with CDT's guiding idea—that it's the expected causal consequences of your actions that matter for rational decision-making—but with the fact that the classic formulation of CDT doesn't pay sufficient attention to the context-sensitivity of counterfactuals. I develop a contextualist version of CDT which better accounts for this context-sensitivity. And I show that my theory avoids the problems faced by the classic formulation of CDT in determinstic worlds.

1 Introduction

Here is a bet—take it or leave it. You win \$1 if a proposition, P, is true, but you lose \$1 if P is false. Before you choose whether to accept or decline this bet, I'll tell you what P is. It's the proposition that the past state of the world, together with the laws of nature, determines that you accept.

Suppose you're certain of determinism. That is, suppose you're certain that the past state of the world, together with the laws of nature, determines whatever it is that you actually do (although in the present case, you're uncertain precisely *what* these things determine you'll do). Then, should you accept my bet? Or should you decline it? It seems perfectly clear that you should accept. After all, by your lights the proposition P is true only if you accept the bet. And it's false only if you decline. So, by accepting, it seems like you're sure to be a dollar better off than you'd otherwise be. Taking the bet is like accepting free money.

Cases similar to this one have come up quite often in the recent philosophical literature. And like the case just described, they're usually cases in which the best course of action is intuitively clear. Surprisingly,

*This paper has been a while in the making, and thanks are due to many people. For helpful discussions, I'm grateful to Arif Ahmed, Dave Baker, Kevin Blackwell, Clara Bradley, Cian Dorr, Adam Elga, Dmitri Gallow, Veronica Gómez-Sánchez, Daniel Herrmann, Josh Hunt, Simon Huttegger, Saira Khan, Boris Kment, Matt Mandelkern, Eduardo Martinez, Aidan Penn, Maria Rosala, Richard Roth, Elise Woodard, audiences at the University of Michigan and the Central APA, and especially Snow Zhang. For comments on earlier drafts of the paper, I'm grateful to Zach Barnett, Francisco Calderón, Melissa Fusco (my wonderful commentator at the APA), Mikayla Kelley, and an anonymous reviewer for this journal. Finally, special thanks are due to the members of my dissertation committee: Gordon Belot, Sarah Moss, Brian Weatherson, and especially Jim Joyce. The possible world in which this paper was written without the unwavering support of these four people is one that's very dissimilar to the actual world indeed.

however, *causal decision theory* (CDT)—a theory that many regard as our best theory of rational decision-making—gets these cases wrong. It recommends courses of action that almost everyone can agree are irrational.

According to CDT, you should make choices by considering the expected *causal* consequences of your actions. Different versions of the theory attempt to make this idea precise in different ways. My preferred version—namely, the version of Stalnaker (1981b), refined by Gibbard and Harper (1978)—appeals to the close connection between causation, on the one hand, and *counterfactuals*, on the other. Roughly, it says that you should choose an option that you think *would* have a good outcome, *were* you to choose it.

However, the standard theory of counterfactuals—to which this version of CDT usually appeals—has a surprising upshot, if the laws of nature are deterministic. Specifically, it says that if anything, including the choice you make, were different in the present, either the laws would be violated or the distant past would be changed. It's this surprising upshot of the standard theory of counterfactuals that leads my preferred version of CDT to give the absurd recommendations in the cases that I mentioned. Other versions of CDT face similar difficulties, for closely related reasons.[1]

My aim here is to slightly refine the Stalnaker-Gibbard-Harper formulation of CDT, so that it avoids the problems posed by the "deterministic cases" I've been talking about. In my view, what these cases show isn't so much that there's a fault with CDT's guiding idea—that it's the expected causal consequences of your actions that matter for rational decision-making—but instead that Stalnaker-Gibbard-Harper CDT, at least as it's usually spelled out, doesn't pay sufficient attention to the *context-sensitivity* of counterfactuals. In response to this, I develop a "contextualist" version of Stalnaker-Gibbard-Harper CDT, which better accounts for this context-sensitivity. And I show that my theory avoids the problems faced by the classic formulation of CDT in determinsitic worlds.[2]

In §2 below, I introduce the Stalnaker-Gibbard-Harper version of CDT, as well as the standard theory of counterfactuals. Then, in §3 I show that this theory gives the wrong recommendation in two well-known deterministic cases, both of which are due to Arif Ahmed (2013, 2014a, 2014b). In §§4–5 I introduce my theory: §4 starts with some background, as well as a general overview of the theory; and §5 gives some important further details. §6 then concludes the paper by returning to Ahmed's cases, and showing that my theory gets the right answer in them, as well as in related cases.

Before we get started, let me make two comments.

First, since nearly all of the cases I'm interested in here appeal to deterministic laws of nature, I'll assume determinism in what follows. More precisely, I'll assume that all the worlds under consideration obey deterministic laws. And I'll assume that this is something about which *you*—the agent facing the decision problems we discuss below—are certain. For present purposes, we can understand a system of laws to be deterministic just in case the following holds: any two worlds that obey those laws are either always exactly alike or never exactly alike, with respect to particular matters of fact (Lewis, 1979, p. 460). I'll leave it as a task for future work to see how well my theory generalizes to cases involving indeterministic laws. But for what it's worth, I think there's reason to be optimistic about its prospects.[3]

[1] See Skyrms (1980, 1982, 1984), Lewis (1981), Sobel (1994), or Joyce (1999) for other versions of CDT. Then, see Ahmed (2013, 2014a, 2014b), Solomon (2021), Elga (2022), and Hedden (2023) for discussions of the problems raised by "deterministic cases" for these other theories.

[2] The approach I advocate for here is briefly suggested by Elga (2022, pp. 211-12) as an approach worth exploring. Also, while this paper was under review, I learned that Robert Stalnaker has recently sketched a response to a deterministic case that's broadly similar to mine (see §6.4, and his MS for details). There are a few important differences between Stalnaker's approach and mine, and I'll point these out as I go along. However, for the most part, I take this over-arching convergence to be good news: as the reader will notice, the view I spell out here is broadly Stalnakerian in spirit.

[3] A couple of other remarks about laws of nature. First, throughout, I use 'laws' and 'laws of nature' as a shorthands for 'fundamental physical laws of nature'. I also assume that laws of nature are inviolable. This assumption is not wholly uncontroversial (see, e.g., Lange (2000), Braddon-Mitchell (2001), and Kment (2006, 2014) for dissent). But I don't think rejecting it makes for a very promising response to the deterministic cases. So I won't explore it here.

Secondly, some authors have recently argued that deterministic cases are not genuine decision problems. For, apparently, no agent who faces one can see herself as *free*.[4] This is something I disagree with. But for now I'll set my disagreement aside. Going forward, I'll assume that any agent facing a deterministic case can see herself as free, in some non-trivial sense. That my approach gets us the right answers in these cases, while also allowing us to make this assumption, is, I think, one of its main draws for those of us with both causalist and compatibilist commitments.

2 CDT and Counterfactuals

Whenever you face a choice, you'll have some *options* available to you, $A_1, ..., A_n$. Here, I'll take your options to be propositions, which—for now—I take to be sets of worlds. I'll also assume that your options form a *partition* of the space of worlds, in the sense that each world w is a member of exactly one A_i. Intuitively, we can think of your options as the finest-grained propositions you believe you can *make* true by deciding (cf. Jeffrey, 1983, p. 84).

You'll also have *outcomes* that can result from your choice, $O_1, ..., O_m$. I'll take these, too, to be propositions that form a partition. And I'll assume they're propositions whose truth would settle everything that you care about.

Now, let cr be your credence function (subjective probability function). Let v be your subjective value function. And let $>$ be an operator, which takes a pair of propositions, P, Q, and returns the counterfactual $P > Q$. Then, CDT—at least in the Stalnaker-Gibbard-Harper formulation—says that you should choose an option, A, that maximizes *utility*, U, defined as follows:

$$U(A) = \sum_i cr(A > O_i) \cdot v(O_i). \tag{1}$$

As I said before, the idea here is that you should choose an option that you think *would* have a good outcome, *were* you to choose it.

Notice that I haven't yet mentioned causation. However, earlier, I said that, according to CDT, it's the expected *causal* consequences of your actions that matter for rational decision-making. So, we still need to say how the *counterfactual* rule above reflects this guiding idea. And to do that, we need to make some additional assumptions about the counterfactuals $A > O_i$.

For starters, let's assume they have the following standard semantics, due to Stalnaker (1968).[5] Let f be a *selection function*: a function that takes a proposition P and a world w as arguments, and returns a world $f(P, w)$, thought of, intuitively, as the "most similar" P-world to w. Then, Stalnaker's semantics says that a counterfactual $P > Q$ is true at w just in case Q is true at this most similar P-world, $f(P, w)$.[6]

Let's also make an assumption about the meaning of 'most similar P-world'. After all, not just any relation of similarity will do for present purposes. To see why, consider an example from Jackson (1977). Imagine that Fred is on the roof of a tall building, teetering on the edge. A moment later, he steps down. So I turn to you and say: "Thank goodness!

[4]See especially Joyce (2016) and Solomon (MS). Note, however, that Joyce has stressed to me in conversation that he doesn't think being certain of determinism precludes the possibility that an agent can see herself as free *simpliciter*. Instead, he thinks this is merely a special feature of certain of the decision problems we'll encounter below.

[5]See also Stalnaker and Thomason (1970). Lewis (1973b) gives a very similar semantics for counterfactuals, although it differs from Stalnaker's in a few crucial ways. It's well known, however, that Lewis's semantics coincides with Stalnaker's, given the assumption of determinism. Thus, since I'm making that assumption in this paper, the differences between Stalnaker's theory and Lewis's aren't relevant here.

[6]This semantics assumes that there always *is* a P-world to be selected. A more general version of the semantics would relax this assumption, with a clause saying what happens when there's no P-world to be selected (see, e.g., Stalnaker (1968)). For present purposes, however, I'll set that case aside.

(1) If Fred had jumped, he would've died."

Puzzled by this, you respond to me: "That's not true; Fred's not suicidal. He would've jumped only if there had been a net below him. So,

(2) if Fred had jumped, he would've lived."

Here, it doesn't seem like either of us has said anything false. But then, it's also clear that the two counterfactuals we've uttered can't be true at the same time. The most plausible explanation of what's going on invokes *context-sensitivity*. When I uttered my counterfactual, we were in a context at which the most similar antecedent-world was one where there's no net below Fred at the time of his jump. When you uttered your counterfactual, we were in a context at which the most similar antecedent-world was one in which a specific causal precursor for Fred's jumping is salient—namely, there being a net below him. The function of your preamble—"That's not true; Fred's not suicidal…"—was to set up this latter context. Thus, my counterfactual is true in the first context, and your counterfactual is true in the second.[7]

Lewis (1979) calls counterfactuals like mine "standard counterfactuals", and counterfactuals like yours "backtracking counterfactuals". Very roughly, we can think of the former as counterfactuals for which the most similar antecedent-world is one that's like the world of evaluation with respect to matters of fact in the past. And we can think of the latter as counterfactuals for which the past varies. (I'll revisit the former gloss later on.) Lewis also argues—convincingly, in my view—that it's only the first kind of counterfactual that can tell us about the *causal* effects of the antecedent on the consequent. And that, in a nutshell, is what we're after here. So, going forward, let's set backtracking counterfactuals aside, and assume that any counterfactual under discussion has a "standard" interpretation.[8]

To pin down the notion of a standard counterfactual more precisely, let's again follow Lewis—at least for now—in saying that, when P is about a nomically possible, dated event, the most similar P-world to w is one that's like w with respect to the following conditions:

(i) it matches w in all particular matters of fact at times before P, and

(ii) it obeys w's laws.

These criteria are plausible, not least because they deliver the right verdict in cases like Jackson's. To see this, just notice that, because there was no net below Fred when he was up on the roof, it follows by (i) that the most similar world at which he jumps is also a world where there's no net below him. Then, by (ii), it follows that Fred dies after jumping off the roof, since the most similar world at which he jumps is a world where gravity works the same as we're used to.

Notice also, however, that if w is a world with deterministic laws of nature, and P is a proposition that's false at w, then the most similar P-world to w can't be a world that satisfies (i) and (ii) perfectly.[9] After all, if the laws are deterministic, then the intrinsic state of the world at any time, together with the laws, determines its state at all times. Thus, if the most similar P-world to w matched w perfectly with respect to both (i) and (ii), it would have to be a world at which $\neg P$ is true. But by assumption, it's a world at which P is true. So at this world, a contradiction is true. And this makes P *counterfactually impossible.*

[7] I'm speaking loosely here. Really, it's the sentences that express counterfactuals that are context-sensitive, and not the counterfactuals themselves. But for present purposes, I'll mostly elide the distinction between propositions and sentences, since it simplifies things to do so.

[8] Some philosophers argue that the distinction between standard and backtracking counterfactuals is merely one of degree, rather than kind (see, e.g., Holguín and Teitel (MS)). To make things simple here, however, I'm going to assume there's a clear-cut distinction between these two kinds of counterfactuals. For a well worked-out theory of this distinction, with which I'm broadly sympathetic, see Khoo (2017, 2022).

[9] The argument I give here closely follows Dorr (2016). Note that there's an unstated closure premise in the argument, as I state it. See Dorr's paper for a more careful presentation.

Since we're interested in spelling out CDT using counterfactuals, this isn't a consequence we can live with. So, we need to reject the claim that the most similar P-world to w is one that satisfies (i) and (ii) *perfectly*. Instead, we need to say something like: the most similar P-world to w is a world that provides the best *trade-off* between (i) and (ii).

The most influential account of this trade-off is, again, given by Lewis (1979). According to him, the best trade-off-world is one that matches w with respect to all matters of particular fact up until a time shortly before P, but which does not obey w's laws. Instead, it obeys a system of laws similar to those that obtain at w, but which permit a "local divergence miracle"—a small violation of w's laws, sufficient to bring P about.[10]

There are other ways we could go with respect to this trade-off, if we wished. For instance, Dorr (2016) gives a different account of similarity, according to which the best trade-off world is one that obeys w's laws perfectly throughout all time, and which is also like w with respect to "macro-history", but not with respect to "micro-history".[11] However, since causal decision theorists almost always work with Lewis's account by default;[12] and since none of my conclusions would change if we adopted Dorr's account instead;[13] I'll take the former as my foil in what follows. From here on out, I'll call it the *miracles account*.

As an example of how CDT works when combined with the miracles account of similarity, consider the following decision problem (Nozick, 1969):

> *Newcomb.* In front of you are two boxes, A and B. Box A is opaque, and contains $\$1,000,000$ ($\$1m$) or nothing, but you don't know which. Box B is transparent, and contains a $\$1,000$ bill ($\$1k$). You have two options: either take just the opaque box (*One-box*); or take both boxes (*Two-box*). The catch is that, yesterday, a highly reliable predictor predicted which of these things you'd do. If she predicted that you'd take just the opaque box, then she put the million dollars inside that box. If she predicted that you'd take both boxes, then she left the opaque box empty. What is your choice?

Here's a table, representing your decision problem. (Note that here and throughout, I assume you value dollars linearly, so that $v(\$i) = i$, for any i.)

	Million	*No million*
One-box	$\$1m$	$\$0$
Two-box	$\$1m + 1k$	$\$1k$

Table 1: *Newcomb*

Causal decision theorists all agree that you should take both boxes in *Newcomb*. After all, while there's a strong correlation between your choice and the predictor's prediction, that prediction is in the past and there's nothing you can do to change it. So, taking both boxes *causes* you to be better off, no matter what the predictor predicted.

To see that the version of CDT I sketched above delivers this verdict, notice that, no matter what you choose to do, the contents of the opaque box *would* be unchanged at the most similar world at which you

[10] See also Jackson (1977), Bennett (2003), Lange (2000), Kment (2006, 2014), and Khoo (2022).

[11] See Nute (1980), Bennett (1984), Albert (2000), Loewer (2007), Maudlin (2007), and Goodman (2014) for related accounts of similarity. Ahmed (2013, 2014b) denies that CDT can be underwritten by Dorr's account of similarity. But see Dorr (2016, §7) for a reply.

[12] See, e.g., Gibbard and Harper (1978, p. 127, and pp. 160-61, n.2), Lewis (1981, p. 22, especially fn. 16), Sobel (1994, p. 42-43), and Joyce (1999, pp. 169-70).

[13] See, e.g., Williamson and Sandgren (forthcoming), Gallow (2022), Hedden (2023), and Kment (2023) for discussion of deterministic counterexamples that affect a version of CDT which makes use of Dorr's account of similarity.

chose differently, by the miracles account of similarity. Thus, taking both boxes gets you a thousand dollars more than taking one box *would*, no matter what the predictor put in the opaque box.

I won't go through the formal details of this argument, because the case is well known, and also because I'll be returning to it in §6 anyway. But the nice thing about mentioning the *Newcomb* problem now is that it illustrates a principle that's at the heart of CDT—the so-called *causal dominance principle*. According to this principle, if you're sure that one option will *cause* you to be better off than another, no matter what the world turns out to be like, then you shouldn't choose the latter option. This principle seems compelling. And it's ultimately what leads CDT to give (what I and many others think is) the right answer in *Newcomb*.

3 Deterministic Cases

CDT gets the right answer in *Newcomb*. But it gets the wrong answer in both of Ahmed's deterministic cases. In this section, I'll briefly review those cases, and spell out the answer that CDT gives in them.

One quick thing, before we get started. In both of the cases that follow, I assume there's a proposition, L, saying that some particular deterministic regularities are the (exceptionless) laws of nature. I also assume that you're almost certain this proposition is true (so, your credence in L is just a little short of 1). As we'll see, this assumption plays a special role in both of the cases to come.

3.1 *Betting on the Laws*

Here is the first case (Ahmed, 2013, 2014a):

> *Betting on the Laws.* You have a choice between two bets, and you must choose one of them. First, there's B_1, which pays \$1 if L is true, but pays nothing if L is false. Second, there's B_2, which pays nothing if L is true, but pays \$1 if L is false. You're certain that nothing you do can causally affect L's truth-value. You care only about winning the dollar.

	L	$\neg L$
B_1	\$1	\$0
B_2	\$0	\$1

Table 2: *Betting on the Laws*

Here, it seems intuitively clear that you should choose the first bet, B_1.[14] After all, you're almost certain of L's truth. And you're certain that nothing you do can causally affect its truth. Still, CDT says that it's permissible to choose the second bet, B_2. In other words, this theory says it's permissible to bet *against* your own credences.

To see why, recall the miracles account of similarity: if P is a proposition that's false at w, then the most similar P-world to w is one that matches w with respect to all matters of particular fact until shortly before P, but which does not obey w's laws. Now, with that in mind, suppose that L is actually true and you actually choose B_1. Then, happily, you win a dollar. But the miracles account says that, if you had chosen B_2 instead, you'd still have won a dollar, since the most similar B_2-world to actuality is one at which the proposition L is false.

[14] Ahmed (2013, pp. 291-92) gives a formal argument for this claim, based on a principle he calls the *causal betting principle*. I think the intuition elicited by the case is sufficient for my purposes.

Having reasoned your way to this conclusion, you should be certain of the following material conditional:[15]

$$(B_1 > \$1) \supset (B_2 > \$1).$$

The laws of probability then require that your credences satisfy this inequality:

$$cr(B_1 > \$1) \leq cr(B_2 > \$1).$$

Now we can plug these credences into CDT's equation (1):

$$\begin{aligned} U(B_1) &= cr(B_1 > \$1) \cdot 1 + cr(B_1 > \$0) \cdot 0 \\ &= cr(B_1 > \$1) \\ U(B_2) &= cr(B_2 > \$0) \cdot 0 + cr(B_2 > \$1) \cdot 1 \\ &= cr(B_2 > \$1). \end{aligned}$$

And since $cr(B_1 > \$1) \leq cr(B_2 > \$1)$, it follows that $U(B_1) \leq U(B_2)$. So, CDT says that choosing B_2 is rationally permissible. In fact, the theory says that choosing B_2 is rationally *required*, if you give any credence at all to the claim that L would be false no matter what you do.

Thus, what *Betting on the Laws* shows is that, sometimes, CDT tells you it's permissible to bet against the truth of a proposition in which you're almost certain, and whose truth-value you think is outside of your causal control. To my mind, however, no plausible decision theory ever says this. So *Betting on the Laws* is a counterexample to CDT, as we've spelled it out so far.

3.2 *Betting on the Past*

Let's now consider Ahmed's second case (2014a, 2014b). It's a bit like the case we considered at the outset:

> *Betting on the Past.* In my pocket, I have a slip of paper on which is written a proposition H. I'm going to offer you two bets, and you must choose one of them. First, there's B_3, which pays \$1 if H is true, but costs \$10 if H is false. Second, there's B_4, which pays \$10 if H is true, and costs only \$1 if H is false. Before you choose between these bets, let me tell you what H is. It's a proposition about the intrinsic state of the world at some particular time in the distant past. Furthermore, you're certain that the truth of H, together with the the truth of L, determines that you accept B_3. And you're certain that the truth of $\neg H$, together with the truth of L, determines that you accept B_4.

	H	$\neg H$
B_3	\$1	$-\$10$
B_4	\$10	$-\$1$

Table 3: *Betting on the Past*

In this case, there's a compelling argument for the claim that you should choose B_3 (Ahmed, 2014a, p. 676; 2014b, p. 127). This is: you're almost certain that if you accept that bet, then you were determined by H and L to do so, and thus you're sure to win \$1. Conversely, you're almost certain that if you accept B_4 instead, then you were determined by $\neg H$ and L to do *that*, and thus you're sure to lose \$1. What better reason could you have for choosing B_3?

[15] Cf. Ahmed (2013, pp. 294–96).

Still, CDT tells you to choose B_4, instead of B_3. The argument showing this is similar to the one we considered in the previous subsection.[16] To see how it works, first suppose that H is actually true. Then, by the miracles account of similarity, the most similar world at which you choose otherwise than you actually do is also a world at which H is true, since H is a proposition about the past. Thus, you should be certain of this material biconditional:

$$(B_3 > \$1) \equiv (B_4 > \$10).$$

By parallel reasoning, you should be certain of this material biconditional, too:

$$(B_3 > -\$10) \equiv (B_4 > -\$1).$$

The laws of probability then require that your credences satisfy these equalities:

$$\begin{aligned} cr(B_3 > \$1) &= cr(B_4 > \$10), \\ cr(B_3 > -\$10) &= cr(B_4 > -\$1). \end{aligned}$$

Now we can plug these credences into CDT's equation (1):

$$\begin{aligned} U(B_3) &= cr(B_3 > \$1) \cdot 1 + cr(B_3 > -\$10) \cdot -10 \\ &= cr(B_3 > \$1) - cr(B_3 > -\$10) \cdot 10 \\ U(B_4) &= cr(B_4 > \$10) \cdot 10 + cr(B_4 > -\$1) \cdot -1 \\ &= cr(B_3 > \$1) \cdot 10 - cr(B_3 > -\$10). \end{aligned}$$

As you'll see, $U(B_3) < U(B_4)$, no matter what your credences in the various counterfactuals. So, by CDT's causal dominance principle, it seems like you should choose B_4.

But this seems absurd. As Kment (2023, p. 7) remarks, for example, choosing B_4 seems hopelessly self-undermining: it's a bit like taking a bet on the claim that you don't accept that very bet. The upshot is that this case, too, is a counterexample to CDT, as it's currently been spelled out.

That said, not everyone agrees. Instead, some philosophers are willing to bite the bullet in *Betting on the Past*, because they think this case is relevantly similar to the *Newcomb* problem (and choosing *Two-box* in that case is something like a fixed point for causal decision theorists). As Elga (2022) says, for instance:

> In a standard Newcomb problem there is a causal dominance argument for taking two boxes: 'The $1 million is either there or it is not, and you have no causal influence on whether it is. Either way (and no matter what else is true), taking two boxes gets you a better outcome than taking just one. So you should take two boxes.'... These conditions are satisfied in [*Betting on the Past*] just as much as they are in a standard Newcomb problem. So those who are sympathetic to the spirit of causal decision theory are under some pressure to endorse [taking B_4 in *Betting on the Past*]. (p. 207)

I disagree. To me, it seems like there are important differences between *Newcomb* and *Betting on the Past*. And in the next section, I'm going to begin spelling out a theory which, I think, helps us to see those differences.

[16] Cf. Ahmed (2014a, pp. 674-75).

4 Counterfactuals, Context, and Causation

Everyone—including Elga—agrees that at least one of Ahmed's cases poses a problem for CDT. However, philosophers sympathetic to that theory are divided about how to respond. For example, some say that we should modify CDT's decision rule (Sandgren and Williamson, 2020; Williamson and Sandgren, forthcoming; Solomon, MS). Others say that we should adopt a new semantics for counterfactuals (Gallow, 2022). And others still say that we should abandon CDT, and embrace a new decision theory in its place (Hedden, 2023; Kment, 2023). For my part, I don't think any of these responses are right—but I don't have space to discuss them here. So what I'll do instead is begin spelling out the alternative response that I favor. In my view, what Ahmed's cases show isn't so much that CDT is wrong, or that we need a new semantics for counterfactuals; it's that, as we've spelled it out so far, Stalnaker-Gibbard-Harper CDT doesn't pay sufficient attention to the *context-sensitivity* of counterfactuals.

Later on, I'll say why I think this is the right response to Ahmed's cases. But first, let me say a bit more about context-sensitivity for counterfactuals in general. To start, consider these sentences from Lewis (1973b), attributed to Quine:

(3) a. If Caesar had fought in Korea, he would've used nuclear weapons.
b. If Caesar had fought in Korea, he would've used catapults.

Here, both sentences seem to have a standard interpretation. So, we should be able to use our current account of similarity—the miracles account—to pin down their respective truth-values. But unlike in other cases that we've seen, I, at least, have a hard time seeing how the miracles account is supposed to apply here. For one thing, it's not obvious just how much of history we're supposed to hold fixed when we're assessing (3-a) and (3-b).

Worse, it seems like there are contexts in which (3-a) would be true, and (3-b) would not (e.g., contexts in which present-day military technology is salient). And it seems like there are contexts in which (3-b) would be true, and (3-a) would not (e.g., contexts in which the military technology available to Caesar in his own day is salient). But again, it's difficult to see how the miracles account can deliver these verdicts on its own. Instead, it seems like special features of the context—which have nothing to do with history before the antecedent-time—have to be cited, if we're going to get the right predictions about these sentences in the contexts in which it would sound natural to utter them.[17]

Thus, ironically, Lewis's case makes trouble for the claim that the miracles account applies straightforwardly to all standard counterfactuals. Here's another case, with an even more striking upshot. This one is from Dorr (2016, p. 265), and it has a similar flavor to *Betting on the Laws*:

> Suppose that L is a simple, true, deterministic law and that Frank, a philosopher of physics, has devoted his career to defending the truth of L. He is having a public debate with Nancy, who maintains (wrongly) that there are isolated exceptions to certain generalizations that follow from L, so that L is false. [The miracles account implies that] if the circumstances of the debate had been different in any way whatsoever—for example, if someone had put a glass of water on Frank's lectern, or rudely interrupted his talk—then Nancy would have been right and Frank wrong. Thus [(4)] and [(5)] are true:

(4) If we had given Frank a glass of water, his whole career would have been devoted to a mistake.

[17] Lewis might have agreed with this. In his 1979, for example, he says that (3-a) and (3-b) are each "true under a resolution of vagueness [viz., context-sensitivity] appropriate to some contexts" (p. 457). Confusingly, however, he then immediately goes on to discuss the distinction between standard and backtracking counterfactuals. And as I say in the main text, (3-a) and (3-b) *both* seem to have a standard interpretation. For related discussion of these examples, see Kment (2006, p. 263, fn. 4) and Ichikawa (2011, pp. 292-93). In any case, even if Lewis would agree with my verdict about (3-a) and (3-b), that's already a major step towards the kind of contextualism about counterfactuals that I prefer.

(5) If you had told Frank that his whole career was devoted to a mistake, you would have been right.

As Dorr rightly says, however, both (4) and (5) seem clearly false in this context. So this case, too, looks like it makes trouble for the miracles account.[18]

As a final example—and one with an additional upshot, as we'll see—consider a case from Slote (1978), credited to Sidney Morgenbesser. Imagine I've just tossed a fair coin, which is genuinely indeterministic, and I've offered you a bet while it's spinning in the air. (For a moment, set aside our earlier assumption about the laws of nature being deterministic.) If the coin lands heads, you win $1; but if it lands tails, you lose $2. Now, suppose you decline the bet, and a moment later the coin lands heads. I say to you: "That's a shame.

(6) If you had accepted, you would've won."

Most people think that this counterfactual is true. But if it is, we must be holding fixed a specific fact about history after the time of the conditional's antecedent. And it's not obvious how to square that verdict with the miracles account, since that account says what we hold fixed when we're assessing standard counterfactuals are facts about the past.[19]

Thus, each of the examples we just looked at seems to have a similar upshot. They each seem to show that there are some contexts in which the miracles account doesn't get the right verdicts about standard counterfactuals. Instead, there are contexts in which that account looks insufficient, on its own, to pin down the truth-values for sentences which we nevertheless judge true or false. And there are other contexts in which it looks like the account gives the wrong results entirely.

The general lesson we should take away from these examples, I believe, is that counterfactuals are more sensitive to context than we earlier made them seem. While it's true that they can be sorted into the standard and backtracking categories, this doesn't exhaust the range of ways in which counterfactuals can be influenced by context. On the contrary, even if we focus on standard counterfactuals alone, it still looks like there's room for variation between contexts about the denotation of 'most similar antecedent-world'.

In philosophy of language, this fact has been widely acknowledged for some time.[20] But causal decision theorists don't seem to have paid it much attention. This is unfortunate, since the fact arguably has important consequences for what we think about Ahmed's cases. After all, in both of those cases, the miracles account played a key role in deriving the absurd recommendations. But then, if that account sometimes makes bad predictions about standard counterfactuals, it's reasonable to suspect that it's *this* that's leading CDT astray. In a case like *Betting on the Laws*, for example, you'd naïvely think that CDT would tell you to bet on the truth of the proposition L, rather than against it. And it seems like it's only because our current version of CDT relies on the miracles account of similarity that it says you should do otherwise.

This, indeed, is why I think CDT goes wrong in Ahmed's cases. So what I'm going to do now is sketch a more "contextualist" view of similarity for standard counterfactuals, and then a new version of Stalnaker-

[18] Does this case constitute an argument for Dorr's alternative account of similarity, according to which we hold the laws fixed when we're assessing standard counterfactuals? Not in my view (although see Dorr (2016, §6)). For one thing, it doesn't follow from the fact that there are *some* counterfactuals for which we naturally hold the laws fixed that *all* standard counterfactuals require us to do this. See Holguín and Teitel (MS) for further discussion.

[19] This is a bit of a simplification. But see Edgington (2003) and Kment (2006, §3) for more in-depth discussions of why cases like this one are problematic for the miracles account.

[20] Thanks here to an anonymous referee, who points out that all of the following authors reject the miracles account in favor of an account of similarity that's more "contextualist" (note, however, that none of these accounts are exactly like the one I'll give below, nor do they ultimately get applied to CDT): Stalnaker (1968, 1981a, 1984, 2021), Ichikawa (2011), Lewis (2015), Ippolito (2016), Steele and Sandgren (2020). Additionally, even some philosophers who are broadly sympathetic to the miracles account—like Kment (2006) and Khoo (2022)—are critical of the version I gave in §2, and opt for a more contextualist version instead. For additional criticisms of the miracles account, different to the ones I've given here, see Elga (2001) and Holguín and Teitel (MS).

Gibbard-Harper CDT to go along with it. First, however, let me note one last thing about the final example we looked at. This is that there's a natural explanation for *why* we hold the outcome of the coin flip fixed in our assessment of (6)—namely, that this outcome is *causally independent* of whether or not you accept the bet.[21] This explanation looks especially plausible when we contrast the sentence (6) with the following sentence:

(7) If I had flipped a different (fair, indeterministic) coin, you would've won the bet.

Unlike (6), most people think that this counterfactual is *not* true. And the most natural explanation for why is that, while in the first case your choice to accept or decline the bet is causally independent of the coin flip's outcome, in the second case my choice of which coin to flip is not causally independent of the outcome. In other words, in (7), but not in (6), there's a causal chain running from the event described by the counterfactual's antecedent to the event described by its consequent. And this is why we think (6) is true, but (7) is not.

Lewis, whose account of similarity we've been working with up until now, wouldn't have accepted this explanation, because he believed that causation could be *analyzed* as a relation of counterfactual dependence between distinct events.[22] (Incidentally, that's why causal notions were nowhere mentioned when I was sketching the miracles account initially, despite the fact that I said standard counterfactuals often tell us about *causal* effects of the antecedent on the consequent.) However, I think examples like this one show strongly that a counterfactual analysis of causation can't succeed. So in what follows, I'm going to assume that causal notions *can* inform the truth-conditions for counterfactuals. Doing so makes available to us some resources that we didn't have before. And as we'll see, the idea that causal notions can influence counterfactuals plays an important role in how I spell out my theory.

4.1 Questions in Context

The examples we just looked at show that similarity relations for standard counterfactuals depend, not just on facts about the world's history before the antecedent-time, but also on more "local" matters, like features of a context that happen to be salient. In the Lewis/Quine case, for instance, the sentence (3-a) would plausibly be true in a context in which the existence of nuclear weapons is salient, but not in a context in which it's not salient. Thus, an adequate account of similarity for standard counterfactuals should give more weight to these features of context than the account we previously had. And what I'm going to do now is spell out one way in which I think this can be accomplished. The account of similarity I'll sketch below isn't so much a *theory* of similarity for standard counterfactuals, as it is a set of constraints which I think any plausible such theory should satisfy. But as we'll see, even this rough-and-ready account of similarity is sufficient to get the right answers in tricky cases like the ones we've seen.[23]

To start off, note that in fields like semantics and philosophy of language, it's common to think of salient features of a context as being represented by salient *questions*.[24] These questions *foreground* the issues that are "live" in the context, and they *background* the issues that aren't live (Yalcin, 2016, p. 30).

[21] This observation is also made by Bennett (2003, chapter 15), Edgington (2003), and Kment (2006, 2014), among others. See those works for further discussion, as well as for related examples.

[22] See Lewis (1973a, 1986, 2000).

[23] The contextualist view of similarity I sketch in this section has a lot in common with views espoused by, e.g., Kaufmann (2004), Ippolito (2016), Khoo (2016), Boylan and Schultheis (2021), and Dorr and Hawthorne (MS). It also has something in common with so-called *causal modelling* approaches to counterfactuals, like those of Hiddleston (2005), Santorio (2019), Gallow (2022), or Khoo (2022). See also Joyce (2009).

[24] The *locus classicus* for this view is Roberts (2012). However, there's an important difference between the way I'm understanding the notion of a salient question—or a *question under discussion*, as Roberts calls it—and the way Roberts herself does. In particular, I'm not going to assume that these questions are always *unanswered* in a context. See Boylan and Schultheis (2021) for a similar understading of salient questions.

In the Slote/Morgenbesser case, for instance, we can think of the salient issue as being represented by the question *How did the coin land?*. Then, contributions to the conversation are deemed relevant, or appropriate, just in case they address that salient question. (Note, however, that we don't have to assume this question is ever explicitly spoken in the context; it may be merely implicit.)

To make this idea precise, let's introduce some formalism. Following Stalnaker (1978), let's first say that any context can be modelled by a set of worlds, $\mathcal{W}$, which we call the *context set*. For simplicity, I'll assume that $\mathcal{W}$ is always a finite set of worlds. And intuitively, we can think of it as the set consisting of all worlds that count as "live options" in the context, for the purposes at hand. For instance, in an ordinary conversational context, $\mathcal{W}$ might consist of all the worlds that you and other conversational participants believe could be actual. And in a deliberational context, $\mathcal{W}$ might consist just of your epistemically possible worlds.

Now, a salient question can be thought of as a *partition* of the context set.[25] Each cell of this partition groups together worlds that are alike with respect to a complete answer to the question. And any union of these cells corresponds to a partial answer. In the Slote/Morgenbesser case, for example, the partition is just the set which groups together *Heads*-worlds and *Tails*-worlds, respectively. Similarly, in Dorr's "Frank vs. Nancy" case, where the question *Is L a law of nature?* is salient, the associated partition consists of worlds that obey the L-law, and worlds that don't, respectively.

Questions like this give us a way of constraining the similarity relations that are appropriate in a context—or *admissible*, as I'll often say. Specifically, when there's some feature of the context that's especially salient, we can think of an admissible similarity relation as being one that "holds fixed" the answers to a corresponding question. To see what this means, let $\mathcal{W}$ again be a context set, and let $Z = \{Z_1, ..., Z_n\}$ be a partition of $\mathcal{W}$, corresponding to such a question. Then, in my view, a similarity relation is admissible in this context only if its associated selection function satisfies the following constraint: for each world $w \in \mathcal{W}$, if $w \in Z_i$, then $f(P, w) \in Z_i$, where P is the antecedent of the counterfactual of interest. In words: a similarity relation is admissible in a context only if it says that the most similar P-world to w is one that lies in the same cell of the salient partition as w itself. (See Figure 1 for an illustration.)

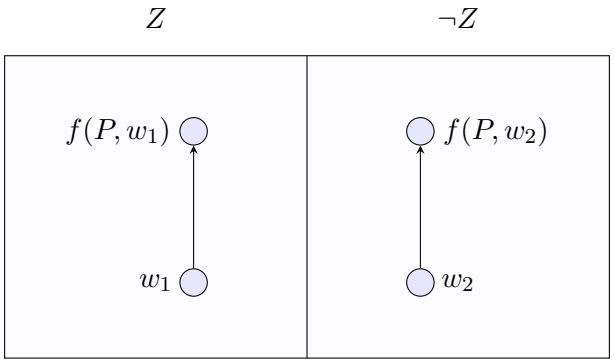

Figure 1: An Admissible Similarity Relation.

We can make this idea clearer by using a concrete example. So, consider again the Slote/Morgenbesser case, and particularly the sentence (6):

(6) If you had accepted, you would've won.

[25] See Groenendijk and Stokhof (1984), and also Hamblin (1973). A different, but equally plausible, way to think about this partition is in terms of *subject matters*. See Lewis (1988a, 1988b). I'll stick with the notion of questions in the main text to streamline the discussion.

As I said before, the salient question here is *How did the coin land?*. And the associated partition consists of *Heads*-worlds and *Tails*-worlds, respectively. Now, at the actual world, we know that we're in the *Heads*-cell of this partition, since the coin actually landed on heads. And we know, further, that the actual world is one at which you didn't accept the bet. Thus, according to the account of similarity I'm sketching, the most similar world at which you *do* accept the bet is also a world in the *Heads*-cell of this partition, on any admissible similarity relation. The upshot is that the sentence (6) comes out true at the actual world. And this is the result for which we were hoping.

Already, then, this broad-brush account of similarity for standard counterfactuals gets a case right, which the miracles account got wrong. There are a few interesting things to note about it. First, it's consistent with the view as I've spelled it out so far that there can be more than one admissible similarity relation in a context. After all, if a similarity relation is admissible just in case it holds fixed the answers to a salient question, then in general there will be many such relations that can do this job. Thus, the account of similarity I'm sketching allows for some *indeterminacy* in the interpretation of counterfactuals. Specifically, when there's more than one similarity relation in play in a context, it will be indeterminate which world is picked out by the phrase 'most similar P-world'. When that's so, a sentence like 'If P, would Q' can be indeterminate in the context, since there will be some admissible similarity relations which make it true, and others which make it false. This fact will be important later. But for now, just note that it's in keeping with the idea that questions represent the distinctions we're interested in making in a context. In other words, since questions group together worlds according to aspects of similarity that are contextually salient, pinning down a similarity relation more precisely than this might give us more information than is relevant.

Additionally, note that nothing I've said so far rules out there being more than one salient question in a context. (We can, however, always assume there's *at least* one salient question in every context, since the trivial question, $\mathcal{W}$ itself, always counts as salient.) In particular, if one question *contains* another—in the sense that every cell of the partition corresponding to the second question is a union of cells of the partition corresponding to the first—then, if the more fine-grained question is salient in a context, the more coarse-grained question will be, too. I'll call the first question here a *refinement* of the second, and the second a *coarsening* of the first.

In cases like this, we can usually think of similarity relations as being constrained by the *most* fine-grained question in a context. This goes also when we have two (or more) such questions, and neither is more fine-grained than the other. For example, suppose that in the Slote/Morgenbesser case, we were interested in the question *How did the coin land?*, but also in the question *Is the coin a nickel or a dime?*. Then, in that case, we could think of the similarity relations for counterfactuals as being constrained, not by either of these questions alone, but by their *conjunction*—or, more precisely, their *coarsest common refinement*. This is the partition each of whose cells corresponds to an intersection of cells from the first question and cells from the second question. For instance, the coarsest common refinement of *How did the coin land?* and *Is it a nickel or a dime?* is the partition: {*Heads and Nickel, Heads and Dime, Tails and Nickel, Tails and Dime*}.

This is useful to know about, but it won't play much of a role in what's to come. There is, however, another notion which will turn out to be important. In some contexts where there's more than one salient question in play, it's appropriate to think of similarity relations as being constrained, not by their coarsest common refinement, but by their *finest common coarsening*. This is the most fine-grained question that's coarser than both of the questions we started with. For instance, the finest common coarsening of *How did the coin land?* and *Is it a nickel or a dime?* is just the trivial partition, $\mathcal{W}$, since the only "question" that's coarser than both {*Heads, Tails*} and {*Nickel, Dime*} is the context set itself. The notion of a finest common coarsening of questions is a bit like the *disjunction* of those questions.[26] But it's a tricky notion

[26]Arguably, it's not exactly like the disjunction, however. One reason is that the disjunction (union) of partitions need not be

to get your head around. So I'll defer further discussion of it until it's needed.

In the meantime, let me note one other constraint that needs to be imposed on questions-partitions, if my account of similarity is going to work as intended. It's important that this account not make counterfactuals vacuously true in a context. But for all I've said so far, nothing rules out this being the case: it may be that $Z = \{Z_1, ..., Z_n\}$ is the salient partition, but $P \cap Z_i = \varnothing$, for some Z_i and counterfactual antecedent P. In cases like this, a counterfactual beginning with P will be vacuously true at any world $w \in Z_i$, since there simply won't be any P-worlds in that cell in the first place. Iin general, however, when we assess counterfactual sentences in a context, we try to do so in ways that don't make them vacuously true. Thus, to capture this idea, I'm going to assume that, if P is a counterfactual's antecedent, then any suitable partition that could constrain relations of similarity for this counterfactual satisfies $P \cap Z_i \neq \varnothing$, for each Z_i. Khoo (2016) calls this the *well-definedness* constraint on question-partitions, and here I'll follow suit.

Now, if you go back and check the examples we've looked at so far, you'll see that it's often easy to spot a candidate question, and that, when similarity relations are constrained by this question, we get the right verdicts about the counterfactuals. In Dorr's case, for instance, the salient question corresponds to the partition $\{L, \neg L\}$. Then, since we're told that the actual world lies in the L-cell of this partition, the sentences (4) and (5) both come out false in this context. This, as I noted before, is the result that we were after.

But there's still one important way in which my account needs to be refined. To see what it is, consider again the sentence (7):

(7) If I had flipped a different (fair, indeterministic) coin, you would've won the bet.

Suppose we analyzed this sentence in the same way we analyzed (6). That is, suppose we took the salient question to be *How did the coin land?* And suppose we took the corresponding partition to consist just of *Heads*-worlds and *Tails*-worlds. Then, since the actual world lies in the *Heads*-cell of this partition, it seems like any admissible similarity relation will say that the most similar world to actuality at which I flipped a different coin is a world at which you win the bet. The sentence (7) thus comes out true. But earlier, I said this sentence is *not* true.

The reason our analysis goes wrong here—as I tried to stress before—is that, in the case of (7), the coin flip's outcome is not causally independent of the counterfactual's antecedent. This immediately suggests one final constraint we need to impose on question-partitions. If a partition is going to constrain the admissible similarity relations for standard counterfactuals in a context, then it has to consist only of propositions that are *causally independent* of the antecedents. Without that constraint, our account will make bad predictions in cases like the one we've just seen.

With this constraint, however, the account gives plausible verdicts. The constraint also gives us a way of characterizing question-partitions for counterfacuals more generally—at least when the antecedents of those counterfactuals are about nomically possible, dated events. Often, we can think of the propositions in these partitions as being ones that describe salient "causal background factors", with which a counterfactual's antecedent would combine to *bring about* the consequent. For example, in the Slote/Morgenbesser case, every contextually-relevant world at which the coin lands heads is one at which taking my bet *causes* you to win \$1; and every contextually-relevant world at which the coin lands tails is one at which taking the bet causes you to lose \$2. Thus, how the coin lands is the only contextually-relevant background factor in this case. And that's why it makes sense to think of the relevant question-partition as holding

a partition. And if we think of questions as being modelled by partitions, then the disjunction of two questions might not itself be a well-formed question. This is plausibly related to the fact that, in natural language, disjunctions of questions often strike us as infelicitious. (For example: 'Did the coin land heads or tails, or is it a nickel or a dime?'.) As we'll see later on, I think some of this applies to Ahmed's *Betting on the Past* case.

only this background factor fixed. After all, as we heard before, standard counterfactuals often tell us about the *causal* effects of the antecedent on the consequent. So what we hold fixed when we're assessing these counterfactuals is often a contextually salient causal background, against which the counterfactual's antecedent takes place.

There's more to be said about this in general. But thankfully, in all the cases we'll consider here, it'll be straightforward to see what this causal background consists in.

4.2 Contextualist CDT: A First Pass

The account of similarity for standard counterfactuals I just sketched is more flexible than the miracles account with which we started. And this should give us some hope that a version of Stalnaker-Gibbard-Harper CDT, equipped with this account of similarity, can avoid the problems posed by Ahmed's cases. That said, there are a few things we still need to figure out. One of them is how we're supposed to think about the notion of a "salient question" in a context where you're making a decision, rather than having a conversation. This isn't yet obvious. But thanfully, it turns that there's a very natural way to think about these questions in decision-making contexts. Daniel Hoek (2019, 2022) has recently investigated this idea at length, and what I'll say here is broadly in line with his suggestions.[27] To see how the idea works, let's go back to the *Newcomb* problem:

	Million	*No million*
One-box	$\$1m$	$\$0$
Two-box	$\$1m + 1k$	$\$1k$

Table 1: *Newcomb*

Here, the rows of the table correspond to your options, and the columns correspond to propositions about "states of the world", which (by your lights) may or may not obtain. The conjunction of any option with any state-proposition entails a unique outcome—that's why we can represent the decision problem using a table like this one. Notice, however, that if this representation is going to work, then it's important that the state-propositions form a partition. Thus, since we're thinking of questions as corresponding to partitions, there's a very natural candidate for the salient question in *Newcomb*. This is the question *How much money is in the opaque box?*, corresponding to the partition $\{$*Million*, *No million*$\}$.[28]

Similar things can be said about the other decision problems we've looked at. That is, in each of the cases we've seen, there's a partition of states given in the corresponding decision table, with the features that (i) each of your options is consistent with each cell of this partition, and (ii) the conjunction of any state-proposition with any one of your options entails a unique outcome. Thus, it's natural to think of these partitions, too, as corresponding to salient questions. And in fact, whenever we have a partition of states in a decision problem with the features (i) and (ii), it seems—prima facie—like we can take that partition to correspond to a salient question. Then, we can use these partitions to fix the admissible similarity relations, in a way analogous to the way we saw before. That is, in parallel to what I said in the last subsection, a

[27] I should note, however, that there are a few important differences between my theory and the one that Hoek develops. For instance, Hoek (2019) seems to think that question-partitions are induced by similarity relations, rather than constraining those relations. Also, Hoek (2019, 2022) doesn't discuss the role of context in making certain questions salient. And most importantly, my theory allows that there can be multiple, competing questions "raised" in a decision-making situation, which is something that Hoek doesn't discuss. As we'll see, this fact also necessitates the distinctive formal apparatus that I introduce in §5. (This is also one of the ways my theory differs from the view of Stalnaker (MS).

[28] Actually, there are really two salient questions in *Newcomb* (and similarly for other decision problems). The second question corresponds to the partition of your options. However, because your "answer" to this question depends on what you believe about the other question, we can generally take the partition of states to be the *most* salient question in a decision problem.

similarity relation is admissible in a context only if it says that the most similar world to w at which you choose some particular option lies in the same cell of the state-partition as w itself.

This works whenever the state-partition consists of propositions that are causally independent of your options. But then again, not every decision problem has this feature. To illustrate, consider a well-known case from Joyce (1999). Imagine that you've parked your car in a seedy neighborhood, when a man approaches you and offers to "protect" your car for the low fee of \$10. You know that people who don't pay the fee invariably come back to find their windshields smashed. And you know that any repairs to your windshield would cost you \$100. Now, the natural way to represent your decision problem is as follows:

	Smashed windshield	*Unsmashed windshield*
Pay	−\$110	−\$10
Don't pay	−\$100	\$0

Table 4: *The Shakedown*

But the salient question constraining the similarity relations in this context can't be the one corresponding to the partition {*Smashed windshield, Unsmashed windshield*}. If it were, then every admissible similarity relation would say that you do better by not paying than by paying. But that seems absurd: by not paying, you *cause* the man to smash your windshield. And by paying, you cause him to leave your windshield alone.

As before, then, we need to assume that any partition that constrains similarity relations for standard counterfactuals in a decision-making context is one which specifies only propositions whose truth-values are causally independent of your options. Sometimes, this will mean that the partitions which constrain these relations don't match up with the corresponding decision tables. There is, however, a general way in which we can think about these partitions, analogous to what I said before. This is: usually, we can think of them as specifying salient "causal background factors" with which an option combines to cause a specific outcome.[29] In Joyce's case, for example, this partition might consist of the propositions *The man is a villain* and *The man is not a villain*, since, here, the man's temperament is the only salient causal background factor determining what the outcome of your options would be. Similarly, in the *Newcomb* problem, the only relevant causal background factor is whether or not the million dollars is in the opaque box. At every contextually relevant world, once you know whether or not the money's in that box, you know everything you need to know about what your choice of an option would cause.

Now, given this way of thinking about question-partitions in decision-making contexts, there's a generic, English language gloss we can give of these questions. This is: *How do the things I care about—viz., outcomes—depend causally on what I do?*. When it's put in these terms, the question might remind you of something. Both Skyrms (1980) and Lewis (1981) give versions of CDT which appeal to propositions called *causal dependency hypotheses*. In Skyrms's words, these are "maximally specific specifications of the factors outside our [causal] influence at the time of decision, which are causally relevant to the outcomes of our actions" (p. 133).[30] This sounds pretty similar to the thing I'm now proposing.

Broadly speaking, this is right. But there are a few crucial differences between "dependency hypotheses", as I'm thinking about them, and the way that Lewis and Skyrms do. First and most obviously, Lewis thinks these propositions hold in virtue of patterns of counterfactual dependence, since his view is that causation just *is* a relation of counterfactual dependence between distinct events. Earlier, however, I said

[29] If the laws are indeterministic, then we might need to say something more general here. After all, in that sort of case, even if we hold fixed the complete past and laws of nature—both of which are causally independent of your options—your choice might nevertheless only be sufficient to causally determine the *chance* of some outcome, rather than the outcome itself. Since I'm assuming determinism here, however, I'll ignore this possibility.

[30] Lewis (1981, p. 11) gives a gloss of 'dependency hypotheses' that's even more similar to the one I just gave.

that my view is that (standard) counterfactuals often hold in virtue of causal relations. So there's a sense in which Lewis and I are approaching things from opposite directions. Whereas he thinks that counterfactuals come before dependency hypotheses in the order of explanation, I think that the reverse is true.

More importantly, unlike both Lewis and Skyrms, I'm not requiring dependency hypotheses to be *maximally specific* propositions. On the contrary, the view that they *are* maximally specific propositions looks untenable, if the laws of nature are deterministic. As Hedden (2023) points out, for example:

> something that doesn't causally depend on which of your present actions you perform can nonetheless entail which one you do. This means that if dependency hypotheses can specify anything that doesn't causally depend on your present action [like history and the laws of nature], then we'll have some dependency hypotheses which are inconsistent with some of your available actions, resulting in actions with undefined [utility]. (p. 744)

The upshot is that, if the laws of nature are deterministic, the theories of Lewis and Skyrms will simply fall silent in certain decision problems. This seems like an even worse problem than the ones we encountered in §3.

But like I said, I'm not requiring "dependency hypotheses" to be maximally specific propositions. All I'm requiring is that they specify salient causal background factors which, by your lights, are sufficient to cause outcomes, in conjunction with your choice. Which factors those are is a context-sensitive matter. And it's this sensitivity to context that allows me to avoid the problems faced by Lewis and Skyrms.

4.3 Loose Ends

Having now sketched most of the background for my theory, you can probably already tell how it's going to work in particular cases. Unfortunately, however, we're not yet in the clear. To see why, consider again the *Betting on the Past* case. Recall that the decision table I used to represent that decision problem was:

	H	$\neg H$
B_3	\$1	$-$\$10
B_4	\$10	$-$\$1

Table 3: *Betting on the Past*

And here, it's clear that the partition $\{H, \neg H\}$ corresponds to a salient question, in the sense I defined above. But notice: given what you know about the salient proposition L in *Betting on the Past*—namely, that it determines you choose B_3 in conjunction with H, and determines you choose B_4 in conjunction with $\neg H$—the following table also seems like a good representation of your decision problem.

	L	$H \wedge \neg L$	$\neg H \wedge \neg L$
B_3	\$1	\$1	$-$\$10
B_4	$-$\$1	\$10	$-$\$1

Table 5: *Betting on the Past*, Version 2

Indeed, even Ahmed acknowledges this. In his 2014a, for example, he says that "[Table 5] is just as accurate as [Table 3] when it comes to representing [your] situation. It represents the same payoffs to the same actions in the same circumstances at all the possible worlds where this could matter to a causalist" (p.

677).[31] If this *is* an adequate representation of your decision situation, however, then it looks like the partition $\{L, H \wedge \neg L, \neg H \wedge \neg L\}$ also counts as a salient question. And in *this* case, it's clear that the causal dominance argument for B_4 doesn't go through. Indeed, if utility is calculated in line with Table 5, then CDT will recommend B_3.

This is peculiar. What we seem to have is a case in which there's more than one salient question raised by your decision situation—something I earlier said was possible. But problematically, depending on which of the questions we focus on, CDT gives different recommendations about what you should do.[32]

Some philosophers will respond to this by saying that *Betting on the Past* isn't a well-posed decision problem. Others will say that there isn't a univocal answer about which option you should choose, since CDT makes different recommendations depending on how the problem is represented. I won't pursue either of those responses (although I have some sympathy with the latter). My chief reservation about them is just that, in *Betting on the Past*, I have very clear intuitions about which option it's rational to choose: as I said before, choosing B_4 seems hopelessly self-undermining.

Besides, there's a more general problem looming here. To see what it is, recall from earlier that I said it's consistent with the contextualist view about similarity that I like that context can sometimes underdetermine which world counts as "most similar". This is a general feature of contextualist views about counterfactuals. But what *Betting on the Past* shows, I think, is that, when there's more than one question that's salient in a context, this kind of indeterminacy needn't be inert. Instead, it will occasionally lead CDT to give conflicting recommendations, depending on how 'most similar P-world' is precisified.

Thus, before I can state my own version of CDT completely, we need to find a way of handling this kind of indeterminacy. That's the task of the next section. And it's that task to which we now turn.

5 Accommodating Indeterminacy

Let's go back to Stalnaker's semantics. Recall that, when I introduced that semantics initially, I appealed to the notion of a selection function: a function from propositions and worlds to possible worlds. Now, it turns out that if we assume selection functions satisfy some natural constraints, then Stalnaker's semantics can be specified in a slightly different way. To see this, let me first state the constraints I have in mind.

[31] Actually, the version of the table that Ahmed considers in his 2014a is the following, since in that paper he assumes, not just that you're highly confident of L, but that you're certain of it.

	L	$\neg L$
B_3	\$1	−\$10
B_4	−\$1	\$10

My Table 5 is a bit more complicated, because, in the version of *Betting on the Past* I've given here, you give a tiny amount of credence to the possibility that L is false. Thus, for you, there some worlds which get positive where you choose B_4 and H is true. And there are some worlds that get positive credence where you choose B_3 and $\neg H$ is true. This is why I've fine-grained the $\neg L$-column in Ahmed's table.

[32] Now's a good time to note that we can't simply focus on the coarsest common refinement of the two partitions in this case either. After all, this coarest common refinement corresponds to the following partition:

	$H \wedge L$	$H \wedge \neg L$	$\neg H \wedge L$	$\neg H \wedge \neg L$
B_3	\$1	\$1	∅	−\$10
B_4	∅	\$10	−\$1	−\$1

This fails my well-definedness condition on question-partitions. See Joyce (2016), Solomon (2021), Elga (2022), Hedden (2023), and Fusco (forthcoming) for further discussion of this table. You can probably see now why the notion of a finest common coarsening of questions is going to be important.

They are:

(i) **Success**. $f(P, w) \in P$.

(ii) **Strong Centering**. If $w \in P$, then $f(P, w) = w$.

(iii) **Reciprocity**. If $f(P, w) \in Q$ and $f(Q, w) \in P$, then $f(P, w) = f(Q, w)$.

(iv) **Accessibility**. If $w \in \mathcal{W}$, then $f(P, w) \in \mathcal{W}$.[33]

(Remember: $\mathcal{W}$ here is the context set.) Each of these constraints is very plausible. For example, Success just says that the most similar P-world to w should *be* a P-world—and that seems obviously right. Strong Centering says that, if $w \in P$, then w should count as the most similar P-world to itself—and that, too, seems right. Reciprocity is needed to validate a host of compelling inference patterns involving counterfactuals. And Accessibility says that selection functions shouldn't "reach outside" the set of worlds that are relevant in the context. A little reflection shows, additionally, that the first three constraints in particular are needed if selection functions are going to track anything like a *similarity* relation between possible worlds. And that, of course, is what we're after here.

Now, it turns out that the constraints (i)–(iv) above suffice to ensure that selection functions totally order the worlds in $\mathcal{W}$. That is, given the choice of a "base world", w, selection functions "rank" the worlds in $\mathcal{W}$ according to how similar they are to w, with w always counting as the most similar world to itself.[34] The upshot is that, given a selection function and choice of base world w, there corresponds a *sequence* of possible worlds, which orders the worlds in $\mathcal{W}$ according to how similar they are to w. Conversely, for every sequence of worlds in $\mathcal{W}$, there corresponds a selection function-world pair, $\langle f, w\rangle$, such that w is the first world in the given sequence. What this means is that everything we could do before, using a selection function, we can now do using a sequence.

In particular, we can give a slightly different definition of Stalnaker's semantics (as I previously said).[35] To do so, let's first introduce some terminology. From here on out, let's say that a "factual" (i.e., non-conditional) proposition P is *true at a sequence*, s, just in case P is true at the *first* world in s. Then, let's say that a counterfactual $P > Q$ is true at s just in case Q is true at the first P-world in s.[36] Finally, let's say that $P > Q$ is true at a world, w, simpliciter just in case it's true at every (admissible) sequence whose first world is w. This, then, is our new definition of Stalnaker's semantics. (I'll return to the topic of admissibility in a moment.)

As an example, to make what I've just said bit more concrete, suppose that the set of worlds we're interested in is $\mathcal{W} = \{w_1, w_2, w_3\}$. Let $\mathcal{S}_\mathcal{W}$ be the set of all the sequences of worlds that we can generate from $\mathcal{W}$, namely:

$$\mathcal{S}_\mathcal{W} = \left\{ \begin{array}{l} \langle w_1, w_2, w_3\rangle, \langle w_1, w_3, w_2\rangle, \\ \langle w_2, w_1, w_3\rangle, \langle w_2, w_3, w_1\rangle, \\ \langle w_3, w_1, w_2\rangle, \langle w_3, w_2, w_1\rangle \end{array} \right\}.$$

[33] The fact that I'm introducing the Accessibility constraint here might surprise you, since this constraint is usually taken to characterize *indicative* conditionals, rather than counterfactuals (see, e.g., Stalnaker (1975)). I'll say more about why I'm introducing accessibility below. In particular, see fn. 40.

[34] To see how this works, first suppose that we have a selection function f and a base world w_1. Then, we can construct a *sequence* of possible worlds, $s = \langle w_1, ..., w_n\rangle$, corresponding to this selection function-world pair as follows. For any worlds w_i and w_j, let w_i come before w_j in the sequence just in case $f(\{w_i, w_j\}, w) = w_i$. Conversely, given s, we can construct a selection function as follows. Let $f(\{w_i, w_j\}, w) = w_i$ whenever w_i comes before w_j in s. The rest of f can then be derived from the constraints (i)–(iv) in the main text. Cf. Mandelkern (2018) and Khoo (2022).

[35] This isn't 100% accurate. As Matthew Mandelkern points out to me, the sequence formulation of Stalnaker's semantics requires a very mild strengthening of his background logic. However, this strengthening has no bearing on anything I'll say here, so it needn't concern us. See Mandelkern (forthcoming, §7.4) for further discussion.

[36] This definition only works for *simple* counterfactuals, i.e., those that don't have counterfactuals (or other modals) as antecedents or consequents. All the counterfactuals I'm interested in here, however, count as "simple" in this sense.

Now suppose that P is a factual proposition true at the worlds w_1 and w_2, and Q is a factual proposition true at w_2 and w_3. Then, P is true at the first four sequences in $\mathcal{S}_{\mathcal{W}}$, as I've written it above; Q is true at the last four sequences; and $P > Q$ is true at the following sequences, since these are the only sequences whose first P-world is a Q-world: $\langle w_2, w_1, w_3\rangle$, $\langle w_2, w_3, w_1\rangle$, and $\langle w_3, w_2, w_1\rangle$. (Note also that, while $P > Q$ is true at the world w_2 simpliciter, it's neither nor true nor false at the world w_3, since there's one sequence beginning with w_3 whose first P-world is a Q-world, and there's one sequence beginning with w_3 whose first P-world is not a Q-world.)

Why, however, am I bothering to introduce this new formulation of Stalnaker's semantics, when the previous formulation seemed perfectly adequate? There are two key reasons. The first is simply that the constraints on selection functions that I gave above are all completely standard. Stalnaker himself assumes them, for example (1968). And so does nearly everyone who's worked with his semantics in the meantime. Thus, by appealing directly to sequences, rather than selection functions, we can forgo the need to keep mentioning the constraints (i)–(iv). They're built right into the sequence formulation of the semantics; they're not extra assumptions that we need to make.

The more important reason, however—as you're probably expecting—is that the sequence-based formulation of Stalnaker's semantics helps us to handle the issue of indeterminacy, which I mentioned at the end of the last section. To begin to see how, start by taking another look at the toy example, where $\mathcal{W} = \{w_1, w_2, w_3\}$. In that case, each world is consistent with two different similarity orderings. So, what this implies is that, even after we've pinned down truth-values for all the factual propositions at a world, we still haven't pinned down the truth-values for all the conditional propositions. Here's a picture, for illustration:

w_1		w_2		w_3	
w_1	w_1	w_2	w_2	w_3	w_3
w_2	w_3	w_1	w_3	w_1	w_2
w_3	w_2	w_3	w_1	w_2	w_1

Figure 2: Logical Space in the Toy Example

This is an idea worth dwelling on a bit. You'll notice that, in the sequence-based set-up I've just introduced, worlds no longer count as the most basic possibilities. Instead, sequences do. And we can think of worlds as sets of sequences, just as we could think, before, of propositions as sets of worlds. The idea is that, while all the "descriptive" facts are settled by the world, the conditional facts need not be. Instead, conditional facts depend for their truth on relations of similarity *between* worlds. Moreover, those relations are fixed by context; they needn't supervene on the descriptive facts that obtain at a world.

In the recent literature, this "fine-grained" view of a conditional's content has gained in popularity. Several authors have shown, for example, that it helps us to defuse puzzles arising from the interaction between our credences, on the one hand, and conditionals, on the other (see, e.g., Khoo and Santorio (2018), Goldstein and Santorio (2021), Khoo (2022), Schultheis (forthcoming), and Mandelkern (forthcoming)). Later in this section, I'll briefly mention one of those puzzles, and allude to how the fine-grained view helps to resolve it. But in the meantime, note that, since we're now working in a more fine-grained setting, we have to say how your credence function, cr, can be extended, so that it's defined over sequences, and

not just over worlds.[37]

There are a number of ways we could make this extension, each with advantages and disadvantages.[38] But for simplicity, I'm here going to work with an idea from Goldstein and Santorio (2021) and Khoo (2022).[39] Specifically, I'll "lift" your credence function, cr, to a new credence function, pr, by means of a recursive procedure. This new function will then allow you to assign credences to arbitrary sets of sequences, and not just to sets of worlds.

To see how this works, let's start with some assumptions. First, let's suppose that the context set, $\mathcal{W}$, is just the set of your epistemically possible worlds. Then, let's assume that your credence function is *regular*, in the sense that, for every world $w \in \mathcal{W}$, your credence in w is such that $cr(w) > 0$. Neither of these assumptions is strictly essential for what I'm doing. But dropping them introduces additional complications which I'd rather not get into.

Now, in the simplest case, where all sequences of worlds count as admissible in the context, we can "lift" your credence function as in the following way (I'll give a slight refinement of this definition in a moment). First, let's write '$[w]$' for the set of sequences beginning with the world w, and '$[w_1, ..., w_k]$' for the set of sequences whose k-length initial segment consists of $w_1, ..., w_k$, in that order. Then, we define the credence function pr as:

(i) $pr([w]) = cr(w)$,

(ii) $pr([w_1, ..., w_k]) = pr([w_1, ..., w_{k-1}]) \cdot cr(w_k \mid \mathcal{W} - \{w_1, ..., w_{k-1}\})$.

Metaphorically, we can think of this as saying that your credence in a sequence $s = \langle w_1, ..., w_n \rangle$ is equal to your credence that you'd draw those worlds from an urn, in that order, and without replacement. For example, the credence that you assign to the sequence $\langle w_1, w_2, w_3 \rangle$ from the toy example is just your credence that you'd draw w_1 first, multiplied by your credence that you'd draw w_2 second, having already drawn w_1, and so on. Since each sequence of worlds is supposed to correspond to an ordering of possible worlds according to how similar they are to a base world, it's easy to see why this lifting procedure is a sensible proposal.

It's also easy to see that pr preserves the credences that cr assigns to factual propositions. To quickly illustrate this anyway, however, consider again the toy example, where $\mathcal{W} = \{w_1, w_2, w_3\}$. Suppose that $cr(w_1) = cr(w_2) = cr(w_3) = 1/3$. Then, it follows that $cr(P) = 2/3$, since $P = \{w_1, w_2\}$. Now, by the definition of pr:

$$\begin{aligned} pr(\langle w_1, w_2, w_3 \rangle) &= pr([w_1, w_2]) \cdot cr(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= pr([w_1]) \cdot cr(w_2 \mid \mathcal{W} - \{w_1\}) \cdot cr(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= cr(w_1) \cdot cr(w_2 \mid \mathcal{W} - \{w_1\}) \cdot cr(w_3 \mid \mathcal{W} - \{w_1, w_2\}) \\ &= 1/3 \cdot 1/2 \cdot 1 \\ &= 1/6. \end{aligned}$$

Similar calculations show that $pr(\langle w_1, w_3, w_2 \rangle) = pr(\langle w_2, w_1, w_3 \rangle) = pr(\langle w_2, w_3, w_1 \rangle) = 1/6$. And taking the sum of your credences in all of these sequences gives $pr(P) = 2/3$, as desired. (Note that your credence in a counterfactual $P > Q$ is also the sum of your credences in all of the sequences at which it's true. But in the present setting, this set of sequences need not always correspond to a set of worlds.)

[37] Do we also need to say how your value function, v, can be extended, so that it's defined over sequences? Not for present purposes, since I'll assume that outcomes are ordinary "factual" propositions. I explore this issue elsewhere, however. See McNamara (MS).

[38] See van Fraassen (1976) and Mandelkern (forthcoming) for proposals different to the one I'll make use of here.

[39] See also Khoo and Santorio (2018).

Now, there's one last piece of the puzzle I need to put in place, before I can state my contextualist version of CDT precisely. Specifically, I need to say how things work out when not all of the possible similarity orderings are admissible in a context. After all, in §4 we heard that a similarity ordering is admissible only if it holds fixed the answers to a salient question. So, how are you supposed to assign credences to sets of sequences, given this constraint?

This turns out to be straightforward. For example, suppose we have a partition, $Z = \{Z_1, ..., Z_n\}$, corresponding to such a question. Then, a sequence of worlds $s = \langle w_1, ..., w_m \rangle$ corresponds to an admissible similarity ordering just in case all the worlds in s comes from a single cell Z_i. To illustrate this, let $\mathcal{W}$ again be the set $\{w_1, w_2, w_3\}$, and suppose the relevant partition is $Z = \{\{w_1, w_2\}, \{w_3\}\}$. Then, the admissible similarity orderings here are $\langle w_1, w_2 \rangle$, $\langle w_2, w_1 \rangle$, and $\langle w_3 \rangle$, respectively, since these are the only orderings we can generate from the cells of the corresponding partition.

Given this constraint, we can give a slightly different definition of our lifting procedure:

(i) $pr([w]) = cr(w)$,

(ii*) $pr([w_1, ..., w_k]) = pr([w_1, ..., w_{k-1}]) \cdot cr(w_k \mid Z_i - \{w_1, ..., w_{k-1}\})$.

Here, Z_i is the partition cell to which w_k belongs. So the only difference between this definition of the lift of cr and our original definition is that, in this new case, once a base world has been chosen, we only consider worlds from the same partition-cell as that world. (In fact, our old definition and this new one agree, whenever the relevant partition is the trivial partition, $\mathcal{W}$.)

All this applies equally when we have more than one salient question in a context. For example, in cases where the similarity relations are constrained by the coarsest common refinement of two (or more) such questions, we can replace Z_i in the above definition with the cells from this coarsest common refinement. Similarly, when similarity relations are constrained by the finest common coarsening of some questions, we can replace Z_i with the cells from this latter partition. Like I said, the first of these cases won't play much of a role in what's to come. But the second one will be important in the next section, and I'll say more about it then.

For now, we're at last in a position to state the version of CDT that I've been working towards. In my view, when you're making a decision, you should choose an option that maximizes the following quantity, which I'll still refer to as *utility*:

$$U(A) = \sum_i pr(A > O_i) \cdot v(O_i). \tag{2}$$

This decision rule looks more-or-less identical to the original Stalnaker-Gibbard-Harper rule. The only difference is that, in the case of (2), the lifted credence function, pr, replaces the original credence function, cr. This means that the counterfactuals $A > O_i$ appealed to in this equation don't always have to correspond to a set of worlds. Instead, they can correspond to the set of all admissible similarity orderings at which that counterfactual is true. As we'll see later on, it's this change that helps us to get the right answer in cases involving indeterminacy.

Let me now close this section by making a few additional comments about the version of CDT I've just introduced.

First, you'll notice that, although I've been speaking throughout about *counterfactuals*, all of the worlds appealed to in my theory, as I've set it up here, are epistemically possible worlds. This, I think, is an important thing to point out, because some authors object to Stalnaker-Gibbard-Harper CDT on the grounds that it requires you to think about epistemically impossible worlds in certain deterministic cases. Kment (2023), for instance, criticizes CDT in this way, saying that epistemically impossible worlds are "irrelevant to a rational assessment of your options… Reflection on such worlds is a form of wishful thinking that has no place in rational choice" (p. 10). I'm not sure I agree with Kment about this for every decision problem

(which worlds are relevant, after all, is a matter of context, in my view). But in any case, the objection has no force against my theory, since, as we'll see below, this theory gets the right answer in deterministic cases, and only appeals to epistemically possible worlds.[40]

Additionally, the formal framework I've set up here owes a lot to Khoo and Santorio (2018), Goldstein and Santorio (2021), Khoo (2022), and Mandelkern (forthcoming), all of whom use a similar framework for a very different purpose—namely, to prove *tenability results* for versions of *Stalnaker's thesis* (Stalnaker, 1970).[41] Recall that Stalnaker's thesis relates your credences in indicative conditionals to your conditional credences. Specifically, it says that, if you're rational, your credence in an indicative conditional $P > Q$ will match your conditional credence in Q given that P (assuming this is well-defined). Formally: $pr(P > Q) = pr(Q \mid P)$.[42] For a long time, it was thought that this thesis couldn't be true, owing to the famous triviality results of Lewis (1976) and others. But as the authors mentioned above have recently shown, versions of Stalnaker's thesis can hold (non-trivially) after all, provided all the sequences of possible worlds count as admissible in a context. (If not all sequences are admissible, then the situation is more complicated. See, e.g., Khoo (2016, 2022) and Mandelkern (forthcoming).)

In the present setting, these tenability results turn out to have an interesting upshot. To see what it is, consider the following alternative to CDT's decision rule. Suppose that the quantity you should maximize when you're making a decision isn't U, but the following:

$$V(A) = \sum_i pr(O_i \mid A) \cdot v(O_i) \tag{3}$$

This quantity is sometimes called the *news value* of A. And it's the quantity that CDT's chief rival, *evidential decision theory* (EDT), tells you to maximize when you're choosing between your options.[43] Thus, what the tenability results I mentioned above imply is that, in any context in which all the similarity orderings are admissible, the version of CDT I've advocated for here will give the same recommendations as EDT. After all, in any case like that, $pr(A > O_i) = pr(O_i \mid A)$ for all O_i, and so $U(A) = V(A)$. This, I think, is a very interesting point of connection between my theory and a rival. And as we'll now see, it has important consequences for some of the decision problems we'll reconsider.

6 Cases Redux

With my version of Stalnaker-Gibbard-Harper CDT now in place, let's return to Ahmed's cases. In this section, I'll show that my theory gets the right answer in those cases. (After seeing this, it should also be

[40] Is it right to say that the conditionals in my theory are really then *counterfactuals*, rather than, say, indicative conditionals? You might be worried that they're the latter. However, I think it's still legitimate to call these conditionals 'counterfactuals' because the relations of similarity that are relevant to their assessment are those that hold fixed facts about causal connections, rather than, say, facts about (mere) epistemic connections. As Stalnaker (1975) and others have argued, the key difference between counterfactuals and indicative conditionals seems not to be anything to do with "counterfactuality" per se, but instead the fact that indicative conditionals are about epistemic possibilities, and counterfactuals are about causal or metaphysical possibilities.

[41] See also van Fraassen (1976) and Bacon (2015).

[42] I use the same symbol, '$>$', for both indicative conditionals and counterfactuals because Stalnaker's semantics is a *uniform* semantics. That is, it says that the truth-conditions for indicative conditionals and counterfactuals are one and the same, and all the differences between these conditionals come down to the salient similarity relations that we use to assess them. When all similarity relations are admissible in a context, however—and the context set consists just of epistemically possible worlds—Stalnaker's semantics says that these types of conditionals coincide. There's some evidence that this is indeed the case of natural language conditionals. For example, so-called "future-directed" counterfactuals often seem to say the same thing as corresponding indicative conditionals. Compare: 'If I were to flip the coin, it would land heads' and 'If I flip the coin, it will land heads'. Plausibly, the reason for this convergence is that we're using the same relations of similarity to assess these conditionals. See, e.g., Edgington (1995) for further discussion.

[43] EDT was first introduced by Richard Jeffrey (1965); see also Jeffrey (1983). Incidentally, Ahmed himself vigorously defends EDT over CDT, and sees his deterministic cases as giving us a reason to favor the former.

obvious how my theory handles analogous deterministic cases, like those recently discussed by Williamson and Sandgren (forthcoming), Gallow (2022), Kment (2023), and others.) I'll also show that my theory gives the two-boxing recommendation in *Newcomb*. And I'll close the paper by considering a case that we haven't yet looked at.

6.1 *Betting on the Laws* Redux

Let's start with *Betting on the Laws*. In that case, you were offered a choice between two bets on the proposition L, namely: B_1, which pays \$1 if L is true, but pays nothing if L is false; and B_2, which pays nothing if L is true, but pays \$1 if L is false. Here, again, is the decision table:

	L	$\neg L$
B_1	\$1	\$0
B_2	\$0	\$1

Table 2: *Betting on the Laws*

Now, given what I said in §4, it should be clear that the salient question here corresponds to $\{L, \neg L\}$. After all, the propositions in this partition are both causally independent of your choice; they're also consistent with each of your options; and any cell of this partition determines the amount of money you'll receive, once you've chosen a particular option. Thus, it follows that every admissible similarity ordering makes one of the following biconditionals true:

$$(B_1 > \$1) \equiv (B_2 > \$0),$$

$$(B_1 > \$0) \equiv (B_2 > \$1).$$

The sequences at which the first biconditional is true partition the proposition L. So your credences satisfy:

$$pr(B_1 > \$1) = pr(B_2 > \$0) = pr(L).$$

Similarly, the sequences at which the second biconditional is true partition the proposition $\neg L$. So your credences also satisfy:

$$pr(B_1 > \$0) = pr(B_2 > \$1) = pr(\neg L).$$

Given these equalities, we have:

$$\begin{aligned} U(B_1) &= pr(B_1 > \$1) \cdot 1 + pr(B_1 > \$0) \cdot 0 \\ &= pr(L) \cdot 1 + pr(\neg L) \cdot 0 \\ &= pr(L) \\ U(B_2) &= pr(B_2 > \$0) \cdot 0 + pr(B_2 > \$1) \cdot 1 \\ &= pr(L) \cdot 0 + pr(\neg L) \cdot 1 \\ &= pr(\neg L). \end{aligned}$$

Then, since $pr(L) \approx 1$ and $pr(\neg L) \approx 0$, it follows that $U(B_1) \approx 1$ and $U(B_2) \approx 0$. So my theory recommends B_1—the right answer.

Notice that, since the admissible sequences in this case partition the propositions L and $\neg L$, our calculations of utility simplified. Specifically, we ended up being able to calculate $U(B_1)$ and $U(B_2)$ directly in terms of your credences in the propositions L and $\neg L$. As it turns out, the same thing goes in any decision problem with similar features. That is, so long as there's just one salient partition in a decision problem, consisting of state-propositions whose truth-values are all causally independent of what you do, we can always calculate utility directly in terms of your credences in the states.

6.2 *Betting on the Past* Redux

Seeing what my theory says about *Betting on the Laws* was straightforward. But seeing what it says about *Betting on the Past* is a little trickier. After all, we saw in §4 that there isn't just one salient question here, but two. I'll repeat the relevant tables for convenience:

	H	$\neg H$
B_3	\$1	$-$\$10
B_4	\$10	$-$\$1

Table 3: *Betting on the Past*

	L	$H \wedge \neg L$	$\neg H \wedge \neg L$
B_3	\$1	\$1	$-$\$10
B_4	$-$\$1	\$10	$-$\$1

Table 5: *Betting on the Past*

Now, given that there's more than one salient question in this case, a natural first thought is that we should take the relevant similarity relations to be constrained, not by either of these questions alone, but by their conjunction—or more precisely, their coarsest common refinement. Unfortunately, however, this won't work, since the partition we end up with is one where your options are inconsistent with some of the cells, violating well-definedness. (See fn. 32 above, as well as Joyce (2016), Solomon (2021), Elga (2022), and Fusco (forthcoming) for further discussion.) So we need to try out something else.

Thus, consider the other thing I said in §4. There, I said that in certain contexts, it's more appropriate to think of similarity relations as being constrained by the *finest common coarsening* of questions, rather than by their coarsest common refinement. *Betting on the Past* shows, I think, why this is sometimes the case. After all, the two questions here "compete" with one another, in the sense that holding fixed the answers to one question means you can't hold fixed the answers to the other—at least not when you're deliberating about what to do.

In a bit more detail: imagine you choose B_3 and then win you \$1. Then, you're almost certainly at a world where both H and L are true. But if that's so, then ask yourself: what would have happened if you had chosen the option B_4 instead? Here, it seems like you're pulled in two different directions. One salient question seems to imply that choosing B_4 would've won you \$10; but another seems to imply that choosing B_4 would've lost you a dollar. Thus, there seem to be different admissible precisifications of 'closest B_4-world in this case, which it makes it indeterminate what your choice of B_4 would've resulted in. In other words, different admissible precisifications say that different outcomes would've occurred, if you'd chosen otherwise than you actually did.

This, I think, is one of the things that makes *Betting on the Past* so interesting. In my view, the counterfactuals in this case admit of a significant amount of indeterminacy, in virtue of the two different questions in play. In order to capture this indeterminacy, we have to allow a whole range of similarity orderings to count as admissible. And the best way to do this, I believe, is to "merge" the salient questions, and think of similarity relations as being constrained by their finest common coarsening. This is the most plausible way I can see to allow, e.g., that there are non-actual B_4-worlds at which you win \$10, but also others at which you lose \$1, as in the above example. Similarly for the different precisifications of analogous counterfactuals.

But what *is* the finest common coarsening of the relevant questions in *Betting on the Past*? Well, since the only partition that's coarser than $\{H, \neg H\}$ is the trivial partition, $\mathcal{W}$, the only partition that's coarser than both of these questions is, again, the context set $\mathcal{W}$. Thus, on my analysis, *every sequence of worlds* is admissible in *Betting on the Past*. And this turns out to have an important upshot. Recall that in the previous section, I said it's implied by the tenability results for Stalnaker's thesis that, when all the sequences of possible worlds are admissible in a context, my theory gives the same recommendations as EDT. After all,

in cases like that, we have $pr(A > O_i) = pr(O_i \mid A)$ for each O_i. So, given these equalities, we have:

$$\begin{aligned}
U(B_3) &= pr(B_3 > \$1) \cdot 1 + pr(B_3 > -\$10) \cdot -10 \\
&= pr(\$1 \mid B_3) \cdot 1 + pr(-\$10 \mid B_3) \cdot -10 \\
&\approx 1 \cdot 1 + 0 \cdot -10 \\
&= 1 \\
U(B_4) &= pr(B_4 > \$10) \cdot 10 + pr(B_4 > -\$1) \cdot -1 \\
&= pr(\$10 \mid B_4) \cdot 10 + pr(-\$1 \mid B_4) \cdot -1 \\
&\approx 0 \cdot 10 + 1 \cdot -1 \\
&= -1.
\end{aligned}$$

So, in the end, my theory says that $U(B_3) \approx 1$ and $U(B_4) \approx -1$. The theory thus recommends you choose B_3—the right answer.

6.3 *Newcomb* Redux

Contrast this with what my theory says about the *Newcomb* problem. In that case, the only salient question is *How much money is in the opaque box?*.[44] So it follows that every admissible similarity ordering is one which makes one of the following material biconditionals true:

$$(\textit{One-box} > \$1m) \equiv (\textit{Two-box} > \$1m + 1k),$$

$$(\textit{One-box} > \$0) \equiv (\textit{Two-box} > \$1k).$$

Given this, things play out much as they did in *Betting on the Laws*. That is, the sequences at which the first biconditional is true partition the proposition *Million*. So your credences satisfy:

$$pr(\textit{One-box} > \$1m) = pr(\textit{Two-box} > \$m + 1k) = pr(\textit{Million}).$$

Similarly, the sequences at which the second biconditional is true partition the proposition *No million*. So your credences also satisfy:

$$pr(\textit{One-box} > \$0) = pr(\textit{Two-box} > \$1k) = pr(\textit{No million}).$$

This means that we can calculate utility straightforwardly using your credences in states:

$$\begin{aligned}
U(\textit{One-box}) &= pr(\textit{Million}) \cdot 1m + pr(\textit{No million}) \cdot 0 \\
&= pr(\textit{Million}) \cdot 1m \\
U(\textit{Two-box}) &= pr(\textit{Million}) \cdot (1m + 1k) + pr(\textit{No million}) \cdot 1k \\
&= pr(\textit{Million}) \cdot 1m + 1k.
\end{aligned}$$

The upshot is that $U(\textit{One-box}) < U(\textit{Two-box})$, no matter what your credences in *Million* and *No million*. So my theory says that you should take both boxes, just as our original version of CDT did.

[44] Zach Barnett asks me why we can't think of the partition {*The predictor is accurate, The Predictor is inaccurate*} as corresponding to a salient question in *Newcomb*. One answer is that we can—although it's not one that can constrain the similarity relations for *standard* counterfactuals. The reason is that counterfactuals whose similarity relations are constrained by this partition are plausibly true only on a backtracking interpretation. And in §2, I set backtracking counterfactuals aside. Along the same lines, it's plausiblethat the the propositions in this partition aren't really causally independent of your options. The reason, as Joyce (2018) points out, is that it's an important feature of any genuine *Newcomb*-type problem that you believe you have the power to make the predictor's prediction wrong—even if you're certain you won't actually do this.

I think this demonstration shows that there's an important distinction between *Newcomb* and *Betting on the Past*. Specifically, in the former case, there's just one salient question raised by your decision situation; but in the latter case, there are two. Moreover, it's only if we focus on one of the questions that a causal dominance argument can be mounted in *Betting on the Past*. If we focus on the other question instead, then causal dominance reasoning doesn't apply. Thus, something Elga said earlier is plausibly mistaken. Recall his remark that:

> In a standard Newcomb problem there is a causal dominance argument for taking two boxes: 'The $1 million is either there or it is not, and you have no causal influence on whether it is. Either way (and no matter what else is true), taking two boxes gets you a better outcome than taking just one. So you should take two boxes.'... These conditions are satisfied in [*Betting on the Past*] just as much as they are in a standard Newcomb problem. (p. 207)

My sense, however, is that this isn't right. In *Newcomb*, *Two-box* causally dominates *One-box* on *every* admissible precisification of the counterfactuals. In *Betting on the Past*, in contrast, B_4 causally dominates B_3 only on some admissible precisifications. Thus, it isn't right to say that the relevant conditions "are satisfied in [*Betting on the Past*] just as much as they are in a standard Newcomb problem". So, in my view, causal decision theorists have a good reason to resist Elga's conclusion.

6.4 A New Case

I'm going to wrap up now by considering a new case. I'm adapting it from one recently discussed by Hedden (2023). And similar cases have been considered by Solomon (2021), Gallow (2022), and Fusco (forthcoming), among others. The case is also a bit like the one we considered right at the outset:

> *Betting on the Past and the Laws.* You have to choose between two bets on a proposition D. First, there's B_5, which pays $1 if D is true, but loses $10 if D is false. Second, there's B_6, which pays $10 if D is true, and loses only $1 if D is false. D is the proposition that the past state of the world, together with the laws of nature, determines that you accept B_5. And because you're certain of determinism, you're certain that the negation of this proposition determines that you accept B_6.

	D	$\neg D$
B_5	$\$1$	$-\$10$
B_6	$\$10$	$-\$1$

Table 6: *Betting on the Past and the Laws*

On its face, this case looks a little bit like *Betting on the Past*.[45] However, there's an important difference between that case and this new one. Unlike in *Betting on the Past*, the proposition D here is *inconsistent* with your choosing B_6, while $\neg D$ is inconsistent with your choosing B_5. So, the only way you won't lose a dollar by choosing B_6 is if a contradiction is true.

Hedden claims that various formulations of CDT tell you to choose B_6 in *Betting on the Past and the Laws*. After all, he says, the proposition D

[45] Some authors—Hedden among them, but also Stalnaker (MS)—say that this case just *is Betting on the Past*. But I disagree. As Ahmed (2014b) is careful to say, for example, the proposition H in *Betting on the Past* is not meant to be a "cheesy" proposition, of the form 'The past and the laws, whatever they are, determine that you accept B_3'. Rather, H is quite specific in what it describes—namely, the intrinsic state of the world at some particular time in the distant past. This is the reason my treatment of *Betting on the Past* got quite complicated. That said, even though I think the case given here is distinct from *Betting on the Past*, my treatment of it is broadly similar to Stalnaker's.

> is a proposition about the [history] of the universe and the laws of nature. The [history] of the universe and the laws of nature are both beyond your causal control; you cannot cause either one to be different. Therefore, no matter how things beyond your causal control might be (i.e. no matter whether D or $\neg D$ is true), B_6 yields a strictly better outcome than B_5. (2023, p. 743, notation adapted)

But more than any case we've seen, this is clearly absurd.

Now, I'm not convinced that CDT, in any formulation, tells you to choose B_6, contrary to what Hedden says. (Although I'll concede that some versions of the theory may simply fall silent—think, for instance, about the "dependency hypothesis" versions of CDT due to Lewis and Skyrms.) However, setting aside my worries about Hedden's applications of CDT, I want to show that my theory gets the right answer in this case anyway.

The reasoning here is straightforward. Once more, recall from §4 that I said that, in any decision situation, the salient question must be one for which each cell of the corresponding partition is consistent with each your options. Thus, given this way of thinking about salient questions, it should be clear that there's only one such "question" that can be said to be raised in this situation. This is just the trivial question, $\mathcal{W}$ itself. So your decision problem looks as follows:

	$\mathcal{W}$
B_5	\$1
B_6	$-\$1$

Table 7: *Betting on the Past and the Laws*

Then, since there's just one outcome consistent with B_5, and one outcome consistent with B_6; and since the first outcome is clearly better than the second; my theory tells you to choose B_5.[46] The upshot is that, even if Hedden is correct to say that other versions of CDT get this case wrong, my theory gets it right. It tells you to choose the option that has the best *possible* outcome.

References

Ahmed, A. (2013). Causal decision theory: A counterexample. *Philosophical Review*, *122*(2), 289–306. https://doi.org/10.1215/00318108-1963725

Ahmed, A. (2014a). Causal decision theory and the fixity of the past. *The British Journal for the Philosophy of Science*, *65*(4), 665–685. https://doi.org/10.1093/bjps/axt021

Ahmed, A. (2014b). *Evidence, decision and causality*. Cambridge University Press.

Albert, D. Z. (2000). *Time and chance*. Harvard University Press.

Bacon, A. (2015). Stalnaker's thesis in context. *The Review of Symbolic Logic*, *8*(1), 131–163. https://doi.org/10.1017/S1755020314000318

Bennett, J. (1984). Counterfactuals and temporal direction. *The Philosophical Review*, *93*(1), 57. https://doi.org/10.2307/2184413

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford University Press.

Boylan, D., & Schultheis, G. (2021). How strong is a counterfactual? *Journal of Philosophy*, *118*(7), 373–404. https://doi.org/10.5840/jphil2021118728

[46] In a bit more detail: since the salient question here is just the trivial question, $\mathcal{W}$, all sequences of possible worlds count as admissible in the context. So, by the tenability results for Stalnaker's thesis, it follows that we can calculate utility using your conditional credences, rather than your credences in counterfactuals. Then, since $pr(\$1 \mid B_5) = 1$ and $pr(-\$1 \mid B_6) = 1$, it follows that $U(B_5) = 1$ and $U(B_6) = -1$. Thus, my theory tells you to choose B_5.

Braddon-Mitchell, D. (2001). Lossy laws. *Noûss*, *35*(2), 260–277. https://doi.org/10.1111/0029-4624.00296
Dorr, C. (2016). Against counterfactual miracles. *Philosophical Review*, *125*(2), 241–286. https://doi.org/10.1215/00318108-3453187
Dorr, C., & Hawthorne, J. (MS). *If…: A theory of conditionals* [Unpublished Manuscript].
Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329. https://doi.org/10.1093/mind/104.414.235
Edgington, D. (2003). Counterfactuals and the benefit of hindsight. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world*. Routledge.
Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *68*(S3), S313–S324. https://doi.org/10.1086/392918
Elga, A. (2022). Confession of a causal decision theorist. *Analysis*. https://doi.org/10.1093/analys/anab040
Fusco, M. (forthcoming). Absolution of a causal decision theorist. *Noûs*.
Gallow, J. D. (2022). Causal counterfactuals without miracles or backtracking. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12925
Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and applications of decision theory* (pp. 125–162). D. Reidel.
Goldstein, S., & Santorio, P. (2021). Probability for epistemic modalities. *Philosophers' Imprint*, *21*(33).
Goodman, J. (2014). Knowledge, counterfactuals, and determinism. *Philosophical Studies*, *172*(9), 2275–2278. https://doi.org/10.1007/s11098-014-0409-6
Groenendijk, J. A. G., & Stokhof, M. J. B. (1984). *Studies on the semantics of questions and the pragmatics of answers* (Doctoral dissertation). University of Amsterdam.
Hamblin, C. L. (1973). Questions in Montague english. *Foundations of Language*, *10*(1), 41–53.
Hedden, B. (2023). Counterfactual decision theory. *Mind*, *132*(527), 730–761. https://doi.org/10.1093/mind/fzac060
Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous*, *39*(4), 632–657. https://doi.org/10.1111/j.0029-4624.2005.00542.x
Hoek, D. (2019). *The web of questions the web of questions: Inquisitive decision theory and the bounds of rationality* (Doctoral dissertation). New York University.
Hoek, D. (2022). Questions in action. *The Journal of Philosophy*, *119*(3), 113–143. https://doi.org/10.5840/jphil202211938
Holguín, B., & Teitel, T. (MS). *On the plurality of counterfactuals* [Unpublished Manuscript].
Ichikawa, J. (2011). Quantifiers, knowledge, and counterfactuals. *Philosophy and Phenomenological Research*, *82*(2), 287–313.
Ippolito, M. (2016). How similar is similar enough? *Semantics and Pragmatics*, *9*, 6–1.
Jackson, F. (1977). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, *55*(1), 3–21. https://doi.org/10.1080/00048407712341001
Jeffrey, R. C. (1965). *The logic of decision.* 1st Edition, University of Chicago Press.
Jeffrey, R. C. (1983). *The logic of decision.* 2nd Edition, Chicago; London: University of Chicago Press.
Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.
Joyce, J. M. (2009). Causal reasoning and backtracking. *Philosophical Studies*, *147*(1), 139–154. https://doi.org/10.1007/s11098-009-9454-y
Joyce, J. M. (2016). Review of Arif Ahmed: Evidence, Decision and Causality. *Journal of Philosophy*, *113*(4), 224–232. https://doi.org/10.5840/jphil2016113413
Joyce, J. M. (2018). Deliberation and stability in Newcomb problems and pseudo-Newcomb problems. In A. Ahmed (Ed.), *Newcomb's problem* (pp. 138–159). Cambridge University Press. https://doi.org/10.1017/9781316847893.008
Kaufmann, S. (2004). Conditioning against the grain. *Journal of Philosophical Logic*, *33*(6), 583–606. https://doi.org/10.1023/b:logi.0000046142.51136.bf

Khoo, J. (2016). Probabilities of conditionals in context. *Linguistics and Philosophy*, *39*(1), 1–43. https://doi.org/10.1007/s10988-015-9182-z
Khoo, J. (2017). Backtracking counterfactuals revisited. *Mind*, fzw005. https://doi.org/10.1093/mind/fzw005
Khoo, J. (2022). *The meaning of 'If'*. New York, USA: Oxford University Press.
Khoo, J., & Santorio, P. (2018). *Lecture notes: Probabilities of conditionals in modal semantics* [Unpublished Manuscript].
Kment, B. (2006). Counterfactuals and explanation. *Mind*, *115*(458), 261–310. https://doi.org/10.1093/mind/fzl261
Kment, B. (2014). *Modality and explanatory reasoning*. Oxford University Press.
Kment, B. (2023). Decision, causality, and predetermination. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12935
Lange, M. (2000). *Natural laws in scientific practice*. Oxford University Press.
Lewis, D. (1973a). Causation. *The Journal of Philosophy*, *70*(17), 556. https://doi.org/10.2307/2025310
Lewis, D. (1973b). *Counterfactuals*. Blackwell.
Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, *85*(3), 297. https://doi.org/10.2307/2184045
Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476. https://doi.org/10.2307/2215339
Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*(1), 5–30. https://doi.org/10.1080/00048408112340011
Lewis, D. (1986). *Philosophical papers: Volume II*. Oxford University Press.
Lewis, D. (1988a). Relevant implication. *Theoria*, *54*(3), 161–174. https://doi.org/10.1111/j.1755-2567.1988.tb00716.x
Lewis, D. (1988b). Statements partly about observation. *Philosophical Papers*, *17*(1), 1–31. https://doi.org/10.1080/05568648809506282
Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, *97*(4), 182–197. https://doi.org/jphil200497437
Lewis, K. (2015). Elusive counterfactuals. *Noûs*, *50*(2), 286–313. https://doi.org/10.1111/nous.12085
Loewer, B. (2007). Counterfactuals and the second law. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited*. Oxford University Press.
Mandelkern, M. (2018). Talking about worlds. *Philosophical Perspectives*, *32*(1), 298–325. https://doi.org/10.1111/phpe.12112
Mandelkern, M. (forthcoming). *Bounds: The dynamics of interpretation* [Unpublished Manuscript]. Oxford University Press.
Maudlin, T. (2007). *The metaphysics within physics*. Oxford University Press.
McNamara, C. (MS). *Desire-as-belief in context* [Unpublished Manuscript].
Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of carl g. hempel* (pp. 114–146). Reidel.
Nute, D. (1980). Conversational scorekeeping and conditionals. *Journal of Philosophical Logic*, *9*(2). https://doi.org/10.1007/bf00247746
Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*. https://doi.org/10.3765/sp.5.6
Sandgren, A., & Williamson, T. L. (2020). Determinism, counterfactuals, and decision. *Australasian Journal of Philosophy*, *99*(2), 286–302. https://doi.org/10.1080/00048402.2020.1764073
Santorio, P. (2019). Interventions in premise semantics. *Philosophers' Imprint*, *19*.
Schultheis, G. (forthcoming). Counterfactual probability. *Journal of Philosophy*.
Skyrms, B. (1980). *Causal necessity: A pragmatic investigation of the necessity of laws*. Yale University Press.
Skyrms, B. (1982). Causal decision theory. *The Journal of Philosophy*, *79*(11), 695. https://doi.org/10.2307/2026547

Skyrms, B. (1984). *Pragmatics and empiricism*. Yale University Press, New Haven.
Slote, M. A. (1978). Time in counterfactuals. *Philosophical Review*, *87*(1), 3–27. https://doi.org/10.2307/2184345
Sobel, J. H. (1994). *Taking chances: Essays on rational choice*. Cambridge University Press.
Solomon, T. C. P. (2021). Causal decision theory's predetermination problem. *Synthese*, *198*(6), 5623–5654. https://doi.org/10.1007/s11229-019-02425-0
Solomon, T. C. P. (MS). *Libertarian decision theory* [Unpublished Manuscript].
Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory (american philosophical quarterly monographs 2)* (pp. 98–112). Oxford: Blackwell.
Stalnaker, R. (1970). Probability and conditionals. *Philosophy of Science*, *37*(1), 64–80. https://doi.org/10.1086/288280
Stalnaker, R. (1975). Indicative conditionals. *Philosophia*, *5*(3), 269–286. https://doi.org/10.1007/bf02379021
Stalnaker, R. (1978). Assertion. *Syntax and Semantics (New York Academic Press)*, *9*, 315–332.
Stalnaker, R. (1981a). A defense of conditional excluded middle. In W. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 87–104). Reidel.
Stalnaker, R. (1981b). Letter to David Lewis. In W. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 151–152). Reidel.
Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.
Stalnaker, R. (2021). Counterfactuals and probability. In L. Walters & J. Hawthorne (Eds.), *Conditionals, paradox, and probability: Themes from the philosophy of dorothy edgington*. Oxford University press.
Stalnaker, R. (MS). *Counterfactuals, compatibilism, and rational choice* [Unpublished Manuscript].
Stalnaker, R., & Thomason, R. (1970). A semantic analysis of conditional logic. *Theoria*, *36*(1), 23–42. https://doi.org/10.1111/j.1755-2567.1970.tb00408.x
Steele, K., & Sandgren, A. (2020). Levelling counterfactual scepticism. *Synthese*, *199*(1-2), 927–947. https://doi.org/10.1007/s11229-020-02742-9
van Fraassen, B. (1976). Probabilities of conditionals. In W. H. C. Hooker (Ed.), *Foundations of probability theory, statistical inference, and statistical theories of science*.
Williamson, T. L., & Sandgren, A. (forthcoming). Law-abiding causal decision theory. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/715103
Yalcin, S. (2016). Belief as question-sensitive. *Philosophy and Phenomenological Research*, *97*(1), 23–47. https://doi.org/10.1111/phpr.12330