

Only Causal Decision Theory Defers to Experts

Calum McNamara*

Forthcoming in the *Journal of Philosophy*

Abstract. The philosophical literature contains numerous proposals saying how you should *defer* to expert opinions. For example, there are *Reflection* principles, *New Reflection* principles, and the more recent principles of *Trust*. All of these principles can be motivated on the grounds that, if you defer to experts in the way they suggest, then you'll make better decisions on average. However, one issue with these arguments supporting Reflection, New Reflection, etc., is that they rely on the tools of expected utility (EU) theory. But many philosophers reject EU theory, since it's insufficiently general. In its place, they endorse some form of *suppositional decision theory*, like causal decision theory (CDT) or evidential decision theory (EDT). A question thus arises: Do the defenses of expert deference principles like Reflection carry over to the suppositional setting? In this paper, I show that the answer is 'No', or at least 'Not straightforwardly'. Only one suppositional decision theory preserves the idea that it's always good to defer to an expert while you're making a decision. This is CDT—or rather, a particular version thereof.

1 Introduction

Recall the *Newcomb* problem (Nozick, 1969):

Newcomb. In front of you are two boxes, labeled 'A' and 'B', respectively. You can take either just box A, or both boxes. Box B is transparent and contains what you can see to be a \$1,000 bill. Box A, in contrast, is

*For helpful discussion, I'm grateful to Zach Barnett, Melissa Fusco, Dmitri Gallow, Ben Holguín, Wes Holliday, Jim Joyce, and audiences at the University of California, Berkeley, the University of Massachusetts, Amherst, and the Pacific APA. Thanks especially to two anonymous reviewers for this journal; to Ryan Doody and Simon Huttegger—my commentators at the APA—for their diligent and thoughtful feedback; and to Ben Levinstein and Bernhard Salow, for helpful discussions of the Total Trust principle. Finally, special thanks go to Richard Roth and Snow Zhang, who helped me understand the background for this paper much better than I could have on my own.

opaque, and you don't know what's inside it. The contents of box A were determined yesterday on the basis of a prediction I made about your behavior. If I predicted that you'd take only box A, then I put \$1,000,000 inside that box. However, if I predicted that you'd take both boxes, then I left the opaque box empty. Note that I'm a highly reliable predictor of your behavior. So, with that in mind: What is your choice?

There's a compelling argument for the claim that you should take both boxes in *Newcomb*.¹ To see how it works, first imagine that you have a friend, Bob, who's managed to sneak a peek inside the opaque box. Bob can't tell you what's inside that box; but he can give you advice about what to do. How, then, would "better-informed-Bob" advise you?

Well, suppose that Bob has spied a million dollars in the opaque box. Then, clearly he'd tell you to take both boxes, since doing this would get you a thousand dollars more than taking only one box would.

Now suppose that Bob has spied nothing in the opaque box. Then, again, it seems like he'd tell you to take both boxes. After all, taking both boxes would leave you with \$1,000, while taking only box A would leave you with nothing.

In both cases, then, better-informed-Bob would tell you to take both boxes. And since he's the *expert* in this situation—he has more information than you do—you'd be wise to listen to his advice. (Of course, in reality, you don't even need to ask for Bob's advice in *Newcomb*, since you already know, in advance, what he would tell you to do. But the point still stands: the fact that the expert here *would* tell you to take both boxes, if you asked him, gives you a strong reason to do so now.)

Like I said, this is a compelling argument for taking both boxes in *Newcomb*. However, I'm spelling it out now not to convince you that "two-boxing" is the right response. Rather, I'm giving this argument because it illustrates a more general idea about rational decision-making. This is that, when you're in a position to do so, it's always sensible to *defer* to the advice of an expert when deciding. (In what follows, I will construe the notion of an *expert* broadly, so as to include, not only better informed people like Bob, but also your future self, the objective chances, and so on.)

Many philosophers endorse this idea about deference to experts in decision-making. In fact, some even say that it's part of what it *means* to call someone an expert that you're disposed to defer to them while deciding. Dorst et al. (2021) say, for example, that characterizing someone as "an expert requires that, for any decision problem, you would prefer to give the expert power of attorney: to use their probabilities to make the decision rather than use your own" (p. 104).²

¹Versions of this argument have been given by, e.g., Nozick (1969, pp. 116–17), Pollock (2009, pp. 70–73), Ahmed (2014b, p. 199), and Kment (2023, pp. 641–42), among others.

²See also Dorst (2019, pp. 604–05) for similar remarks.

But what exactly does it mean to “use an expert’s probabilities” in your decision-making? Without some specific answer to that question on the table, it’s not clear how we can work the idea into a general theory of rational choice.

In the philosophical literature, there are numerous proposals saying how you should “use” an expert’s probabilities. For example, there are *Reflection* principles, *New Reflection* principles, and the more recent principles of *Trust*.³ All of these principles can be motivated on the grounds that, if you align your opinions with the expert’s in the way that they suggest, then you’ll make better decisions on average.

But there’s a catch. The arguments deployed in support of these principles typically rely on the tools of *expected utility (EU) theory*. In particular, they show that an agent who chooses by maximizing EU will never go wrong by deferring to an expert, in the way described by the principle in question.⁴ In philosophy, however, EU theory is widely thought to be insufficiently general. The issue is that, when the options you can choose and the states of the world aren’t *independent*—causally, evidentially, or otherwise—EU theory falls silent, or else can give very bad advice, about what you should do. For this reason, many philosophers endorse some *generalization* of EU theory, like *causal decision theory* (CDT) or *evidential decision theory* (EDT).⁵

CDT and EDT are both members of a wide class of decision theories known as *suppositional decision theories*. Roughly speaking, these theories say that you should choose an option, not by considering how good its outcome might be, but by considering how good its outcome might be *under the supposition that it’s chosen*. Making decisions in this way avoids the problems faced by standard EU theory. But it also raises an important question: Do the arguments mentioned above, which show a link between principles of expert deference and rational decision-making, hold up in the suppositional setting? (This question is important not least because certain proponents of principles like Reflection are also staunch supporters of suppositional decision theories—see, e.g., Skyrms (1980, 1982, 1984, 1990) and Huttegger (2013a,b, 2023).)

In this paper, my goal is to answer that question. Which—if any—suppositional decision theories preserve the idea that it’s always good to defer to an expert while you’re making a decision? In particular, I want to show that, among all the suppositional decision theories that are out there, there’s only one that can be

³My terminology here follows Dorst et al. (2021, p. 102), who use ‘Reflection’ to capture a wide class of expert deference principles. ‘Reflection’ is also used in a more specific sense in the literature—namely, to denote a principle proposed by van Fraassen (1984, p. 244). Given this dual usage, as well as van Fraassen’s historical precedence, it would arguably be more appropriate to use another name for the more general class of principles that authors like Dorst et al. discuss. However, to keep my discussion consistent with theirs, I’ll stick with ‘Reflection’ in this paper. Thanks to an anonymous referee for pushing me to clarify this. (Note also that all this holds equally for the term ‘New Reflection’.)

⁴For arguments of this kind, see Skyrms (1990, pp. 97–98), Huttegger (2013b, pp. 280–83), Dorst (2019, pp. 604–08), and Dorst et al. (2021, pp. 104–05).

⁵As an anonymous reviewer for the *Journal* notes, it’s possible to think of EU theory as an implicitly causal decision theory. I discuss this more in fn. 23 below.

plausibly said to preserve the foregoing idea. This is CDT—or rather, a particular version thereof. All other suppositional decision theories say that, even if you recognize someone as an expert now, you can ignore their opinion while you’re making a decision.⁶

This, I think, provides a strong argument in support of CDT, and a strong argument against other suppositional decision theories.⁷ After all—as the *Newcomb* problem illustrates—it’s generally a good idea to defer to the advice of an expert when deciding, if you can.

My argument also adds to a stock of arguments in favor of CDT in the recent literature. To put things in context: initially, CDT was introduced as an alternative to EDT, intended to give the right answer in the *Newcomb* problem.⁸ Since then, the debate surrounding CDT and EDT has proceeded mostly by considering various cases: fans of CDT give decision problems in which EDT seems to go wrong—like *Newcomb*—and fans of EDT do the same in the opposite direction.⁹ Recently, however, Andrew Bacon (2022) and Michael Nielsen (2023) have shown that CDT can be derived from more basic assumptions. In Bacon’s case, he shows that CDT is the only suppositional decision theory that tells you to maximize the expectation of a quantity called *actual value*—a quantity that’s played an important role in decision theory since its inception (Bacon, 2022, p. 617).¹⁰ Similarly, Nielsen shows that CDT is the only suppositional decision theory that respects the so-called *Value of Knowledge* principle—a widely-accepted principle which says that it’s always good to collect free information before making a decision.¹¹ Thus, in light of these results, philosophers who reject CDT seem like they’re committed to rejecting more than is initially obvious: they

⁶Although see Herrmann and Rothfus (forthcoming, §2.2), who’ve recently argued that there’s a sense in which EDT also says that you should defer to expert opinions—provided the expert in question is *your current evidence*. While I think this is an interesting idea, it’s quite different from the sense in which I use the term ‘deference to experts’ in this paper.

⁷Note, however, that this is not the only way in which my results can be read. On the contrary, as an anonymous reviewer mentions, some philosophers believe that deference principles like Reflection should *not* hold in full generality. Thus, an alternative way of reading my results is as constituting an argument *against* Stalnakerian CDT, since that theory requires unconditional deference to experts. See, e.g., Kadane, Schervish, and Seidenfeld (2008) or Ahmed and Salow (2019) for further, helpful discussion of this and related ideas.

⁸For a nice discussion of the history, see Joyce (1999, pp. 1–8). By ‘right answer’, I mean, of course, the two-boxing answer. CDT gives this answer; EDT does not.

⁹In the latter case, for example, see Ahmed (2013, pp. 290–96); Ahmed (2014a, pp. 666–76), or Spencer and Wells (2019, p. 34).

¹⁰Bacon credits one of his main results to unpublished work of Snow Zhang. See the Appendix for further details

¹¹As I’ll explain in §6, there’s a tight connection between the Value of Knowledge principle, on the one hand, and certain principles of expert deference, on the other. Thus, there’s also a tight connection between Nielsen’s results and mine. In fact, Nielsen’s paper is one of the key sources of inspiration for this paper.

must reject the notion of actual value, in light of Bacon’s results, and the Value of Knowledge principle, in light of Nielsen’s.¹²

My results are meant to constitute an argument for CDT in a similar spirit. They show, in effect, that anyone who rejects CDT is plausibly committed to rejecting several widely-accepted principles of expert deference—at least in the context of decision-making. The upshot is that the costs associated with rejecting CDT are arguably higher than fans of EDT, and other philosophers, might have initially expected. And the benefits of accepting CDT are higher as well.

In §2, I’ll introduce one well-known class of expert deference principles—the *Reflection* principles—and I’ll say how these principles can be motivated using standard EU theory. In §3 I’ll sketch the main shortcoming of EU theory, and I’ll introduce the notion of a suppositional decision theory, which gets around this shortcoming. In §4 I’ll ask whether the arguments tying Reflection to decision-making work when we transition to suppositional decision theories, and I’ll state the first technical result of this paper. This is: the only suppositional decision theory that says you should defer to experts, in the sense described by Reflection, is a particular version of CDT. In §5, I’ll motivate generalizations of Reflection—namely, New Reflection and the Trust principles—and I’ll state results that are analogous to the one given in §4. Once again, we can show that the only suppositional decision theory that respects New Reflection, or Trust, is a particular version of CDT. §6 concludes by drawing out a connection between my results and Nielsen’s (2023), and by sketching a generalization of his main result. Proofs of my results can then be found in the Appendix.

2 Expert Deference in Decision-making—A First Pass

Let’s begin with some set-up.

First, let $\mathcal{W}$ be the set of all your epistemically possible worlds (a set that I’ll assume to be finite, mostly for simplicity). Let $\mathbb{P}(\mathcal{W})$ be the set of all subsets of $\mathcal{W}$, the elements of which we call *propositions*. Let $\Delta(\mathcal{W})$ be the set of all probability functions defined over $\mathbb{P}(\mathcal{W})$, and let c be *your* probability function in particular. Throughout, I’ll call c your *credence function*. It measures the degree to which you believe each proposition $X \subseteq \mathcal{W}$.

Next, let $\mathcal{P}$ stand for the definite description ‘the expert’s credence function’ (whatever it is).¹³ Formally, $\mathcal{P} : \mathcal{W} \rightarrow \Delta(\mathcal{W})$ is a function that takes a possible world, w , as its argument, and maps this to a probability function, $\mathcal{P}_w \in \Delta(\mathcal{W})$, which represents the expert’s credences at w . For example, if $\mathcal{P}$ is the definite

¹²Indeed, some evidential decision theorists say exactly this. See, e.g., Ahmed and Spencer (2020, p. 1157) and Ahmed (2021, p. 59) for rejections of the notion of actual value, and Ahmed and Salow (2019, p. 348) for a rejection of the Value of Knowledge principle.

¹³By the way, when I say ‘the expert’ in this paper, I mean to denote some particular expert that’s salient in the context. Obviously, there are multiple people (things, etc.) deserving of the honorific ‘expert’. But it’s a tricky matter to say how expert deference works when there are multiple experts, each of whom is equally worthy of deference. See, e.g., Gallow (2018) for relevant discussion.

description ‘Bob’s credence function’, then in the *Newcomb* problem, every world w where there’s a million dollars in the opaque box is such that $\mathcal{P}$ maps w to a probability function satisfying $\mathcal{P}_w(\text{Million}) = 1$. Conversely, at every world v where there’s *not* a million dollars in the opaque box, $\mathcal{P}$ maps v to a probability function satisfying $\mathcal{P}_v(\text{Million}) = 0$.

Now, given the function $\mathcal{P}$, we can define a class of propositions saying that the expert has this-or-that probability function. To do so, we collect together all the worlds $w \in \mathcal{W}$ such that $\mathcal{P}$ outputs a certain probability function when given that world as an argument. Formally:

$$\langle \mathcal{P} = p \rangle := \{w \in \mathcal{W} : \mathcal{P}_w = p\}$$

Thus, the proposition $\langle \mathcal{P} = p \rangle$ says that the expert’s probability function is $p \in \Delta(\mathcal{W})$. And once we have propositions like this in place, we can define our first class of expert deference principles. As I said before, these are the principles known as *Reflection* principles:

Reflection A probability function c *reflects* the expert’s probabilities $\mathcal{P}$ if and only if, for all propositions $X \in \mathbb{P}(\mathcal{W})$ and for all values p of $\mathcal{P}$:¹⁴

$$c(X \mid \langle \mathcal{P} = p \rangle) = p(X)$$

Thus, Reflection says that, conditional on the proposition that the expert’s credences are given by the probability function p , your credence in any proposition should match the expert’s credence in that proposition. (Note here that, given the standard ratio formula for conditional probability, the quantity $c(X \mid \langle \mathcal{P} = p \rangle)$ is defined only when $c(\langle \mathcal{P} = p \rangle) > 0$. However, in what follows, I will assume that all conditional probabilities under discussion *are* defined.)

The version of Reflection just stated entails another, useful version of that principle—one that will be helpful to have in hand later on. This version says that, if your credences reflect those of the expert, then your credence in any proposition X can be expressed as a *mixture* of the expert’s credences, in the sense that there exist weights $\lambda_i \in [0, 1]$, with $\sum_i \lambda_i = 1$, such that $c(X) = \sum_i \lambda_i \cdot p_i(X)$ for any X . Indeed, in the case of Reflection, these weights are just your credences $c(\langle \mathcal{P} = p_i \rangle)$. So, Reflection implies the following:¹⁵

$$c(X) = \sum_i c(\langle \mathcal{P} = p_i \rangle) \cdot p_i(X)$$

¹⁴The version of Reflection I’ve given here is sometimes known as *Global Reflection*. There’s also a corresponding *local* version of Reflection, which is implied by the global version. See Gallow (2023, p. 2754) for details. Where possible, I’ll stick with the global versions of expert deference principles in this paper.

¹⁵To see how we can derive this version of Reflection from our original version, just note that, since the propositions $\langle \mathcal{P} = p_i \rangle$ form a partition, the law of total probability tells us that, for any X , $c(X) = \sum_i c(\langle \mathcal{P} = p_i \rangle) \cdot c(X \mid \langle \mathcal{P} = p_i \rangle)$. Then, by Reflection, we have $c(X \mid \langle \mathcal{P} = p_i \rangle) = p_i(X)$, for each p_i . Thus, substituting the latter terms into the equation above gives us the expectational version of Reflection.

In words: your credence in any proposition X is your *expectation* of the expert's credence in X .

Many well-known principles in philosophy count as Reflection principles, in these senses. To illustrate, take David Lewis's (1980) *Principal Principle*. Let $\langle Ch = ch \rangle$ be the proposition saying that the objective chances are given by the probability function ch . Then, the Principal Principle says (approximately):¹⁶

$$c(X \mid \langle Ch = ch \rangle) = ch(X)$$

So, according to the Principal Principle, you should defer to the objective chances.

Principles like this one have been widely defended in the literature. One common argument in their favor (as I've said) ties them to rational decision-making. Lewis himself defended his Principal Principle along these lines, for example, saying that "It is reasonable to let one's choices be guided... by one's firm opinions about objective chances or, when firm opinions are lacking, by one's degrees of belief about chances" (p. 287).

Thus, to turn remarks like these into a concrete argument, let me introduce a few key ideas from expected utility (EU) theory. First, then, suppose you have a collection of *acts* or *options* between which you can choose, $\mathcal{A} = \{A_1, \dots, A_n\}$. Suppose there's also a collection of *states of the world*, which you believe may or may not obtain, $\mathcal{S} = \{S_1, \dots, S_m\}$. I will assume throughout that both acts and states are propositions that partition the set $\mathcal{W}$; in particular, the former are the finest-grained propositions you believe you can *make* true at will.¹⁷ I will also assume that the state-propositions are grained sufficiently finely so that, for any act-state intersection $A \cap S$, the truth of such a proposition settles everything that you care about. (For this reason, in fact, I'll often refer to such intersections as *outcomes*. Note that I'll also frequently drop subscripts to ease notation. That is, I'll write, e.g., ' $A \cap S$ ' in place of ' $A_i \cap S_j$ '—as I did above.)

Now, let u be your *utility function*: a function which takes each act-state intersection, $A \cap S$, and maps it to a real number, $u(A \cap S)$, which represents the amount of utility you'd receive if this proposition were true. Then, EU theory says that the *expected* utility (EU) of choosing an option A , calculated according to c and u , is the following:¹⁸

$$V(c, A, u) := \sum_i c(S_i) \cdot u(A \cap S_i)$$

¹⁶This isn't *exactly* Lewis's Principal Principle. Instead, Lewis's official version of the principle has a clause saying what you should do in cases where you have "inadmissible information" (see, e.g., his 1980, pp. 272–73 for discussion). Since inadmissible information won't play a role in what's to come, however, I'll ignore that extra clause here.

¹⁷Cf. Jeffrey (1983, pp. 83–85).

¹⁸The ' V ' notation I employ here follows Bacon (2022, p. 626) and Nielsen (2023, p. 68).

So, the expected utility of an option, A , is a probability weighted average of utilities.¹⁹ And then, given this definition, the core idea underpinning EU theory is the following (below, a *decision problem* refers to any partition of option-propositions):

Maximize Expected Utility Whenever you’re facing a decision problem $\mathcal{A}$, you should choose an option, $A \in \mathcal{A}$, that *maximizes* your EU.

In a slogan, EU theory says that you should always choose an option whose outcome you *expect* to be best.

Now, the quantity $V(c, A, u)$ above is the benefit you expect to receive, if *you* were to choose A . However, what about the expected benefit of letting the expert, $\mathcal{P}$, decide on your behalf? Following Dorst (2019, p. 605) and Dorst et al. (2021, p. 105), I’ll now say that you *value* an expert’s opinions, in a given decision problem, just in case you think that deferring to that expert, in that problem, would leave you at least as well off as would choosing on your own. I’ll then say that you *value* an expert’s opinions simpliciter just in case you value their opinions in *every* decision problem. More precisely, we have the following:²⁰

Value A probability function c *values* the expert’s probabilities $\mathcal{P}$ if and only if there exist weights $\lambda_i \in [0, 1]$, with $\sum_i \lambda_i = 1$, such that $c(\cdot) = \sum_i \lambda_i \cdot p_i(\cdot)$ (where the p_i are $\mathcal{P}$ ’s values); and furthermore, for all decision problems $\mathcal{A}$ and utility functions u :

$$\sum_i \lambda_i \cdot \max_j V(p_i, A_j, u) \geq \max_j V(c, A_j, u)$$

In words, this says (i) that c can be represented as a mixture of the expert’s possible probability functions, and (ii) the corresponding weighted sum of the experts’ best options is always at least as great as the value of *your* best option. Indeed, in the special case of Reflection, the relevant weights are simply $c(\langle \mathcal{P} = p_i \rangle)$. And so, in that case, the inequality given in Value can be stated

¹⁹The version of EU theory I’ve given here is slightly non-standard. In particular, in the classical version of this theory due to Savage (1954, pp. 13–17), acts are not propositions, but rather *functions* from states to “consequences” (roughly, propositions corresponding to what I’ve called outcomes). The reason I’ve adopted this slightly unusual version of EU theory here, however, is that it’s much closer to my set-up for suppositional decision theories. As we’ll see later, for example, acts are considered to be propositions in the suppositional decision theory setting, since those theories require you to weigh options by *supposing* that they’re chosen; and we usually think of the object of a supposition as a proposition (not a function). Nothing much turns on the formal set-up for EU theory given here, however; everything I do in this section could equally well be done with Savage’s original version. Thanks to an anonymous referee for very useful feedback on this topic, and for pushing me to clarify my set-up for EU theory.

²⁰Dorst et al. (2021, pp. 104–5) give a slightly different definition of Value; the definition I give here is much closer to Nielsen’s formulation of the Value of Knowledge principle (2023, p. 72). As I discuss in the Appendix, however, it’s possible to give versions of my results using a definition of Value much closer to Dorst et al.’s.

alternatively as follows:

$$\sum_i c(\langle \mathcal{P} = p_i \rangle) \cdot \max_j V(p_i, A_j, u) \geq \max_j V(c, A_j, u)$$

Thus, the expected utility of choosing on your own is never greater than the expected value of choosing, using the expert’s probabilities. (That said, it will be useful to have the more general, “mixture definition” of Value in hand for later on.)

Now, with this definition of Value in place, we can state a first result linking Reflection to decision-making. This result was first proved by Brian Skyrms:²¹

Fact 1 (Skyrms, 1990, pp. 97-98). *For all probability functions c and expert functions $\mathcal{P}$, if c reflects $\mathcal{P}$, then c values $\mathcal{P}$.*

In rough terms, this says that, if you recognize someone as an expert, in the sense described by Reflection, then you expect that deferring to that person about what to do will always leave you at least well off as would choosing on your own. Thus, given what it plausibly *means* for someone to count as an expert, this seems like a point in favor of Reflection.

3 Suppositional Decision Theories

Although Fact 1 seems to provide strong support for Reflection, it is, unfortunately, not general enough. One issue is that it relies on the tools of EU theory. But as I mentioned before, many philosophers reject EU theory, in light of decision problems like the following—this one is from Joyce (1999):

The Extortionist. Suppose you have just parked in a seedy neighborhood when a man approaches and offers to “protect” your car from harm for \$10. You recognize this as extortion and have heard that people who refuse “protection” invariably return to find their windshields smashed. Those who pay find their cars intact. You cannot park anywhere else because you are late for an important meeting. It costs \$400 to replace a windshield. Should you buy “protection”? (pp. 115–16)

Here is a table, which summarizes your decision:

	<i>Smashed windshield</i>	<i>No smashed windshield</i>
<i>Pay</i>	−\$410	−\$10
<i>Don’t pay</i>	−\$400	\$0

Table 1: *The Extortionist*

²¹Huttegger (2013b, pp. 286–88) shows that the converse direction of Fact 1 also holds, provided the expert in question is “immodest”—see §5 for further details about “immodesty”.

Now, for simplicity, let's assume that you value dollars linearly (with the obvious choice of units), so that every dollar is worth just as much to you as the last. Let's also assume that you're a miser, so that money is all you care about. Then, it should be clear that $-\$410$ is a worse outcome than $-\$400$; and $-\$10$ is a worse outcome than $\$0$. A natural application of EU theory thus tells you that not paying is better than paying. After all, just look at the table: if the *Smashed windshield* state obtains, then you're better off by $\$10$ by refusing to pay. And the same thing goes if the *No smashed windshield* state obtains. Consequently, not paying seems to leave you better off than paying, *no matter how the world turns out to be*. In decision theorists' jargon: not paying *dominates* paying.

But of course, the reasoning we just employed is absurd. As Joyce himself describes things:

[In *The Extortionist*] Your choice has a *direct influence* on the state of the world; refusing to pay *makes it likely* that your windshield will be smashed while paying *makes this unlikely*. The extortionist is a despicable person, but he has you over a barrel and investing a mere $\$10$ now saves $\$400$ down the line. You should pay now (and alert the police later). (p. 116, with emphasis added)

Thus, the issue with standard EU theory—as Joyce's remark evinces—is that it doesn't take into account how the states of the world, on which the outcomes of your options depend, can themselves depend on your choices. The result is that, in *The Extortionist*, EU theory delivers a patently absurd result—it says that not paying is better than paying, when the opposite is plainly true.

Faced with examples like this one, the most common reaction among philosophers has been to say that EU theory simply doesn't apply when there's a dependence between your options and the world's state.²² This, in turn, solves the immediate problem raised by cases like *The Extortionist*, but only at the cost of severely restricting the scope of EU theory. What we'd like, then, is some *generalization* of EU theory, which applies even in cases where there is a dependence between your choice and the state of the world. The question now is: How exactly are we to do this?²³

Suppositional decision theories are one proposal for how to make this generalization.²⁴ They agree with EU theory whenever there *isn't* a dependence

²²See, e.g., Jeffrey (1983, pp. 83–85) for relevant discussion.

²³The historical framing I give here follows, e.g., Joyce (1999, pp. 1–8) and Ahmed (2014b, pp. 35–53). As an anonymous referee notes, however, an alternative framing is available. Indeed, there is some evidence to suggest that authors like Savage (1954) conceived of their theory as a causal decision theory all along. In particular, faced with an example like Joyce's *Extortionist*, these authors might say that this problem has really been *misdescribed*, because the states have not been chosen appropriately to be causally independent of options (see, e.g., Jeffrey (1983, pp. 21–22) for relevant discussion). On this reading, then, EDT was really introduced as an alternative to CDT, rather than both of these theories being introduced as generalizations of standard EU theory.

²⁴However, suppositional decision theories are arguably not the *first* proposal meant to solve the issue posed by cases like *The Extortionist*. See Fishburn (1964).

between your options and the states of the world. But they also say what you should do in cases where this kind of dependence obtains. The basic idea underpinning these theories is that, rather than choosing an option that maximizes EU, you should choose an option that maximizes EU *under the supposition that it's chosen*.

There are different ways of making this idea formally precise; but to spell out any of them, we need to introduce the notion of a *supposition function* (Joyce, 1999, p. 184; Elga, 2022, p. 204; Nielsen, 2023, p. 68). In formal terms, this is a function $s : \Delta(\mathcal{W}) \times \mathbb{P}(\mathcal{W}) \rightarrow \Delta(\mathcal{W})$, which takes a probability function, c , and a proposition, X , and maps these to a probability function, $s(c, X)$, which is such that $s(c, X)(X) = 1$. Informally, we can think of $s(c, X)$ as the probability function you get by starting with c , and then *supposing* the truth of X .

For example, one kind of supposition function corresponds to *conditionalization*.²⁵ This says, for any proposition Y :

$$s(c, X)(Y) = c(Y \mid X).$$

In the literature, it's common to think of conditionalization as corresponding to *indicative*, or *matter of fact*, supposition.²⁶ Thus, supposing X by conditionalizing is like saying to yourself: “Suppose X is, *as a matter of fact*, true. Then, how likely is it that such-and-such is also true?”

Another important supposition function corresponds to what Lewis (1976) called *imaging*. To define it, we'll need to introduce the notion of a *selection function* (Stalnaker, 1968, p. 34; Gärdenfors, 1982, pp. 749–51). For present purposes, a selection function is a function $f : \mathbb{P}(\mathcal{W}) \times \mathcal{W} \rightarrow \Delta(\mathcal{W})$, which takes a (non-empty) proposition X and a world w as arguments, and maps these to a specific probability function, $f(X, w)$, which concentrates all its probability on a subset of the X -worlds, so that $f(X, w)(X) = 1$. Intuitively, we can think of these X -worlds as the “closest” X -worlds to w .

Now, given the choice of a selection function f , we can define *the image of c on X* —written ' c^X '—as follows, for any proposition Y :

$$s(c, X)(Y) = c^X(Y) := \sum_w c(w) \cdot f(X, w)(Y).$$

²⁵There are some complications here. Strictly speaking, the definition of a supposition function I gave a moment ago requires these functions to be *total*. In the case of conditionalization, however, the corresponding definition must be partial, since $c(\cdot \mid X)$ is undefined when $c(X) = 0$. Thus, one thing we could do here is expand our definition of a supposition function to allow such functions to be partial. But doing so would severely complicate the proofs of my results in the Appendix. Thus, what I'll do instead is retain the official definition of ‘supposition function’, and treat conditionalization as a merely “honorary” supposition function in the paper's main body. For present purposes, this doesn't affect any of my arguments, since (i) I assumed in §2 that all conditional probabilities under discussion are defined, and (ii) my objections to EDT go through anyway: as we'll see below, we can come up with concrete decision problems showing that EDT will fail to defer to an expert. (Note also that, in this respect, my treatment of conditionalization and EDT exactly parallels Nielsen's (2023).)

²⁶See, e.g., Joyce (1999, pp. 195–96) or Bradley (2017, p. 95–96) for discussion.

In this case, it's common to think of c^X as corresponding to *subjunctive*, or *counterfactual*, supposition. So, supposing X by imaging is like saying to yourself: "Suppose X were true. Then how likely is it that such-and-such *would be* true?"²⁷ (In fact, if we assume that f concentrates all its probability on a single possible world, when given X and w as arguments, then it turns out that $c^X(Y) = c(X \Box \rightarrow Y)$. In other words, your credence in any proposition, imaged on X , is equal to your credence in a corresponding counterfactual, whose antecedent is X . See Lewis (1976, pp. 310–11), Lewis (1981, p. 23), or Joyce (1999, pp. 196–97) for additional discussion.)

Now, according to any suppositional decision theory, the *suppositional expected utility* (SEU) of an option, A , available to you in a decision problem is the following:

$$V_s(c, A, u) = \sum_i s(c, A)(S_i) \cdot u(A \cap S_i)$$

On its face, this equation looks very similar to the one we used to define the expected utility of A , according to standard EU theory. The chief difference is that, rather than using your "ordinary" credence function c to calculate the expectation, we now use your credence function *under the supposition that you choose A* .

The equation above can then be precisified, depending on how we understand the supposition function, s , in question. For example, if we take s to be conditionalization, then the suppositional decision theory we get back is none other than EDT (Jeffrey, 1965, 1983):

$$V_s(c, A, u) := \sum_i c(S_i | A) \cdot u(A \cap S_i)$$

Thus, EDT says that you should choose an option, A , whose outcome you expect to be best, while supposing, *as a matter of fact*, that A is chosen. Alternatively: you should choose an option A that provides the best *evidence* for an outcome you value (which is why this theory is known as *evidential* decision theory).

In contrast, if we take s to correspond to imaging, then the suppositional decision theory we get back is a specific version of CDT:

$$V_s(c, A, u) = \sum_i c^A(S_i) \cdot u(A \cap S_i)$$

This version of CDT was first described by Stalnaker (1981, pp. 151–52), and developed in detail by Gibbard and Harper (1978),²⁸ Sobel (1994), and Joyce

²⁷Again, this characterization of imaging as corresponding to subjunctive supposition is standard. See especially Joyce (1999, pp. 196–98) for helpful discussion.

²⁸The version of CDT given by Stalnaker and Gibbard and Harper originally looks a little different to the one I've just presented. Specifically, they define: $V_s(c, A, u) = \sum_i c(A \Box \rightarrow S_i) \cdot u(A \cap S_i)$. However, this version of CDT is equivalent to the one given in the main text, if we assume that selection functions concentrate all their probability on a single world (as I

(1999). I'll refer to it throughout as *Stalnakerian CDT*. Roughly, it says that you should choose an option, A , whose outcome you expect to be best *under the subjunctive supposition that you choose A* . Then, since there's a tight connection between subjunctive supposition and causation—for example, when you want to know whether X caused Y , you'll often ask yourself questions like 'Suppose X hadn't happened. Then, *would* Y have happened?'—we can think of this theory as saying that you should choose an option A that you expect to *cause* the best results.²⁹

Now, there are other versions of CDT, in addition to this one. But I won't spend much time on those other theories here. Indeed, the results I give below don't apply to them. And in my view, this gives us a reason to think that Stalnakerian CDT is preferable.³⁰

Before we move on, let me note one final thing. This is that the version of CDT sketched above, as well as EDT, adequately handle cases like Joyce's *Extortionist*. To see this, consider how things go in the case of EDT. In that case, since there's a strong correlation between your choice and the state of the world, we have that $c(\text{No smashed windshield} \mid \text{Pay}) \approx 1$, and $c(\text{Smashed windshield} \mid \text{Don't pay}) \approx 1$. So, conditional on choosing *Pay*, you're almost certain that *No smashed windshield* obtains; and conditional on choosing *Don't pay*, you're almost certain that *Smashed windshield* obtains. Thus, according to EDT:

$$\begin{aligned} V_s(c, \text{Pay}, u) &= c(\text{Smashed windshield} \mid \text{Pay}) \cdot -410 \\ &\quad + c(\text{No smashed windshield} \mid \text{Pay}) \cdot -10 \\ &\approx 0 \cdot -410 + 1 \cdot -10 \\ &= -10, \\ V_s(c, \text{Don't pay}, u) &= c(\text{Smashed windshield} \mid \text{Don't pay}) \cdot -400 \\ &\quad + c(\text{No Smashed windshield} \mid \text{Don't pay}) \cdot 0 \\ &\approx 1 \cdot -400 + 0 \cdot 0 \\ &= -400. \end{aligned}$$

mentioned before). Again, see Lewis (1976, pp. 310–11), Lewis (1981, p. 23), or Joyce (1999, pp. 196–97) for details.

²⁹For more on the relation between subjunctive supposition and causation, and how this relates to CDT, see the discussions of Lewis (1981), Joyce (1999), Hitchcock (2013), Kment (2023), McNamara (2024), and especially Gallow (2024). Note also that not every philosopher agrees that Stalnakerian CDT is best interpreted as a *causal* decision theory. See, e.g., Dorr (2016) and Hedden (2023) for more on this.

³⁰For example, another well-known version of CDT is the "Bayesian" version, given by Skyrms (1980, 1982, 1984) and Lewis (1981). These versions appeal to a partition of so-called *causal dependency hypotheses*. In particular, if $K_1, \dots, K_m$ is a partition of such hypotheses, then the Skyrms/Lewis version of CDT says: $V_s(c, A, u) = \sum_i \sum_j c(K_i) \cdot c(S_j \mid A \cap K_i) \cdot u(A \cap S_j)$. Lewis claims (1981, p. 313) that this is equivalent to the imaging version of CDT given in the main text, since we can simply set $c^A(S_j) = \sum_i c(K_i) \cdot c(S_j \mid A \cap K_i)$. However, in the definition of imaging I gave, the selection function f is chosen independently of your credence function. And this property is not satisfied in Lewis's definition of imaging—so the two versions of CDT diverge. As it turns out, this is a key property of imaging, needed to derive my result. See Gärdenfors (1982) and Bacon (2022) for closely related discussion.

Then, since the former quantity is greater than the latter, EDT says that you should pay the extortionist. (Note also that CDT gives the same answer in this case—although I won't spell out *how* CDT gets that answer here.)

Thus, suppositional decision theories get around the main issue faced by standard EU theory—namely, the issue of what to say when there's a dependence between your options and the states of the world. However, different suppositional decision theories disagree about what *kinds* of dependencies matter for decision-making. One upshot of this, as we'll see, is that not all suppositional decision theories agree that it's always sensible to defer to an expert while deciding.

4 Revisiting Reflection

We've now moved past the setting of standard EU theory, and into the setting of suppositional decision theories. And what we'd now like to know, having done this, is whether there's an analogue of Fact 1 that can be proved in this latter setting. It turns out that there is; but to state it, we will need to alter one of our preceding definitions.

The definition in question is the definition of Value. In particular, we now need to state it in terms of suppositional decision theories (since, earlier, it was stated in terms of standard EU theory). To distinguish this new notion of Value from what we had before, then, I will call it *Suppositional Value*. The definition we need is the following:

Suppositional Value Let s be a supposition function. Then, a probability function c *suppositionally values* the expert's probabilities $\mathcal{P}$ if and only if, there exist weights $\lambda_i \in [0, 1]$, with $\sum_i \lambda_i = 1$, such that $c(\cdot) = \sum_i \lambda_i \cdot p_i(\cdot)$ (where the p_i are $\mathcal{P}$'s values); and furthermore, for all decision problems $\mathcal{A}$ and utility functions u :

$$\sum_i \lambda_i \cdot \max_j V_s(p_i, A_j, u) \geq \max_j V_s(c, A_j, u)$$

Thus, you suppositionally value an expert's opinions just in case (i) your credence function can be represented as a mixture of the expert's possible credence functions, and (ii) the corresponding weighted sum of the experts' best options is at least as great as the value of *your* best option. Additionally, in parallel to what we saw in §2 above, in the case of Reflection, the relevant weights are simply $c(\langle \mathcal{P} = p_i \rangle)$, which means that you suppositionally value the expert's opinions if and only if:

$$\sum_i c(\langle \mathcal{P} = p_i \rangle) \cdot \max_j V_s(p_i, A_j, u) \geq \max_j V_s(c, A_j, u)$$

Thus, with this definition in hand, we can state our first analogue of Fact 1, in the suppositional decision theory setting:

Fact 2. Let s be a supposition function. Then, the following are equivalent:

- (i) s determines a Stalnakerian causal decision theory.
- (ii) For all probability functions c and expert functions $\mathcal{P}$, if c reflects $\mathcal{P}$, then c suppositionally values $\mathcal{P}$, with respect to the suppositional decision theory determined by s .

Informally: if you suppositionally value the expert, then you think that using their probabilities to make your decision will always leave you at least as well off as would choosing on your own—if and only if the suppositional decision theory in question is a Stalnakerian CDT. To put this differently: out of all the suppositional decision theories that are out there, *only* Stalnakerian CDT preserves the idea that it’s always good to defer to an expert when you’re deciding, in the sense defined by Reflection.

A proof of Fact 2 can be found in the Appendix. (Actually, as you’ll see there, this result really follows from a more general result, stated in the next section.) I think it constitutes a strong argument for Stalnakerian CDT. After all, other suppositional decision theories say that, even if you recognize someone as an expert now, because you reflect their opinions, you can sometimes decline the opportunity to defer to them while deciding. As Dorst et al. said in §1, however, it’s arguably part of what it *means* to call someone an expert that you’re disposed to defer to them while choosing.

To illustrate the problem for non-Stalnakerian suppositional decision theories, let’s take another look at the *Newcomb* problem.³¹ In that problem, EDT says that one-boxing is the best option. After all, there’s a strong correlation between taking only the opaque box, and the million dollars being present; and there’s a strong correlation between taking both boxes, and the million dollars being absent. Thus:

$$\begin{aligned} V_s(c, \text{One-boxing}, u) &= c(\text{Million} \mid \text{One-boxing}) \cdot 1,000,000 \\ &\quad + c(\text{No Million} \mid \text{One-boxing}) \cdot 0 \\ &\approx 1 \cdot 1,000,000 + 0 \cdot 0 \\ &= 1,000,000. \end{aligned}$$

A similar calculation shows that $V_s(c, \text{Two-boxing}, u) \approx 1,000$. So, in a straight choice between one-boxing and two-boxing, EDT says that you should choose the former option (and thereby give yourself *evidence* that you’re a millionaire).

However, what does EDT say about deferring to Bob in *Newcomb*? To see what it says, first recall that Bob would recommend that you take both boxes

³¹Note here that Stalnakerian CDTs say that deferring to Bob in *Newcomb* is *equally* as good as choosing on your own. The reason for this, however, is that, according to these theories, two-boxing *dominates* one-boxing. As it turns out, decision problems with this feature are the only kind of case in which deferring to an expert isn’t deemed *strictly* better than choosing on your own, according to Stalnakerian CDTs.

in this case, no matter what it is that he's seen. In particular, at every world in which there's a million dollars in the opaque box, Bob's probability function, p_1 , satisfies $p_1(\text{Million}) = 1$. And so, according to EDT:

$$\begin{aligned}
 V(p_1, \text{One-boxing}, u) &= p_1(\text{Million} \mid \text{One-boxing}) \cdot 1,000,000 \\
 &\quad + p_1(\text{No Million} \mid \text{One-boxing}) \cdot 0 \\
 &= p_1(\text{Million}) \cdot 1,000,000 + p_1(\text{No Million}) \cdot 0 \\
 &= 1 \cdot 1,000,000 + 0 \cdot 0 \\
 &= 1,000,000 \\
 V(p_1, \text{Two-boxing}, u) &= p_1(\text{Million} \mid \text{Two-boxing}) \cdot 1,001,000 \\
 &\quad + p_1(\text{No Million} \mid \text{Two-boxing}) \cdot 1,000 \\
 &= p_1(\text{Million}) \cdot 1,001,000 + p_1(\text{No Million}) \cdot 1,000 \\
 &= 1 \cdot 1,001,000 + 0 \cdot 1,000 \\
 &= 1,001,000
 \end{aligned}$$

Thus, in the case in which Bob's seen a million dollars in the opaque box, we have $V(p_1, \text{Two-boxing}, u) > V(p_1, \text{One-boxing}, u)$. And similar calculations show that $V(p_2, \text{Two-boxing}, u) = 1,000 > V(p_2, \text{One-boxing}, u) = 0$, at all the worlds where Bob's spied nothing in the opaque box (and thus his credence function, p_2 , is such that $p_2(\text{Million}) = 0$). Consequently—and as we heard before—Bob would recommend that you take both boxes, no matter what it is that he's seen in the opaque box. (And this is true, even if he abides by EDT.)

But now, notice that your credence that Bob has the first kind of probability function, p_1 , is just your credence in the proposition *Million*. And your credence that he has the other kind of probability function, p_2 , is just your credence in the proposition *No Million*. Thus, the result is that, in any case in which $c(\text{Million}) < 0.999$, you will *not* suppositionally value Bob's credences. To illustrate, in the case in which your credences are such that $c(\text{Million}) = c(\text{No Million}) = 0.5$, it follows that the value you assign to deferring to Bob is the following:

$$\begin{aligned}
 0.5 \cdot V(p_1, \text{Two-boxing}, u) + 0.5 \cdot V(p_2, \text{Two-boxing}, u) &= 0.5 \cdot 1,001,000 + 0.5 \cdot 1,000 \\
 &= 501,000
 \end{aligned}$$

This is far lower than the value you assign to one-boxing straight away. So, the upshot is that, if you follow EDT's recommendation in *Newcomb*, then you'll decline the expert's advice, if you can get it.

To many of us, however, this seems absurd. After all, *by your own lights* Bob is someone with more evidence than you have. Indeed, he *knows* what's inside the opaque box. And so it seems crazy not to take advantage of his evidence, if you're given the opportunity to consult with him. Instead, to borrow a phrase from Lewis: it seems like EDT is here “commending an irrational policy of managing the news so as to get good news about matters which you have no control over...

[But] To decline the [expert’s advice] lest taking it brings bad news is to play the ostrich” (1981, pp. 5–8). I think Lewis is right about this.

5 Modest Experts

Once more: Fact 2 seems like a strong argument in favor of Stalnakerian CDT—and in turn, a strong argument against other suppositional decision theories. Additionally, Reflection principles are the most widely defended principles of expert deference in the literature.

At the same time, however, these principles have a strange upshot. Specifically, as Dorst (2019, p. 591) has recently emphasized, they don’t allow for a plausible form of *epistemic modesty*. To see what this means, first let p be the credence function of a particular expert (i.e., some value of the corresponding function $\mathcal{P}$). Then, ask yourself: Must this expert be *certain* of their own expertise? In other words, must they satisfy the following condition, in order to qualify as an expert in the first place?

$$p(\langle \mathcal{P} = p \rangle) = 1$$

In some cases at least, the answer seems to be ‘No’; for example, it doesn’t seem like better-informed-Bob must be certain that he’s deserving of the honorific ‘expert’, in order to qualify as an expert, in the relevant sense, in *Newcomb*. However, if that’s true, then we should allow that:

$$p(\langle \mathcal{P} = p \rangle) < 1$$

But if we do, then Reflection leads to inconsistencies.

To see this, just notice that, by Reflection, we should have:

$$c(\langle \mathcal{P} = p \rangle \mid \langle \mathcal{P} = p \rangle) = p(\langle \mathcal{P} = p \rangle) < 1$$

But of course, the probability axioms require that:

$$c(\langle \mathcal{P} = p \rangle \mid \langle \mathcal{P} = p \rangle) = 1$$

Thus, if we allow that experts can be “modest”, in the relevant sense, then Reflection principles conflict with the probability axioms. We thus have to give up on Reflection principles, or else give up on the rationality of epistemic modesty.

Faced with results like this one, some philosophers might give up on epistemic modesty. However, others will argue that we should retain epistemic modesty, and weaken the Reflection principles instead. A particularly well-known suggestion for how to do this comes from Hall (1994, p. 511), Lewis (1994, p. 487), and Elga (2013, p. 130). As they say, you shouldn’t defer to an expert’s *unconditional* credences, but instead to their conditional credences—*conditional on the proposition that they’re the expert*. The corresponding class of principles consists of ones I’ll call *New Reflection* principles:

New Reflection A probability function c *new reflects* the expert's probabilities, $\mathcal{P}$, if and only if, for all propositions X and for all values p of $\mathcal{P}$:

$$c(X \mid \langle \mathcal{P} = p \rangle) = p(X \mid \langle \mathcal{P} = p \rangle)$$

Equivalently, if we define $\mathcal{P}^*$ to be the definite description 'the *informed* expert's credences'—i.e., we let each value of $\mathcal{P}^*$ be a function $p^*(\cdot) = p(\cdot \mid \langle \mathcal{P} = p \rangle)$ —then we can state New Reflection as a version of Reflection (cf. Stalnaker, 2019, p. 100):

Informed Reflection A probability function c reflects the *informed* expert's probabilities, $\mathcal{P}^*$, if and only if, for all propositions X and for all values p^* of $\mathcal{P}^*$:

$$c(X \mid \langle \mathcal{P}^* = p^* \rangle) = p^*(X)$$

Thus, both of these principles say that you should defer to the expert's *informed* credences.³²

Unlike Reflection, New Reflection and Informed Reflection are consistent with epistemic modesty. After all, suppose that $p(\langle \mathcal{P} = p \rangle) < 1$. Then, by New Reflection:

$$c(\langle \mathcal{P} = p \rangle \mid \langle \mathcal{P} = p \rangle) = p(\langle \mathcal{P} = p \rangle \mid \langle \mathcal{P} = p \rangle) = 1$$

Thus, there's no conflict between New Reflection and the probability axioms, even though the expert in question is modest. (And the same goes, of course, for Informed Reflection.)

New Reflection/Informed Reflection also collapse to (ordinary) Reflection when the expert in question is *immodest*. To see this, just notice that, if $p(\langle \mathcal{P} = p \rangle) = 1$, then for any proposition X :

$$c(X \mid \langle \mathcal{P} = p \rangle) = p(X \mid \langle \mathcal{P} = p \rangle) = p(X)$$

In a real sense, then, these principles *generalize* Reflection. They get us everything the old principle had, and more as well.

However, does an analogue of Fact 2 hold when we move from Reflection to New Reflection (or Informed Reflection)? In other words, does CDT continue to say that you should defer to expert opinions, even when the expert in question is

³²Note, here, that we can use a similar trick to the one we used in §2 to show that New Reflection/Informed Reflection imply that your current credence in any proposition X should be your *expectation* of the informed expert's credence in X . Additionally, note that one way we can motivate these principles—aside from the fact that they accommodate epistemic modesty—is via Hall's (2004) distinction between *database* and *analyst experts*. As Hall says, a database expert is one who simply has more evidence than you have, while an analyst expert is one who's better at reasoning with your evidence. The former kind of expert, Hall contends, is naturally construed as an immodest expert, while the latter kind may be modest. Thus, by adopting New Reflection, we make room for this latter kind of expert. This also helps to explain why Informed Reflection is not *merely* the Reflection principle. On the contrary, with respect to that principle, 'The expert' can still refer to the modest expert; but the principle itself tells us only to defer to them *once they've learned that they're the expert*.

modest? The answer is that it does; indeed, it's the only suppositional decision theory that says this:

Fact 3. *Let s be a supposition function. Then, the following are equivalent:*

- (i) *s determines a Stalnakerian causal decision theory.*
- (ii) *For all probability functions c and informed expert functions $\mathcal{P}^*$, if c reflects $\mathcal{P}^*$, then c suppositionally values $\mathcal{P}^*$, with respect to the suppositional decision theory determined by s .*

Fact 3 is another bit of good news for fans of CDT. But—just like our original Reflection principle—New Reflection/Informed Reflection have been criticized. One thing that's been noted, for example—particularly by authors like Dorst (2019, p. 601), Dorst et al. (2021, p. 103), Levinstein (2023, p. 59), and Levinstein and Spencer (2024, p. 146)—is that it's possible to satisfy these principles even if you think an expert's unconditional credences are *anti-correlated with the truth*.

To see what this means, let's consider an example. Suppose there are just two possible worlds that you believe could be actual, namely w_1 and w_2 . Suppose also that, at w_1 , p_1 is the expert, and at w_2 , p_2 is the expert. Next, suppose that both p_1 and p_2 are modest—in fact, radically so. For concreteness, suppose that $p_1(\langle \mathcal{P} = p_1 \rangle) = 0.1$ and $p_1(\langle \mathcal{P} = p_2 \rangle) = 0.9$, and similarly for p_2 . Then, both experts here are extremely confident that the *other* probability function is the expert, and so their unconditional credence in this proposition is anti-correlated with the truth. (See below for a picture, which should make this set-up easier to grasp. The two nodes there represent the worlds w_1 and w_2 , respectively. And the arrows represent the credences that the experts, at those worlds, assign to the different worlds being actual.)

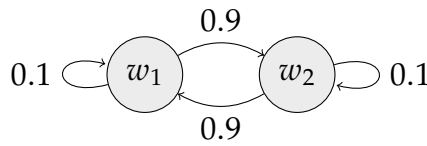


Figure 1: Anti-reliable Experts

Now, despite the fact that p_1 and p_2 are both (unconditionally) anti-correlated with the truth, you can satisfy New Reflection (and Informed Reflection) with respect to both of these functions. To see this, notice that:

$$c(\langle \mathcal{P} = p_1 \rangle \mid \langle \mathcal{P} = p_1 \rangle) = c(w_1 \mid w_1) = p_1(w_1 \mid w_1) = p_1(\langle \mathcal{P} = p_1 \rangle \mid \langle \mathcal{P} = p_1 \rangle) = 1$$

The same thing goes for the function p_2 . And so, you satisfy New Reflection with respect to both p_1 and p_2 . But it's questionable whether these functions are really deserving of the honorific 'expert'. As Dorst et al. (2021) put it, the issue

is that “you can [satisfy New Reflection] *even if you are certain that the expert is mistaken*” (p. 103, emphasis added).³³

In light of cases like this one, then, several authors have concluded that, while satisfying New Reflection is *necessary* for epistemic deference, it’s not entirely *sufficient* either.³⁴ In other words, these authors say that, in order to be deserving of the honorific ‘expert’, we shouldn’t allow that some candidate expert function can be (unconditionally) anti-correlated with the truth. Conversely, our original Reflection principles seemed sufficient for epistemic deference, but not, unfortunately, necessary, since they ruled out epistemic modesty. Thus, what we’re looking for, at this stage, is a kind of *Goldilocks principle*—one whose satisfaction is both *necessary and sufficient* for epistemic deference.³⁵

Dorst et al. (2021, p. 107) claim to have such a principle. They call it the principle of *Total Trust*.³⁶ To state it, we’ll need to introduce some additional notation. Thus, going forward, let’s write ‘ $\mathbb{E}_c$ ’ for expectation, calculated with respect to the probability function c .³⁷ Furthermore, let’s write $\langle \mathbb{E}_{\mathcal{P}}[\mathcal{X}] \geq x \rangle$ for the proposition that the expert expects the value of the random variable $\mathcal{X}$ to be greater than or equal to x . Formally speaking, this proposition is:

$$\langle \mathbb{E}_{\mathcal{P}}[\mathcal{X}] \geq x \rangle := \{w \in \mathcal{W} : \mathbb{E}_{\mathcal{P}_w}[\mathcal{X}] \geq x\}.$$

The Total Trust principle can then be stated:

³³This is worse in some cases than in others. For example, in discussing the idea of *objective chance*-qua-expert, Levinstein and Spencer (2024, pp. 145–46) say that, if we allow objective chance to be radically modest in this way, then it isn’t clear we’re talking about objective chance in the first place. After all, as Lewis himself implores: “Don’t call any alleged feature of reality ‘chance’ unless you’ve already shown that you have something, knowledge of which could constrain rational credence” (1994, p. 484). The problem, however, is that if chances can be radically modest, then it isn’t clear that chances themselves constrain rational credences; rather, it seems like the chances *conditionalized on the proposition that they’re the chances* do (as per New Reflection). But once more, this seems to violate Lewis’s own criterion, for calling something ‘objective chance’ in the first place.

³⁴See Dorst (2019, p. 603), Dorst et al. (2021, p. 103), Levinstein (2023, p. 60), and Levinstein and Spencer (2024, pp. 135–136). Also, as an anonymous reviewer notes, the kind of case I just described is related to cases given by Hild (1998). See Hild’s paper for an alternative solution.

³⁵Note that I’m not necessarily endorsing this line of argument. Instead, my aim here is only to motivate alternative principles of expert deference, in the way that Dorst and others do. Officially, I’m trying to remain neutral in this paper about which principles of expert deference are correct.

³⁶There are several principles in the Trust-category of expert deference principles. However, all of these principles are entailed by the Total Trust principle, described in the main text. See Dorst (2019) and Dorst et al. (2021) for further discussion.

³⁷In a bit more detail, the expectation of a (discrete) random variable $\mathcal{X}$, with respect to c , is defined as:

$$\mathbb{E}_c[\mathcal{X}] = \sum_x c(\mathcal{X} = x) \cdot x,$$

where x ranges over $\mathcal{X}$ ’s possible values. (Recall also that a *random variable* $\mathcal{X}$ is a function $\mathcal{X} : \mathcal{W} \rightarrow \mathbb{R}$.) Similarly, the conditional expectation of $\mathcal{X}$, given that the random variable $\mathcal{Y}$ takes on value y , is:

$$\mathbb{E}_c[\mathcal{X} \mid \mathcal{Y} = y] := \sum_x c(\mathcal{X} = x \mid \mathcal{Y} = y) \cdot x$$

Total Trust A probability function c *totally trusts* the expert's probabilities, $\mathcal{P}$, if and only if, for all random variables $\mathcal{X}$ and all real numbers x such that the relevant conditional expectation is defined:

$$\mathbb{E}_c[\mathcal{X} \mid \langle \mathbb{E}_{\mathcal{P}}[\mathcal{X}] \geq x \rangle] \geq x$$

Informally: conditional on the expert's expectation of $\mathcal{X}$ being at least x , your expectation of $\mathcal{X}$ should be at least x , too.³⁸

This principle will come up again in the next section. For now, however, just note that if we think of it as the correct principle of expert deference, then we can prove another result, relating this principle to suppositional decision theories. The following result is, in some sense, the main technical result of this paper, since it entails both Facts 2 and 3. You can find a proof of it in the Appendix:

Theorem 1 (Main Result). *Let s be a supposition function. Then, the following are equivalent:*

- (i) *s determines a Stalnakerian causal decision theory.*
- (ii) *For all probability functions c and all expert functions $\mathcal{P}$, if c totally trusts $\mathcal{P}$, then c suppositionally values $\mathcal{P}$, with respect to the suppositional decision theory determined by s .*

Thus, you think that deferring to the expert, in the sense described by Total Trust, will always leave you at least as well off as would choosing on your own, if and only if you abide by Stalnakerian CDT. (Note that it's also because of Total Trust that I stated the Suppositional Value principle in terms of general mixtures, rather than in terms of expectations. After all, as you'll see in the Appendix, if your credences satisfy Total Trust (but not the stronger Reflection), then your credence in any proposition is *not*, in general, your expectation of the expert's credence. It is, however, a *mixture* of the credences that you think the expert can have—it's just that the weights λ_i will not generally equal $c(\langle \mathcal{P} = p_i \rangle)$. I will say more about this fact in the next section, and, again, I will expand on it fully in the Appendix.)

This rounds out my discussion of principles of expert deference and suppositional decision theories. To sum up what we've established: it seems like a good idea to defer to the advice of an expert, when you're deciding; and there are various principles saying how you should do this. Each of these principles can be motivated using the tools of standard EU theory; but when we move into

³⁸Total Trust is related to, but strictly weaker than, the higher-order coherence principles of Gaifman (1986, Axiom VI) and Skyrms (1980, Constraint 2, p. 163). Those principles entail the martingale condition (see Gaifman's Theorem 2), which exhibits c as a mixture of the expert's credences with the specific weights $c(\langle \mathcal{P} = p_i \rangle)$. Total Trust, by contrast, requires merely that c lie somewhere in the convex hull of the expert's credences (see Lemma 1 in the Appendix). It is thus a consequence of those principles, not conversely. Thanks to an anonymous referee for this useful comparison.

the more general setting of suppositional decision theories, it turns out that only one such theory preserves the idea that it's always good to defer to experts when you have to make a choice. This, as we've heard, is Stalnakerian CDT. I think this constitutes a strong argument for that theory.

6 Expert Deference and The Value of Knowledge

Before we close, there's one last thing I want to make note of. This is the connection between my results, and an interesting result proved by Michael Nielsen (2023). As I said previously, what Nielsen shows is that Stalnakerian CDT is the only suppositional decision theory that satisfies the *Value of Knowledge principle*—and this, too, seems like a point in that theory's favor.³⁹

To see what this means, let me introduce one last bit of formalism. Thus, let $\mathcal{E} = \{E_1, \dots, E_n\}$ be a set of propositions, and suppose that, in a moment, you're going to learn that one of these propositions is true (and you're also going to update your credences by Bayesian conditionalization accordingly). Now, if the set of propositions $\mathcal{E}$ forms a partition, then we can state the Value of Knowledge principle as follows:

Value of Knowledge For all probability functions c , for all utility functions u , and for all decision problems $\mathcal{A}$:

$$\sum_i c(E_i) \cdot \max_j V(c(\cdot \mid E_i), A_j, u) \geq \max_j V(c, A_j, u).$$

Here's what this says, informally. If the propositions in $\mathcal{E}$ are ones that you can learn on some occasion—and if, again, these propositions form a partition, and you'll update your credences by conditionalization after learning the true such proposition⁴⁰—then the expected value of *first* learning the true member of $\mathcal{E}$, and then making your decision, is always at least as great as the expected utility of choosing straight away. It's a simple exercise to show that standard EU theory satisfies this principle, with respect to any partition.⁴¹

This principle seems plausible, on the face of it; indeed, some philosophers have argued that it's part of what it means to call something a *genuine learning*

³⁹The Value of Knowledge principle is also sometimes known as the Value of Information principle, or Good's principle (after I.J. Good (1967), who proved a version of it). Indeed, given that the principle doesn't really have anything to do with *knowledge*, 'Value of Information' is arguably more appropriate. Thanks here to an anonymous referee.

⁴⁰In this paper, I'll adopt the standard view that you should update your credences by conditionalization. This assumption is relatively uncontroversial in cases where the propositions you might learn form a partition. But it's more controversial in cases where this condition doesn't hold. For discussion, see, e.g., Schoenfield (2017) or Gallow (2021).

⁴¹There's a corresponding version of this principle for suppositional decision theories, which Nielsen (2023) defines. The only change we need to make is to replace $V(\cdot, \cdot, u)$ in the definition with $V_s(\cdot, \cdot, u)$ —i.e., the options are now evaluated using a suppositional decision theory, rather than standard EU theory. This is the version of the principle appealed to in Nielsen's result, stated below.

experience that you expect this principle to hold (Huttegger, 2013b, p. 285). Moreover, Skyrms (1990, pp. 97–98) and Huttegger (2013b, pp. 286–87) have shown that, given the tools of standard EU theory—and assuming that experts are immodest—the Value of Knowledge principle is *equivalent* to Reflection: the one holds if and only if the other does. Dorst et al. (2021, pp. 107–8) then generalize these results to their principle of Total Trust: they show that, if experts are modest, then satisfying (a generalized version of) the Value of Knowledge principle is equivalent to satisfying Total Trust. Thus, in a real sense, these results establish a tight connection between certain principles of expert deference and the Value of Knowledge. In effect, those principles are really one and the same.

Still, as Nielsen (2023) shows, the Value of Knowledge principle does not hold in general, once we transition from the setting of standard EU theory to the setting of suppositional decision theories.⁴² Instead, the only suppositional decision theory that satisfies the Value of Knowledge principle is Stalnakerian CDT:

Fact 4 (Nielsen, 2023, Theorem 1). *Let s be a supposition function. Then the following are equivalent:*

- (i) *s determines a Stalnakerian causal decision theory.*
- (ii) *The suppositional decision theory determined by s satisfies the Value of Knowledge principle.*

Once again, this result seems like good news for fans of Stalnakerian CDT. But at the same time, Nielsen himself says that it’s weaker than we might have hoped for. In particular, the Value of Knowledge principle, as stated

is qualified so as to apply only in certain ideal circumstances. First, *the principle applies only when the possible outcomes of knowledge seeking form a partition...* [But] In general, the possible outcomes of knowledge seeking do not form a partition, and when this occurs even ideal Bayesians can fail to value knowledge... The principle stated above sidesteps the important question of whether knowledge should be valued when the possible outcomes of learning are non-partitional. (p. 73, with emphasis added)

Interestingly, however, the issue of whether the Value of Knowledge principle holds when the propositions in $\mathcal{E}$ do *not* form a partition turns out to be intimately connected to the question of whether experts should be modest. In particular, Dorst et al. (2021, p. 125, n. 5) have recently argued that satisfying the Value of Knowledge when the relevant propositions do not form a partition is equivalent

⁴²Actually, this fact has been known for quite some time. See, e.g., Skyrms (1990) for discussion. Nielsen’s contribution is to show that the Value of Knowledge principle holds in the suppositional decision theory setting *only* if the decision theory in question is a Stalnakerian CDT.

to satisfying their principle of Total Trust.⁴³ In our setting, it's fairly easy to see why this should be. To do so, let the values of the definite description $\mathcal{P}$ be functions of the form $c(\cdot \mid E_i)$ —i.e., *your* credence function conditionalized on a piece of evidence in $\mathcal{E}$. Then, 'The expert's credence function' here is just *your* credence function, after some learning experience. And if you totally trust $\mathcal{P}$ —i.e., you view your future self as a kind of expert—then it follows from Theorem 1 that you'll think deferring to your future self's opinions is a good idea if and only if you abide by a Stalnakerian CDT. Interpreted in this way, then, what Theorem 1 shows is that an even more general version of Nielsen's important result holds. You can suppositionally value the outcomes of knowledge-seeking, even when these propositions do not form a partition, provided the suppositional decision theory in question is a Stalnakerian CDT.

Appendix

Proof of the Main Result

In this Appendix, I prove Theorem 1, and then show that Facts 2 and 3 follow as corollaries. Before I do so, however, let me note one thing. In the main text, I assumed that utility functions are defined over *outcomes*, i.e., propositions of the form $A \cap S$. But in this Appendix, it will be notationally simpler to work with utility functions defined over *worlds*, so that $u : \mathcal{W} \rightarrow \mathbb{R}$. This shift is harmless. After all, any outcome-level utility function determines a world-level utility function, if we set $u(w) = u(A \cap S)$ whenever $w \in A \cap S$. Conversely, for the purposes of the proof below, we may assume (without loss of generality) that the state partition is *so* fine-grained that any conjunction $A \cap S$ consists only of a single world. This brings our set-up from the main text into line with our set-up here.

Now, with that remark having been made, we can move to Theorem 1. To prove that result, we will need a definition and some lemmas.

Definition 1 (Linear Supposition function). *A supposition function s is linear if and only if, for every proposition X , every finite collection of probability functions $p_1, \dots, p_n \in \Delta(\mathcal{W})$, and every collection of weights $\lambda_1, \dots, \lambda_n \in [0, 1]$ with $\sum_i \lambda_i = 1$:*

$$s\left(\sum_i \lambda_i \cdot p_i, X\right)(\cdot) = \sum_i \lambda_i \cdot s(p_i, X)(\cdot)$$

Lemma 1 (Dorst et al., Theorem 4.1). *A probability function c totally trusts $\mathcal{P}$ if and only if:*

⁴³This isn't to say, however, that just *any* set of propositions can be used in the Value of Knowledge principle. On the contrary, an important result due to Geanakoplos (1989, Theorem 1) shows that these propositions must have certain properties, if they're to feature in this principle—in particular, they need to have a property called *nestedness*. It's an easy exercise to show that propositions of the form $\langle \mathbb{E}_{\mathcal{P}}[\mathcal{X}] \geq x \rangle$ satisfy Geanakoplos's constraints. See Dorst et al. (2021, p. 125, n. 5) for further discussion.

- (i) c can be expressed as a mixture of $\mathcal{P}$'s values, p_i , in the sense that there are weights $\lambda_i \in [0, 1]$, with $\sum_i \lambda_i = 1$, such that:

$$c(\cdot) = \sum_i \lambda_i \cdot p_i(\cdot)$$

- (ii) Each p_i satisfies a condition that Dorst et al. call modest informedness.⁴⁴

(Note that in what follows, we will be interested mostly in condition (i) of this lemma.)

Lemma 2 (Gärdenfors, 1982, Theorem 1). *A supposition function $s : \Delta(\mathcal{W}) \times \mathbb{P}(\mathcal{W}) \rightarrow \Delta(\mathcal{W})$ is linear if and only if it is imaging.*

With this definition and these lemmas in hand, we can now prove Theorem 1. We repeat the result here, for convenience:

Theorem 1 (Main Result). *Let s be a supposition function. Then, the following are equivalent:*

- (i) s determines a Stalnakerian causal decision theory.
- (ii) For all probability functions c and all expert functions $\mathcal{P}$, if c totally trusts $\mathcal{P}$, then c suppositionally values $\mathcal{P}$, with respect to the suppositional decision theory determined by s .

Proof. To see that (i) implies (ii), first suppose that s determines a Stalnakerian CDT. We want to show that, for all c and $\mathcal{P}$, if c totally trusts $\mathcal{P}$, then there exist weights λ_i such that c is a mixture of $\mathcal{P}$'s values with those weights, and for all utility functions u and decision problems $\mathcal{A}$:

$$\sum_i \lambda_i \cdot \max_j V_s(p_i, A_j, u) \geq \max_j V_s(c, A_j, u) \quad (1)$$

Thus, to start off, let c and $\mathcal{P}$ be arbitrary, and suppose that c totally trusts $\mathcal{P}$. Then, the first claim—that c can be expressed as a mixture of $\mathcal{P}$'s values—follows automatically by Lemma 1. Next, choose an arbitrary utility function u and decision problem $\mathcal{A}$. Stalnakerian CDT then says you should choose an option, $A_j \in \mathcal{A}$, that maximizes $V_s(c, A_j, u)$, which in turn means that the best you can do by choosing on your own is the following:

$$\max_j V_s(c, A_j, u), \quad (2)$$

(Note that this is just the right-hand side of (1).) Now, by definition, (2) is equal to the below:

$$\max_j \sum_w c^{A_j}(w) \cdot u(w) \quad (3)$$

⁴⁴The modest informedness condition says that each p_i is itself a mixture of the other values of $\mathcal{P}$, together with its informed self, $p_i^*(\cdot) = p_i(\cdot \mid \langle \mathcal{P} = p_i \rangle)$.

And by Lemma 2, we know that, since c is λ_i -mixture of $\mathcal{P}$, it then follows that c^{A_j} can be expressed as a similar mixture of the $p_i^{A_j}$. This allows us to re-write (3) like this:

$$\max_j \sum_w \sum_i \lambda_i \cdot p_i^{A_j}(w) \cdot u(w) \quad (4)$$

However, re-arranging now gives us:

$$\max_j \sum_i \lambda_i \cdot \sum_w p_i^{A_j}(w) \cdot u(w) \quad (5)$$

And here, the right-hand sum is equal to $V_s(p_i, A_j, u)$. This implies that (5) is equal to the following:

$$\max_j \sum_i \lambda_i \cdot V_s(p_i, A_j, u) \quad (6)$$

But now, compare the quantity in (6) to the left-hand side of (1). The latter is a weighted average of maxima, while the former is a maximum of weighted averages (using the same weights). The former can never exceed the latter, on general mathematical grounds. In turn, this establishes (ii).

We now show that (ii) implies (i). Thus, suppose that the conditions in (ii) hold. Then, fix an arbitrary probability function c and an arbitrary (non-empty) proposition A , and let $\mathcal{P}$ be the *maximally opinionated* expert function, defined by $\mathcal{P}_w = \delta_w$ for every world w , where δ_w is the probability function defined by $\delta_w(w) = 1$. Since c is a probability function, it follows that $c(\cdot) = \sum_w c(w) \cdot \delta_w(\cdot)$. In other words, c can be expressed as a mixture of the δ_w , with weights $c(w)$, and each δ_w is (trivially) modestly informed. Together, then, these facts imply that c totally trusts $\mathcal{P}$. (Indeed, something stronger is true: c reflects $\mathcal{P}$.)

Now, with c so defined, consider the decision problem $\mathcal{A} = \{A, \bar{A}\}$ (where $\bar{A}$ is A 's complement). From our assumption of (ii), it follows that, for all utility functions u :

$$\sum_w c(w) \cdot \max \{V_s(\delta_w, A, u), V_s(\delta_w, \bar{A}, u)\} \geq \max \{V_s(c, A, u), V_s(c, \bar{A}, u)\} \quad (7)$$

However, choose an arbitrary world $v \in A$ and let u_v be the utility function defined by $u_v(v) = 1$ and $u_v(w) = 0$ for all worlds $w \neq v$. Then, since $s(c, A)(A) = 1$ for any c (by our definition of a supposition function), we have that $s(c, A)(v) \geq 0$ while $s(c, \bar{A})(v) = 0$; and similarly $s(c, A)(w) = 0$ for all $w \notin A$. Thus, $V_s(c, A, u_v) = s(c, A)(v)$ and $V_s(c, \bar{A}, u_v) = 0$; and likewise $V_s(\delta_w, A, u_v) = s(\delta_w, A)(v)$ and $V_s(\delta_w, \bar{A}, u_v) = 0$. Given this, then, (7) becomes:

$$\sum_w c(w) \cdot s(\delta_w, A)(v) \geq s(c, A)(v) \quad (8)$$

However, since our choice of $v \in A$ was arbitrary, the foregoing must hold for every choice of world. In turn, this implies that the above inequality can't be

strict, for, if it were, summing over all worlds in A would yield $1 > 1$, since each of $\sum_w c(w) \cdot s(\delta_w, A)$ and $s(c, A)$ is a probability function concentrated on A . Thus, (8) must hold with equality for every $v \in A$; and for every $v \notin A$ both sides must be zero. Consequently:

$$s(c, A)(v) = \sum_w c(w) \cdot s(\delta_w, A)(v) \quad (9)$$

for all choices of $v \in \mathcal{W}$. Finally, then, since c and A were arbitrary, (9) holds for every probability function and every proposition. However, this implies that s is linear, and by Lemma 2 it follows that s is imaging. From this we get that s must determine a Stalnakerian CDT, which establishes (i). $\square$

Importantly, note here that Facts 2 and 3 now follow as corollaries of Theorem 1. We get the (i) implies (ii) direction for free, since Reflection and Informed Reflection each entail Total Trust—see Dorst et al. (2021, p. §1 and §4.1). And the (ii) implies (i) direction follows since, in the corresponding direction of the proof of Theorem 1, the probability functions δ_w we appealed to satisfy both Reflection and Informed Reflection.

An Alternative Framing

Before closing, there's one other thing I would like to note. This is: as I mentioned in the main text (fn. 20), my definition of Value differs a little bit from Dorst et al.'s. In particular, my definition is closer to Nielsen's (2023) statement of the Value of Knowledge principle, whereas Dorst et al. define Value using the notion of a *strategy*—roughly, a function that maps each world w to an option that's recommended by the expert at that world. In the suppositional decision theory setting, we can define the analogous notion as follows. First, fix a decision problem $\mathcal{A}$. Then, a *strategy* for this decision problem is a function $\mathcal{S} : \mathcal{W} \rightarrow \mathcal{A}$ such that $\mathcal{S}_w = \mathcal{S}_v$ whenever $\mathcal{P}_w = \mathcal{P}_v$. (Here, I write ' $\mathcal{S}_w$ ' in lieu of ' $\mathcal{S}(w)$ '.) A strategy $\mathcal{S}$ is then *recommended* by $\mathcal{P}$ if, for each value $\mathcal{P}_w$ of $\mathcal{P}$, the option outputted by $\mathcal{S}$ is one that maximizes $V_s(\mathcal{P}_w, \cdot, u)$.

Thus, in parallel to Dorst et al., let us now define—slightly abusing notation—the expected value of following a strategy, $\mathcal{S}$, like this:

$$V_s(c, \mathcal{S}, u) := \sum_w c(w) \cdot V_s(\delta_w, \mathcal{S}_w, u) \quad (10)$$

Here, each δ_w is, again, the probability function defined by $\delta_w(w) = 1$. Thus, the foregoing quantity is something like your expectation of the *actual value* of abiding by the strategy $\mathcal{S}$ at each world (Bacon, 2022). In other words, (10) defines the expected value of selecting an option that's recommended by the expert at w , *at the world w itself*. (Note here that this definition requires $V_s(\delta_w, \mathcal{S}_w, u)$ to be well-defined for every w and $\mathcal{S}_w$. And when s is imaging, this is unproblematic: we can simply set $s(\delta_w, \mathcal{S}_w) = f(\mathcal{S}_w, w)$, which makes

$V_s(\delta_w, \mathcal{S}_w, u) = \sum_v f(\mathcal{S}_w, w)(v) \cdot u(v)$. However, when s is conditionalization, $s(\delta_w, \mathcal{S}_w) = \delta_w(\cdot \mid \mathcal{S}_w)$ is undefined whenever $w \notin \mathcal{S}_w$. So, our definition of the expected value of following a strategy does not obviously encompass EDT—that’s the reason I didn’t adopt this definition in the main text.)

Now, with the foregoing definitions in hand, we can prove a variation of Theorem 1, using something much closer to Dorst et al.’s own characterization of Value. Specifically, in their framework, Dorst et al. define Value using the notion of a strategy: roughly, c values $\mathcal{P}$ ’s probabilities just in case the expected utility of following any strategy recommended by $\mathcal{P}$ is at least as great as the expected utility of choosing on your own. Translated into our framework, however, this amounts to saying that, for all decision problems $\mathcal{A}$, all utility functions u , and all strategies $\mathcal{S}$ recommended by $\mathcal{P}$:

$$V_s(c, \mathcal{S}, u) \geq \max_j V_s(c, A_j, u) \quad (11)$$

Then, given this slightly modified definition of Suppositional Value, we can show that only Stalnakerian CDT satisfies this condition:

Fact 5. *Let s be a supposition function. Then, the following are equivalent:*

- (i) *s determines a Stalnakerian causal decision theory.*
- (ii) *For all probability functions c and all expert functions $\mathcal{P}$, if c totally trusts $\mathcal{P}$, then c suppositionally values $\mathcal{P}$, in the strategy sense given in (11).*

Proof Sketch. First, suppose that (i) holds. Then, the suppositional decision theory induced by s is a Stalnakerian CDT. Now, let c , $\mathcal{P}$, u , and $\mathcal{A}$ be arbitrary; suppose that c totally trusts $\mathcal{P}$; and let $\mathcal{S}$ be a strategy recommended by $\mathcal{P}$ in $\mathcal{A}$. Since s is imaging, Lemma 2 gives us that:

$$s(c, A_j) = \sum_w c(w) \cdot s(\delta_w, A_j) \quad (12)$$

In turn, this implies that, for any utility function u , the right-hand side of (11) is the following:

$$\max_j \sum_w c(w) \cdot V_s(\delta_w, A_j, u) \quad (13)$$

Similarly, by definition (10), the left-hand side of (11) can be written thus:

$$\sum_w c(w) \cdot V_s(\delta_w, \mathcal{S}_w, u). \quad (14)$$

But the desired inequality now follows immediately from Dorst et al. (2021, Theorem 2.2),⁴⁵ which means that c suppositionally values $\mathcal{P}$, in the strategy sense.

⁴⁵Dorst et al. define options as random variables—functions from worlds to real numbers. Specifically, the option corresponding to A_j is the function that maps each world v to the quantity

Next, for the converse direction, suppose that (ii) holds. Then, fix an arbitrary c , proposition A , and world $v \in A$, and let $\mathcal{P}$ be the maximally opinionated expert function, defined by $P_w = \delta_w$ for every world w . As before, c totally trusts $\mathcal{P}$. So, let u_v be defined again by $u_v(v) = 1$ and $u_v(w) = 0$ for all worlds $w \neq v$. Additionally, let $\mathcal{A} = \{A, \bar{A}\}$, and let $\mathcal{S}$ be the constant strategy, defined by $S_w = A$ at every world w . Then, $\mathcal{S}$ is recommended by $\mathcal{P}$, since at each world w , we have $V_s(\delta_w, A, u_v) = s(\delta_w, A)(v) \geq 0 = V_s(\delta_w, \bar{A}, u_v)$. So, applying (ii) to $\mathcal{S}$, the inequality $V_s(c, \mathcal{S}, u_v) \geq \max_j V_s(c, A_j, u_v)$ holds. Furthermore, by definition (10), its left-hand side is $\sum_w c(w) \cdot s(\delta_w, A)(v)$, while its right-hand side is $s(c, A)(v)$. Hence:

$$\sum_w c(w) \cdot s(\delta_w, A)(v) \geq s(c, A)(v) \quad (15)$$

This is exactly the inequality obtained in the proof that (ii) implies (i) in Theorem 1. The rest of the argument thus proceeds precisely as it did before. $\square$

Thus, a version of our main result follows, even if we adopt a definition of Suppositional Value that's based on the one given by Dorst et al. themselves. One other thing that's nice about this result is that it draws out the connection between Dorst et al.'s result, and Bacon's result about actual value (since, again, the quantity $V_s(\delta_w, \mathcal{S}_w, u)$ is really just the *actual value* of choosing the option recommended by the expert at w , at w itself).

References

- Ahmed, Arif (2013). "Causal Decision Theory: A Counterexample". In: *Philosophical Review* 122.2, pp. 289–306. DOI: [10.1215/00318108-1963725](https://doi.org/10.1215/00318108-1963725). URL: <https://doi.org/10.1215/00318108-1963725>.
- Ahmed, Arif (2014a). "Causal Decision Theory and the Fixity of the Past". In: *The British Journal for the Philosophy of Science* 65.4, pp. 665–685. DOI: [10.1093/bjps/axt021](https://doi.org/10.1093/bjps/axt021). URL: <https://doi.org/10.1093/bjps/axt021>.
- Ahmed, Arif (2014b). *Evidence, Decision and Causality*. Cambridge University Press.
- Ahmed, Arif (2021). *Evidential Decision Theory*. Cambridge University Press.
- Ahmed, Arif and Bernhard Salow (2019). "Don't Look Now". In: *The British Journal for the Philosophy of Science* 70.2, pp. 327–350. DOI: [10.1093/bjps/axx047](https://doi.org/10.1093/bjps/axx047). URL: <https://doi.org/10.1093/bjps/axx047>.

$V_s(\delta_v, A_j, u)$. However, when s is imaging, it follows from Lemma 2 that any probability function c can be written as the mixture $c(\cdot) = \sum_v c(v) \cdot \delta_v(\cdot)$, and so $V_s(c, A_j, u) = \sum_v c(v) \cdot V_s(\delta_v, A_j, u)$. This implies that, when s is imaging, the suppositional expected utility of the option A_j is just c 's expectation of the corresponding random variable, in Dorst et al.'s sense. Thus, the (i) implies (ii) direction of our proof follows straightforwardly from their result.

- Ahmed, Arif and Jack Spencer (2020). "Objective Value Is Always Newcom-bizable". In: *Mind* 129.516, pp. 1157–1192. DOI: [10.1093/mind/fzz070](https://doi.org/10.1093/mind/fzz070). URL: <https://doi.org/10.1093/mind/fzz070>.
- Bacon, Andrew (2022). "Actual value in decision theory". In: *Analysis* 82.4, pp. 617–629. DOI: [10.1093/analys/anac014](https://doi.org/10.1093/analys/anac014). URL: <https://doi.org/10.1093/analys/anac014>.
- Bradley, Richard (2017). *Decision Theory with a Human Face*. Cambridge University Press.
- Dorr, Cian (2016). "Against Counterfactual Miracles". In: *Philosophical Review* 125.2, pp. 241–286. DOI: [10.1215/00318108-3453187](https://doi.org/10.1215/00318108-3453187). URL: <https://doi.org/10.1215/00318108-3453187>.
- Dorst, Kevin (2019). "Evidence: A Guide for the Uncertain". In: *Philosophy and Phenomenological Research* 100.3, pp. 586–632. DOI: [10.1111/phpr.12561](https://doi.org/10.1111/phpr.12561). URL: <https://doi.org/10.1111/phpr.12561>.
- Dorst, Kevin, Benjamin A. Levinstein, Bernhard Salow, Brooke E. Husic, and Branden Fitelson (2021). "Deference Done Better". In: *Philosophical Perspectives* 35.1, pp. 99–150. DOI: [10.1111/phpe.12156](https://doi.org/10.1111/phpe.12156). URL: <https://doi.org/10.1111/phpe.12156>.
- Elga, Adam (2013). "The puzzle of the unmarked clock and the new rational reflection principle". In: *Philosophical Studies* 164.1, pp. 127–139. DOI: [10.1007/s11098-013-0091-0](https://doi.org/10.1007/s11098-013-0091-0). URL: <https://doi.org/10.1007/s11098-013-0091-0>.
- Elga, Adam (2022). "Confession of a causal decision theorist". In: *Analysis*. DOI: [10.1093/analys/anab040](https://doi.org/10.1093/analys/anab040). URL: <https://doi.org/10.1093/analys/anab040>.
- Fishburn, Peter C. (1964). *Decision and Value Theory*. New York: Wiley.
- Fraassen, Bas van (1984). "Belief and the Will". In: *Journal of Philosophy* 81.5, pp. 235–256.
- Gaifman, Haim (1986). "A Theory of Higher Order Probabilities". In: *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the 1986 Conference*. Ed. by Joseph Y. Halpern. Los Altos, CA: Morgan Kaufmann, pp. 275–292.
- Gallow, J. Dmitri (2018). "No One Can Serve Two Epistemic Masters". In: *Philosophical Studies* 175.10, pp. 2389–2398. DOI: [10.1007/s11098-017-0964-8](https://doi.org/10.1007/s11098-017-0964-8).
- Gallow, J. Dmitri (2021). "Updating for Externalists". In: *Noûs* 55.3, pp. 487–516. DOI: [10.1111/nous.12307](https://doi.org/10.1111/nous.12307).
- Gallow, J. Dmitri (2023). "Local and global deference". In: *Philosophical Studies* 180.9, pp. 2753–2770. DOI: [10.1007/s11098-023-02003-8](https://doi.org/10.1007/s11098-023-02003-8). URL: <https://doi.org/10.1007/s11098-023-02003-8>.
- Gallow, J. Dmitri (2024). "Counterfactual Decision Theory Is Causal Decision Theory". In: *Pacific Philosophical Quarterly* 105.1, 115–156. ISSN: 1468-0114. DOI: [10.1111/papq.12451](http://dx.doi.org/10.1111/papq.12451). URL: <http://dx.doi.org/10.1111/papq.12451>.

- Gärdenfors, Peter (1982). "Imaging and Conditionalization". In: *The Journal of Philosophy* 79.12, p. 747. DOI: [10.2307/2026039](https://doi.org/10.2307/2026039). URL: <https://doi.org/10.2307/2026039>.
- Geanakoplos, John (1989). "Game Theory without Partitions, and Applications to Speculation and Consensus". In: *Cowles Foundation Discussion Papers* 1158.
- Gibbard, Allan and William Harper (1978). "Counterfactuals and Two Kinds of Expected Utility". In: *Foundations and Applications of Decision Theory*. Ed. by A. Hooker, J. J. Leach, and E. F. McClennen. D. Reidel, pp. 125–162.
- Good, I.J. (1967). "On the Principle of Total Evidence". In: *The British Journal for the Philosophy of Science* 17.4, pp. 319–321. ISSN: 1464-3537. DOI: [10.1093/bjps/17.4.319](https://doi.org/10.1093/bjps/17.4.319). URL: <http://dx.doi.org/10.1093/bjps/17.4.319>.
- Hall, Ned (1994). "Correcting The Guide to Objective Chance". In: *Mind* 103.412, pp. 505–518. ISSN: 1460-2113. DOI: [10.1093/mind/103.412.505](https://doi.org/10.1093/mind/103.412.505). URL: <http://dx.doi.org/10.1093/mind/103.412.505>.
- Hall, Ned (2004). "Two Mistakes About Credence and Chance". In: *Australasian Journal of Philosophy* 82.1, pp. 93–111. DOI: [10.1080/713659806](https://doi.org/10.1080/713659806). URL: <https://doi.org/10.1080/713659806>.
- Hedden, Brian (2023). "Counterfactual Decision Theory". In: *Mind* 132.527, pp. 730–761. DOI: [10.1093/mind/fzac060](https://doi.org/10.1093/mind/fzac060). URL: <https://doi.org/10.1093/2Fmind%2Ffzac060>.
- Herrmann, Daniel and Gerard Rothfus (forthcoming). "Deference and Decision". In: *Theory and Decision*.
- Hild, Matthias (1998). "Auto-Epistemology and Updating". In: *Philosophical Studies* 92.3, pp. 321–361. DOI: [10.1023/a:1004229808144](https://doi.org/10.1023/a:1004229808144).
- Hitchcock, Christopher (2013). "What is the 'Cause' in Causal Decision Theory?". In: *Erkenntnis* 78.S1, pp. 129–146. DOI: [10.1007/s10670-013-9440-9](https://doi.org/10.1007/s10670-013-9440-9). URL: <https://doi.org/10.1007/2Fs10670-013-9440-9>.
- Huttegger, Simon (2013a). "In Defense of Reflection". In: *Philosophy of Science* 80.3, pp. 413–433. DOI: [10.1086/671427](https://doi.org/10.1086/671427). URL: <https://doi.org/10.1086/2F671427>.
- Huttegger, Simon (2013b). "Learning experiences and the value of knowledge". In: *Philosophical Studies* 171.2, pp. 279–288. ISSN: 1573-0883. DOI: [10.1007/s11098-013-0267-7](https://doi.org/10.1007/s11098-013-0267-7). URL: <http://dx.doi.org/10.1007/s11098-013-0267-7>.
- Huttegger, Simon (2023). "Reconciling Evidential and Causal Decision Theory". In: *Philosophers' Imprint* 23.0. ISSN: 1533-628X. DOI: [10.3998/phimp.931](https://doi.org/10.3998/phimp.931). URL: <http://dx.doi.org/10.3998/phimp.931>.
- Jeffrey, Richard C. (1965). *The Logic of Decision*. 1st Edition, University of Chicago Press.
- Jeffrey, Richard C. (1983). *The Logic of Decision*. 2nd Edition, Chicago and London: University of Chicago Press.
- Joyce, James M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

- Kadane, Joseph B., Mark Schervish, and Teddy Seidenfeld (2008). "Is Ignorance Bliss?" In: *Journal of Philosophy* 105.1, pp. 5–36.
- Kment, Boris (2023). "Decision, Causality, and Predetermination". In: *Philosophy and Phenomenological Research*. DOI: [10.1111/phpr.12935](https://doi.org/10.1111/phpr.12935). URL: <https://doi.org/10.1111%2Fphpr.12935>.
- Levinstein, Benjamin A and Jack Spencer (2024). "Bigger, Badder Bugs". In: *Mind*. ISSN: 1460-2113. DOI: [10.1093/mind/fzae039](https://doi.org/10.1093/mind/fzae039). URL: <http://dx.doi.org/10.1093/mind/fzae039>.
- Levinstein, Benjamin A. (2023). "Accuracy, Deference, and Chance". In: *Philosophical Review* 132.1, pp. 43–87. DOI: [10.1215/00318108-10123774](https://doi.org/10.1215/00318108-10123774). URL: <https://doi.org/10.1215%2F00318108-10123774>.
- Lewis, David (1976). "Probabilities of Conditionals and Conditional Probabilities". In: *The Philosophical Review* 85.3, p. 297. ISSN: 0031-8108. DOI: [10.2307/2184045](https://doi.org/10.2307/2184045). URL: <http://dx.doi.org/10.2307/2184045>.
- Lewis, David (1980). "A Subjectivist's Guide to Objective Chance". In: *Philosophical Papers: Volume II*. Oxford, pp. 263–293.
- Lewis, David (1981). "Causal decision theory". In: *Australasian Journal of Philosophy* 59.1, pp. 5–30. ISSN: 1471-6828. DOI: [10.1080/00048408112340011](https://doi.org/10.1080/00048408112340011). URL: <http://dx.doi.org/10.1080/00048408112340011>.
- Lewis, David (1994). "Humean Supervenience Debugged". In: *Mind* 103.412, pp. 473–490. ISSN: 00264423, 14602113. URL: <http://www.jstor.org/stable/2254396>.
- McNamara, Calum (2024). "Causal decision theory, context, and determinism". In: *Philosophy and Phenomenological Research* 109.1, 226â260. ISSN: 1933-1592. DOI: [10.1111/phpr.13021](https://doi.org/10.1111/phpr.13021). URL: <http://dx.doi.org/10.1111/phpr.13021>.
- Nielsen, Michael (2023). "Only CDT values knowledge". In: *Analysis*. DOI: [10.1093/analys/anad041](https://doi.org/10.1093/analys/anad041). URL: <https://doi.org/10.1093%2Fanalys%2Fanad041>.
- Nozick, Robert (1969). "Newcomb's Problem and Two Principles of Choice". In: *Essays in Honor of Carl G. Hempel*. Ed. by Nicholas Rescher. Reidel, pp. 114–146.
- Pollock, John L. (2009). "A resource-bounded agent addresses the newcomb problem". In: *Synthese* 176.1, pp. 57–82. ISSN: 1573-0964. DOI: [10.1007/s11229-009-9484-1](https://doi.org/10.1007/s11229-009-9484-1). URL: <http://dx.doi.org/10.1007/s11229-009-9484-1>.
- Savage, Leonard J. (1954). *The Foundations of Statistics*. Dover Publications.
- Schoenfield, Miriam (2017). "Conditionalization Does Not (in General) Maximize Expected Accuracy". In: *Mind* 126.504, pp. 1155–1187. ISSN: 1460-2113. DOI: [10.1093/mind/fzw027](https://doi.org/10.1093/mind/fzw027). URL: <http://dx.doi.org/10.1093/mind/fzw027>.
- Skyrms, Brian (1980). *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. Yale University Press.
- Skyrms, Brian (1982). "Causal Decision Theory". In: *The Journal of Philosophy* 79.11, p. 695. ISSN: 0022-362X. DOI: [10.2307/2026547](https://doi.org/10.2307/2026547). URL: <http://dx.doi.org/10.2307/2026547>.

- Skyrms, Brian (1984). *Pragmatics and Empiricism*. Yale University Press, New Haven.
- Skyrms, Brian (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.
- Sobel, Jordan Howard (1994). *Taking Chances: Essays on Rational Choice*. Cambridge University Press.
- Spencer, Jack and Ian Wells (2019). "Why Take Both Boxes?" In: *Philosophy and Phenomenological Research* 99.1, pp. 27–48. ISSN: 0031-8205. DOI: [10.1111/phpr.12466](https://doi.org/10.1111/phpr.12466). URL: <http://dx.doi.org/10.1111/phpr.12466>.
- Stalnaker, Robert (1968). "A Theory of Conditionals". In: *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*. Ed. by Nicholas Rescher. Oxford: Blackwell, pp. 98–112.
- Stalnaker, Robert (1981). "Letter to David Lewis". In: *Ifs*. Ed. by William Harper, Robert C. Stalnaker, and Glenn Pearce. Reidel, pp. 151–152.
- Stalnaker, Robert (2019). "Rational Reflection, and the Notorious Unmarked Clock". In: *Knowledge and Conditionals: Essays on the Structure of Inquiry*. Oxford University Press, pp. 99–112.