Epistemic Time Bias in Newcomb's Problem*

Melissa Fusco

Causal decision theorists like David Lewis hold that, while an agent should choose acts by using her current, rather than *anticipated*, credences over causal dependency hypotheses, she should also update by conditionalization on her own act as she performs it (Lewis 1981a, 1976). This package leads to an unflattering form of time-bias, which can be highlighted by considering iterated Newcomb problems. After presenting the puzzle, I discuss a CDTer's best response.

1. Introduction

Suppose you will either experience an hour of pain on Tuesday and an hour of pleasure on Thursday, or an hour of pleasure on Tuesday and an hour of pain on Thursday. Neither possibility seems clearly better than the other—unless, that is, today is Wednesday. For a fixed menu of pleasures and pains, *time-biased agents* prefer distributions wherein more pains are in the past, and more pleasures are in the future.

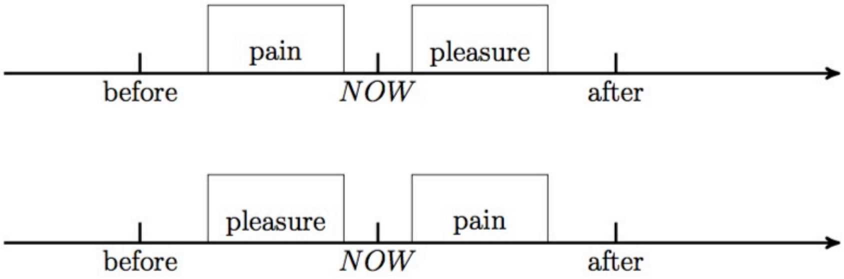


Figure 1: Basic Time Bias.

The two possibilities in Figure 1 illustrate this basic commitment.

A classic discussion of time bias can be found in Chapter 8 of Parfit's *Reasons and Persons* (Parfit 1984).[1] Parfit imagines that he wakes up in the hospital, and is told by a nurse that *either* he has just survived a very long and painful operation (which he does not remember, since patients undergoing his surgery are regularly

* Warm thanks to Arif Ahmed, Lara Buchak, Alex Kocurek, and John MacFarlane (via pandoc) for their help on this paper.

[1] Parfit distinguishes two forms of bias, *future* bias—being more concerned about one's future welfare than for one's past welfare—and *near* bias—being more concerned for the near future than for the distant future. Future bias is thought to be the more rationally defensible than near bias (see op. cit., pg. 158 ff.) I focus only on future bias here.

given post-operative medication which induces amnesia), *or* he has yet to undergo a shorter version of an operation of the same type. While Parfit's character generally prefers to suffer less, rather than more, he now finds that he would much prefer to learn that he has already had the longer, earlier operation. His preferences seem to have been reversed by the mere passage of time. Nor is this a reversal between outcomes which are intuitively equal. Rather, Parfit's character now prefers a possibility in which his life contains *more* total hours of suffering.

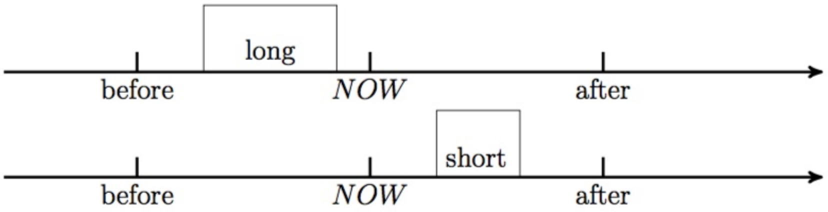


Figure 2: Parfit's Surgeries.

In this paper I investigate a form of time bias which is entailed by a common form of causal decision theory. It can be seen in the causalist's response to Newcomb's problem.

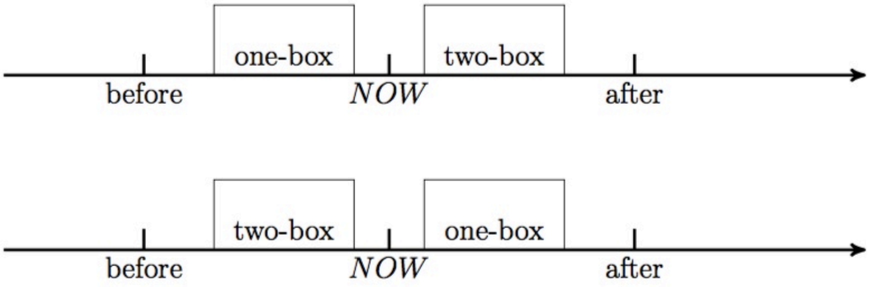


Figure 3: Newcomb Time Bias.

A standard causal decision theorist in Newcomb's problem—an agent who prefers to two-box, given the chance—will nonetheless prefer to learn that Newcomb problems which involve her, but over which she currently has *no agency*, are ones in which she one-boxed. When Parfittian amnesia is on the scene, time bias easily takes root: learning about one's past actions is a paradigm of being invested in a decision over which one does not currently exercise agency.[2] Hence an agent, located at NOW, who sees the past as fixed but the future as hers to choose, will prefer the upper state in the figure to the lower one. As we will see, the CDTer's time bias is, if

[2] Arif Ahmed points out to me that an act's *being in the past* is not *essential* to the difference between choosing and learning. The relevant type of CDTer will also prefer to learn that she *will* one-box in the future. This is true, but such "learnings" are puzzling because of the difficulty many people (myself included) will have reconciling such foreknowledge with free agency. For this reason, I will focus in this essay on learning about the past.

anything, stranger than the time bias investigated by Parfit, because it is traceable to a normative, rather than affective, source.

One puzzle about this setup will immediately present itself. I say that, in Figures 1-3, time-biased agents would pay to switch from the lower possibility into the upper. But—it will be objected—an agent located at NOW cannot *really* make such a switch, since that would involve changing her past. This is true, but I show in section 3 that the basic structure of the time-biased preferences illustrated in Figure 3 can, with suitable manipulation, be turned into actionable preferences which operationalize the basic dispute between the biased and unbiased perspectives.

Nonetheless, this is a confessional essay: I *maintain* the CDTer's commitments in both cases, despite the (literal) costs. Damage control commences in section 4, where I argue that *Evidential* Decision Theory is also committed to a form of time bias. The question, then, is whether the two forms of bias have an equal claim to be features of rational behavior.

2. Evidence and Decision

In this section I sketch the shared background of causal and evidential decision theory with an eye to two issues, screening-off and conditionalization, which bear on the argument to come.

I will assume that in all decision problems, agents have the goal of maximizing expected utility. They come to these problems equipped with a value function $Val(\cdot)$ on possible worlds, which can be lifted onto propositions called *outcomes*;[3] following standard idealization, I will occasionally speak as if the units in the range of the value function are equivalent to dollars. An agent's decision problem at a time t can be represented as a triple $\langle \mathcal{A}, \mathcal{S}, Cr^t \rangle$, where

- $\mathcal{A} = \{a_1, \ldots, a_n\}$ is a set of *acts* available to the agent at t,
- $\mathcal{S} = \{s_1, \ldots, s_m\}$ is a set of admissible *states of nature*, where a set of states is admissible in a decision problem only if each act-state conjunction $(a_i \wedge s_j)$ is assigned a determinate image $Val(a_i \wedge s_j)$ under the value function, and
- $Cr^t(\cdot)$ is the agent's time-t credence function.[4]

[3] The lifting of the value function can be achieved as follows. Where $val(\cdot)$ is a worldly value function in $W \to \mathbb{R}$, (i) $Val(\{w\}) = val(w)$, and (ii) $Val(p \cup q)$ is defined only if $Val(p) = Val(q)$, in which case $Val(p \cup q) = Val(p)$.

[4] In what follows, I treat the members of $\mathcal{A}$, $\mathcal{S}$, and the domain of $Cr^t(\cdot)$ alike as *propositions*. This differs from Savage's original picture of the relevant primitives, on which acts are *functions* from states to outcomes. See Joyce (1999 Ch. 2) for discussion of this shift.

At a first pass, the agent's credence function $Cr^t(\cdot)$ will be conceived of in Bayesian terms. I assume that it is a probability function, and that the agent is disposed to respond to new information by conditioning this function on the strongest proposition of which she is informed. I'll also assume that her expectations are generally *frequentist*, in the sense that when they concern events in the present or future, they will, *ceteris paribus*, conform to observed past frequencies over event-classes of the same type.

Against this background, we can introduce Savage (1972)'s decision theory. It applies the general statistical notion of expectation to the value of an act $a \in \mathcal{A}$.[5]

Equation 1 (Savage Expected Utility). $SEU^t(a) = \sum_i C\,r^t(s_i)Val(a \wedge s_i)$.
Savage's norm: maximize expected utility at t by choosing an act $a \in \mathcal{A}$ such that $SEU^t(a)$ is maximal: that is, such that $SEU^t(a) \geq SEU^t(a')$, for any $a' \in \mathcal{A}$.

EDT and CDT can then be framed as two different ways of cashing out a final condition Savage put on well-formulated decision problems: the requirement that the set $\mathcal{S}$ of states in the problem are, at time t, *independent* of the agents' acts.

EDT construes act-state independence in terms of probabilistic confirmation. In general, the EDTer holds that rationality requires calculating expected utilities in terms of act-conditioned probabilities $Cr^t(s|a)$:

Equation 2 (Evidential Expected Utility). $EEU^t(a) = \sum_i C\,r^t(s_i|a)Val(a \wedge s_i)$.
Conditional Probability norm: maximize expected utility by choosing an act $a \in \mathcal{A}$ such that $EEU^t(a) \geq EEU^t(a')$, for any $a' \in \mathcal{A}$.

This does justice to one attractive gloss on Savage's independence requirement on $\mathcal{S}$: it entails that Savage's equation will apply to any problem in which $Cr^t(s|a)$ is equal to $Cr^t(s)$—in other words, to any problem in which the agent takes a to be *probabilistically* independent of s.

CDT, on the other hand, favors an interpretation of act-state independence that is *causal*. She rejects the EDTer's gloss because probabilistic independence—a lack of probabilistic (dis)confirmation—is neither necessary, nor sufficient, for causal independence.[6] In this paper, I will use Lewis (1981a)'s "dependency hypothesis" characterization of the CDT. This takes as primitive a partition $\mathcal{K}$ of dependency hypotheses, which, are *stipulated* to be causally independent of the

[5] Where F is a function, the generic notion of expected value says that $E[F] = \sum_i f_i\,Pr(F = f_i)$. Here, our probability function is the subjective credence function $Cr^t(\cdot)$, and the function F is $Val(\cdot)$ across act-state pairs, where the act is held fixed.

[6] For non-necessity, we can consider any case (such as Newcomb) in which an act a probabilistically confirms some state s whose truth-value was fixed before a occurred. For non-sufficiency, we can appeal to examples such as Hesslow (1976)'s well-known thrombosis case: while birth control pills cause thrombosis, the effect is statistically masked because birth control pills counteract pregnancy, which *itself* causes thrombosis.

agent's available acts, as well as sufficient to fix an outcome value for each $(a \wedge k_i)$ conjunction.[7]

Equation 3 (Causal Expected Utility). $CEU^t(a_j) = \sum_i C\,r^t(k_i)Val(a_j \wedge k_i)$, where $\mathcal{K} = \{k_1 \ldots k_n\}$ is a partition of dependency hypotheses.
Causal Support norm: maximize expected utility by choosing an act $a \in \mathcal{A}$ such that $CEU^t(a) \geq CEU^t(a')$, for any $a' \in \mathcal{A}$.

This way of stating CDT leaves open the model-theoretic implementation of what is special about the k's—leaves open, that is, how doxastic states might be modeled so as to *reflect* an agent's conviction that the k's are causally independent of acts in $\mathcal{A}$. It is clear, however, that these modeling requirements will exceed those of austere Bayesianism.[8] Note that, since EEU is partition-invariant, the evidential decision theorist can also calculate expected utility using the $\mathcal{K}$-partition.[9]

We are now in a position to define *NC Problems*, decision problems with the general structure of Newcomb's puzzle. For my purposes, an NC Problem will be a problem with two acts $\mathcal{A} = \{a^*, \neg a^*\}$ and two causal dependency hypotheses $\mathcal{K} = \{k_1, k_2\}$, in which a dominance argument across $\mathcal{K}$ gives the agent a one-utile incentive to choose a^*; nonetheless, because $\neg a^*$ indicates a much more valuable outcome (which I'll refer to generically as "the big prize"), the conditional probability norm favored by the EDTer recommends $\neg a^*$. Hence from the CDTer's point of view, the distinctive thing to say about NC problems is that maximizing expected utility will entail getting some "bad news on the side" (Joyce 2007 pg. 542)—news that statistically disconfirms one's getting the big prize.

Here is an example which we will revisit throughout.

Mood Candles. Lighting aromatic mood candles (= L) slightly increases happiness. Bob is deciding whether to light a mood candle. He also finds himself unable to remember whether he suffers from depression (= D). In his current state of ignorance and indecision, his possible outcomes, and their values, are:

	Not Depressed ($\neg D$)	Depressed (D)
Don't light ($\neg L$)	9	0
Light (L)	10	1

[7] For a probabilistic version of the same idea—which does not require that dependency hypotheses *entail* outcomes, but merely that they fix their chances—see §10 of Lewis (1981a).

[8] Candidates include the directed acyclic graphs of Pearl (2000), and the worldly distance metrics of Lewis (1973)'s treatment of natural language counterfactuals.

[9] Evidential expected utility is *partition invariant* in the sense that the expected utility assigned to all acts $a \in \mathcal{A}$ w.r.t. a credal state $Cr^t(\cdot)$ will be the same in any problems $\langle \mathcal{A}, \mathcal{S}_1, Ct^t(\cdot) \rangle$ and $\langle \mathcal{A}, \mathcal{S}_2, Ct^t(\cdot) \rangle$ so long as $\cup\mathcal{S}_1 = \cup\mathcal{S}_2$. See Joyce (1999 pg. 176 ff.), and references therein, for discussion of this feature.

> There's a catch, however. Knowledge of the therapeutic value of mood candles is rare. Only people who are depressed tend to have it, since only they are regularly updated by their doctors about mood therapies. Hence Bob's credences are such that $Cr^t(D|L) > Cr^t(D)$: while lighting mood candles modestly conduces to happiness, it is a strong indicator of *unhappiness*.

A flatfooted (perhaps *naive*; see below) calculation by the EDTer's conditional probability norm will recommend that Bob refrain from lighting a mood candle, so long as L raises the statistical probability of D by more than 1/9.[10] Given that Bob also believes that lighting mood candles cannot *cause* depression, however, it is unclear that this is the right recommendation.

2.1 Screening Off

The formal description of an NC problem is clearly coherent, and *Mood Candles* prima facie fits the bill. Some philosophers have argued, however, that there *are* no NC problems in real life—at least, not for rational agents. This is a common factor of two venerable threads in the literature: the so-called "Ramsey thesis" ((Ahmed 2014 Ch. 8), (Ramsey 1990)), and Ellery Eell's "tickle defense" of EDT (Eells 1982). The shared idea is that as long as a responsible agent takes her full pre-decision evidence into account, and is not irrationally influenced by factors outside of her control, she will be able to access evidence Z that *screens off* the statistical support that her act would apparently provide for any state k:[11]

> **Screening off.** Z screens off k from a relative to a probability function $Cr^t(\cdot)$ just in case $Cr^t(k|a \wedge Z) = Cr^t(k|Z)$.

[10] Calculation:

$$\begin{aligned} EEU^t(\neg L) &> EEU^t(L) \text{ iff} \\ \sum_i Cr^t(k_i|\neg L)Val(k_i \wedge \neg L) &> \sum_i Cr^t(k_i|L)Val(k_i \wedge L) \text{ iff} \\ &> [Cr^t(\neg D|L)\times 10 + Cr^t(D|L)\times 1] \text{ iff} \\ &> [Cr^t(\neg D|L)\times 10 + Cr^t(D|L)] \end{aligned}$$

Since $Cr^t_\phi(\cdot) = Cr^t(\cdot\ |\phi)$ is itself a probability function, this is equivalent to

$$[Cr^t_{\neg L}(\neg D)\times 9] \quad > [Cr^t_D(\neg D)\times 10 + (1 - Cr^t_L(\neg D))]$$

Letting $n = Cr^t_{\neg L}(\neg D)$ and $m = Cr^t_L(\neg D)$, this obtains just in case

$$\begin{aligned} 9n &> 10m + (1 - m) \text{ iff} \\ n &> m + 1/9. \end{aligned}$$

[11] Eells's discussion emphasizes the first idea—screening evidence is always available pre-decision—while Ramsey emphasizes the second—that an agent not irrationally influenced by factors outside of her control in fact *cannot* learn anything from her acts. Beyond Eells, the issue is discussed in Ahmed (2014 Ch. 8) and Briggs (2010).

> **A screening problem.** An agent's decision problem $\langle \mathcal{A}, \mathcal{K}, Cr^t(\cdot) \rangle$ is a screening problem just in case the agent's evidential position at t makes accessible some Z which screens off all $a \in \mathcal{A}$ from all $k \in \mathcal{K}$, relative to $Cr^t(\cdot)$.

When some such Z is available pre-decision, the additional evidential value of actually acting—actually making a instead of $\neg a$ the case, or vice versa—is nothing. $Cr^t(k) = Cr^t_Z(k) = Cr^t_Z(k|a)$ for all a and k, so the characteristic decision rules of EDT and CDT will always endorse the same act.

Mood Candles is designed to suggest screening-off. There are two particularly appealing candidates for the proposition Z which would screen of L from D. The first will likely have occurred to the reader up above: it comes from Bob's ability to reflect on what he already believes about the problem he is facing. In *Mood Candles*, Bob believes that lighting up gives him an extra utile, and that it is causally independent from whether he gets the big prize. (As noted above, exactly how his doxastic state reflects this conviction about causal independence is a question I am currently leaving unanswered, but he *does* believe it.) It follows that Bob already has evidence that he is a member of a certain class: the class of people who are aware that lighting mood candles is efficacious for happiness. This is statistical evidence for D—by the description of the problem, only depressed people tend to be that informed about (small, but) effective ways of elevating one's mood. The first candidate "screener" proposition, then, supports a high credence in D. On this interpretation, Bob's mere middling credence in D looks rather like a bit of epistemic negligence.[12]

A second candidate is also introspective, but pulls in the other direction. It focuses on Bob's introspective access to how he feels, rather than what he believes. D, after all, is the proposition that *Bob is depressed at* t. Surely agents have some privileged access to whether they are depressed, which comes from concentrating on how they feel. By the description of the problem, though, Bob doesn't feel strongly depressed.[13] Reflection on this supports a *low* credence in D. On *this* interpretation, Bob's mere middling credence in D might still be remiss—but because it should be *lower* than it is, rather than higher.

Of course, there is not really an "either-or" in this case: Bob should be epistemically responsible in taking into account *both* what he (already) believes and how he (already) feels. Bob may be in such a singular evidential situation vis-à-vis

[12] In the Newcomb literature, this thought finds an ally in argument that if an agent can reflect that her "decision algorithm" recommends two-boxing, she can already conclude that the predictor left no money in the opaque box. For a version this argument, see Yudkowsky (2010 pg. 26 ff.).

[13] I construe this to be underlying the fact that $Cr^t(D)$ isn't high, in the original description of the case.

his total evidence that it is just difficult to say what a justified credence in D would be.

This is why the suggestion that real-life NC problems are all screening problems has had a equivocal effect on the literature. Briggs weighs in on the thesis as follows:

> Whether the tickle defense rules out all cases of conflict between EDT [and] CDT...[is] controversial. Perhaps agents can be reasonable enough to look to decision theory for advice without being as rational and self-aware as [screening-off] requires...[However,] it seems clear that a *great many* suitable agents will have information that renders their actions evidentially irrelevant to the dependency hypotheses. To the extent that situations [to the contrary] are rare, the task of deciding among various possible decision theories is rendered much less urgent. (Briggs 2010, pg. 28, emphasis added)

Two nondenominational morals, I think, emerge from considerations of the screening-off debate. The first is that screening-off arguments in NC Problems will always support the dominant option–that is, the act prima facie favored by the CDT, rather than the EDTer.[14] The second is that an agent's pre-decision reflection on her full information—an activity I will call *epistemic entrenchment*—must be considered by *both* theories in any putative NC problem, even if full screening-off is controversial. This often leads to subtle disputes about what introspection can deliver, from the agent's point of view.

For CDTers, entrenchment serves as a reminder that causalists do not actually ignore their evidence. In NC problems, dominance reasoning can obscure the fact that even a CDT-rational agent must have some *particular* credence over each dependency hypothesis in $\mathcal{K}$ before she acts. Decision procedure aside, she will not make the best decisions she could make if these credences fail to reflect her total evidence.[15]

[14] Why? In a 2 × 2 NC Problem, let $\mathcal{A} = \{a^*, \neg a^*\}$ and $\mathcal{K} = \{k_1, k_2\}$. By screening off, there is some Z which the agent can update on before acting such that $\forall k$: $Cr_Z^t(k|a) = Cr_Z^t(k)$. It follows from this that $Cr_Z^t(k|a^*) = Cr_Z^t(k|\neg a^*)$ for all $k \in \mathcal{K}$. Calling this number n, we have $EEU(a^*) = nVal(a^* \wedge k_1) + (1-n)Val(a^* \wedge k_2)$, and $EEU(\neg a^*) = nVal(\neg a^* \wedge k_1) + (1-n)Val(\neg a^* \wedge k_2)$. By dominance, $Val(a^* \wedge k_1) > Val(\neg a^* \wedge k_1)$ and $Val(a^* \wedge k_2) > Val(\neg a^* \wedge k_2)$. Hence $EEU(a^*) > EEU(\neg a^*)$. This is not an embarrassment to the EDTer's theory, since screening off makes going for dominant options compatible with that theory.

[15] For a similar observation about how dominance arguments can obscure a CDTer's evidential duties, see Joyce (2012 pg. 239). Here I take the traditional view that epistemic norms such as the Principle of Total Evidence are *sui generis* principles of epistemic rationality, whose justification is prior to the justification of any practical maxims they inform (see, for example, (Hare and Hedden 2016 pg. 615)). For argument that inverts this order of priority—offering a practical vindication of the Principle of Total Evidence—see Good (1967); for an argument that EDT, in contrast to CDT, has trouble underwriting Good's Theorem, see Skyrms (1990).

EDTers, for their part, need epistemic entrenchment to combat the objection that their view recommends patently absurd acts that would apparently result from giving too much weight to general statistical correlations. Consider Pearl's complaint:[16]

> [According to EDT,] [p]atients should avoid going to the doctor to reduce the probability that one is seriously ill; workers should never hurry to work, to reduce the probability of having overslept...remedial actions should be banished lest they increase the probability that a remedy is indeed needed. (Pearl 2000 pg. 108-9)

The reasonable reply here, on behalf of the evidentialist, appeals to an agent's need to incorporate her full information into her credal state before she acts. Focus on Pearl's example of hurrying and lateness (which we can call *Hurrying to Class*). Grant Pearl's assumption, that hurrying (e.g. to a 10am class) on any given day is correlated with lateness on that day:

Hurrying to Class. My outcome matrix and credences are as follows:

	Not Late ($\neg L$)	Late (L)
hurry (H)	9	0
don't hurry ($\neg H$)	10	1

$$Cr^t(\text{late} \mid \text{hurry}) > Cr^t(\text{late})$$

Why doesn't the EDTer's norm recommend that I avoid hurrying today, right now, as I head to campus?

In the simplest case, I have a timekeeping device with me, and either do, or easily could, glance at it, to see that it is (e.g.) 9:46am. Relative to any specific time—*a fortiori*, relative to 9:46am—the correlation between lateness and hurrying is reversed:

$$Cr^t(\text{late} \mid \text{hurry} \land 9{:}46\text{am}) < Cr^t(\text{late} \mid 9{:}46\text{am})$$

Hence relative to my full information, hurrying is *not* contraindicated by my act-conditionalized credences.

What happens if I do not have a watch? It seems that the same reasoning ought to hold—after all, forebearing to hurry when one is unsure of the time seems like an even *worse* idea than forebearing to hurry when one *does* know the time.

Here is a way to generalize the strategy. The argument sketched above clearly did not depend on the time's being exactly 9:46. Rather, it depended on the fact that the agent (here, me) was located at *some* particular time, which she could think about in a direct way. Suppose that agents generally have the power to pick

[16] I take the quote from Pearl via Ahmed (2014 pg. 82).

the current times rigidly by means of some demonstrative token—say, μ.[17] Relative to my full information in *Hurrying to Class*, I accept

$$Cr^t(\text{late} \mid \text{hurry} \wedge \text{the time is } \mu) < Cr^t(\text{late} \mid \text{the time is } \mu)$$

Just in virtue, then, of the patently introspectable fact that the time is μ, an EDT agent in a problem like *Hurrying to Class* can conclude that there is no in-respect-of-lateness reason not to hurry. I will return to this argument in §4.

2.2 Conditionalization and Lewis's Package

Up above, I described credence by saying that agents are presumed to have Bayesian dispositions—including a disposition to update by conditionalizing on the any new evidence they receive. This presentation, however, was neutral with respect the normative status of that tendency. Bayesianism's additional, normative claim is that conditional probabilities provide a *diachronic norm* of belief revision—that an agent *ought* to revise her credence in light of evidence E by moving from $Cr^t(\cdot)$ to $Cr^{t+}(\cdot) = Cr^t(\cdot\,|E)$. I will call this norm *Conditionalization* (with a capital 'C'.)

Conditionalization entails, as a special case:

(CNC) In an NC problem, one ought to update by conditionalizing on one's chosen act a.

(CNC) appears to dovetail well with the CDTer's conciliatory position on news values. Relative to a fixed credal state, propositions like L—*a mood candle is lit*—or H—*the agent hurries*—carry a stable news value both before, and after, that proposition is an available act. The CDT position is simply that, when e.g. L *is* an available option, it should be *chosen* (or passed up) according to its causal expected utility rather than its news value.

Lewis himself seems to have embraced (CNC).[18] The conjunction of (CNC) and the CDTer's expected utility equation comprise a joint CDTer position which we can call "Lewis's package" (LP):

(LP) At all times t, one ought to:

(i) assign expected utility to an available act a using one's time-t credences across the dependency hypotheses $\mathcal{K}$, selecting the CEU-maximal act;

[17] See Moss (2012)'s "Dr. Demonstrative" example (§1) for a similar maneuver, in the case of self-locating beliefs.

[18] In Lewis (1976 pg. 302), he writes that "there are good reasons why the change of belief that results from coming to know an item of new evidence should take place by conditionalizing on what was learned." A footnote from the quoted sentence refers the reader to Teller (1973)'s Dutch Book argument in favor of Conditionalization. Lewis advances his own Dutch Book for Conditionalization in Lewis (1999).

(ii) when making propositions in $\mathcal{A}$ true by one's actions, plan to evolve $Cr^t(\cdot)$ according to Conditionalization.

Conceptually, there is something odd about (LP). If an agent with Lewis's form of CDT is self-aware, she can *anticipate* what her credence in the various dependency hypotheses $k \in \mathcal{K}$ will be at any future time t^+ when she has performed a given available act a.[19] By Conditionalization, these future credences are just

$$\{Cr^t(k|a): k \in \mathcal{K}\}$$

By Equation 2, these are the same numbers the $EDTer$ will *currently* apply to calculate the expected value of a. In favoring CEU over EEU as a guide to action, then, the Lewisian CDTer *de facto* holds that one's current credences in the dependency hypotheses overrule the credences one plans to adopt. This seems in tension with the idea that there is something more informed about the more opinionated credence function $Cr^t(\cdot\,|a)$, *in virtue of which* it is the one an agent should plan (following Conditionalization) to transition to upon performing a.

This tension can be seen in the juxtaposition of norms governing action and belief change. For the Lewisian CDTer, a lack of causal efficacy makes it the case that one should, in choosing an act, be *unmoved*, so to speak, by the fact that the act provides statistical evidence for a desirable state. To return to our first example, in light of its lack of efficacy in promoting D, Bob should be *unmoved*, in his decision-making, by the fact that L provides good statistical evidence for D. But because the Lewisian CDTer also thinks Conditionalization applies, he believes that if Bob *does* bring about L, he must, in a way, be "moved" by its evidential force after all: Bob is rationally required to increase his confidence in D upon processing the fact that he made L true.[20]

3. The Newcomb Hospital

Can this be right? The present section will be devoted to sketching an odd consequence of (LP), drawing on an analogy with Parfit's amnesia cases. The setting is a hospital funded by an eccentric billionaire, where amnesiac patients—beset by various handicaps—get to face daily Newcomb puzzles.

Suppose you wake up in the Newcomb hospital today, knowing that you faced Newcomb's problem yesterday, but unable to remember what you did. You are completely evidentially indifferent on the matter, assigning a probability 1/2 to

[19] And has learned nothing more: a is, at t^+, her *total* evidence.

[20] Again, Lewis seems to have been comfortable with this result. Addressing Newcomb, he writes with an air of resignation that a two-boxer must expect one-boxers to come out ahead: "[w]e [two-boxers]...did not plead surprise. We knew what to expect." (Lewis 1981b pg. 378). See Byrne and Hájek (1997) for more discussion of Lewis's package view.

the proposition that you one-boxed and a probability 1/2 to the proposition that you two-boxed. The news that you did the former would be a near-to-perfect indication that you made $1,000 yesterday, and the news that you did the latter would be a near-to-perfect indication that you made \$1M yesterday. Hence in your state of total ignorance you estimate your winnings at $500,500. A Lewisian rational amnesiac, while a two-boxer, much prefers the news that she one-boxed in the past.

A trade reflecting this preference can be engineered.

> **Newcomb Past**. Fry (a CDTer) awakens, bedridden and with amnesia, in the Newcomb Hospital. Today is Wednesday, and Fry is in bed 336. Bender, Fry's hospital roommate, in bed 335. While Fry can't remember whether he one-boxed on Tuesday, Bender truthfully tells him that *he*, Bender, one-boxed on Tuesday. Fry's winnings from yesterday are in a lockbox marked "336-Tuesday". Bender's are in a lockbox marked "335-Tuesday". For \$$\Delta$, Bender offers to switch lockbox keys with Fry (call this trade σ).

What is the value σ? Because Fry conditionalizes on his evidence, when Bender tells him that he one-boxed yesterday, Fry becomes 90% confident that Bender's lockbox, 336-Tuesday, has $1M in it. He estimates the value of his *own* lockbox at a mere \$500,500. So σ, the deal Bender is offering, is attractive unless the amount of money he demands for the trade, Δ, is greater than \$399,500.

The attractiveness of the trade in *Newcomb Past* reflects the fact that, on Lewis's picture, it is very fine thing to have the past of a one-boxer. Of course, it is also a fine thing to *be* a two-boxer, since two-boxing maximizes causal expected utility. We can easily engineer a bet whose appeal reflects this preference as well, by inducing some uncertainty over which act is being performed:

> **Newcomb Present**. It's still Wednesday in the hospital with daily Newcomb rounds. But now it's time for Fry to choose today's move. It works like this: the predictor has already deposited either $1M or nothing in opaque lockbox "Wednesday-336", depending on whether she predicted that Fry would, today, choose act $2B$ (which dumps an extra thousand into Wednesday-336). Likewise, she has already deposited either $1M or nothing in Bender's lockbox, "Wednesday-335", depending on whether she predicted Bender would perform $2B$ (which dumps an extra thousand into Wednesday-335).
>
> Unfortunately, Fry has an injury that makes speech impossible, and due to a mix-up, the nurse incorrectly believes he is a monolingual Dingbats speaker. The nurse gives him the day's Newcomb menu with two options printed entirely in Dingbats: he can either choose "@" or "#". He does not know which one corresponds to $2B$. In frustration, Fry randomly circles the first option, "@".
>
> Bender's form is in English—he can either choose "$1B$" or "$2B$"—and it is still blank. For \$$\Delta$, Bender offers to switch forms with Fry (call this trade τ).

How should Fry evaluate τ? The value of the trade depends on what he would do with Bender's form if he got it. But that part is easy: from the Causal Point of View, a two-box form in a Newcomb problem is always worth $1000 more than a one-box form (it ensures that the extra thousand is dumped into one's box). So a CDTer like Fry, who chooses τ and gets a blank form will certainly go on to pick the two-box option. Indeed, the trade Bender is offering is appealing to him so long as $\Delta < \$500$.

Argument: let β be the amount of money already in the box. As in the original Newcomb puzzle, because the identity of β is causally independent of anything Fry does now, we can frame the problem from the CDTer’s point of view via an argument that abstracts from the value of β:

	$p\ (@ = 1B)$	$\neg p\ (@ = 2B)$
τ	β + (1000-Δ)	β + (1000-Δ)
$\neg\tau$	β + 0	β + 1000

If Fry is completely indifferent on the identity of @, then $Cr(p)$ =.5. Hence

$$\begin{aligned} CEU^t(\tau) &= .5(\beta + (1000 - \Delta)) + (1 - .5)(\beta + (1000 - \Delta)) \\ &= \beta + (1000 - \Delta) \\ CEU^t(\neg\tau) &= .5(\beta + 0) + (1 - .5)(\beta + 1000) \\ &= \beta + 500 \end{aligned}$$

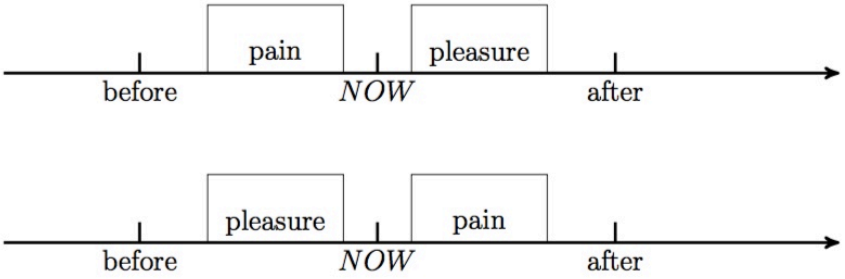
Hence $CEU^t(\tau) > CEU^t(\neg\tau)$ iff $\Delta <$ \$500. (For a presentation of the argument without the variable β, see Appendix.)

3.1 Putting it all together

Our last step to getting the preferences reflected in Figure 3 is to combine the betting behavior of the CDTers in *Newcomb Past* and *Newcomb Present*. The following setup, while baroque, will do the job:

> **Newcomb Hospital Fusion.** As before, it’s Wednesday. The lockboxes, Fry’s amnesia, and mutual knowledge that Bender one-boxed yesterday are as in *Newcomb Past*. As in *Newcomb Present*, Fry has been given a Dingbats form for today’s round and has randomly marked “@" on it, while Bender was given a form in English. Bender offers Fry a “combo deal": for \$$n$, he will switch lockbox keys *and* forms with Fry (call this trade ω).
>
> Before Fry makes his decision, he learns one additional piece of information from the Dingbats-language chart by his bed: yesterday, the act he chose was #.

How will Fry weigh the combo deal? The causal expected utility of choosing “@"on this round is equal to the causal expected utility of choosing “#" given his current credences, since each act has a 50% chance of being the act of one-boxing and a 50% chance of being the act of two-boxing. Because he knows he chose # yesterday and has currently marked @ on his form, he knows, even before he trades, that he is one of the two situations in Figure 3. For an additional \$$\Delta$—by choosing ω—he can guarantee the outcome a Lewisian CDTer confidently desires: that he has a one-box-*past* and two-box-*future* as opposed to the reverse.



Newcomb Time Bias: Figure 3, repeated.

As in *Newcomb Present*, the part of the combo deal that ensures a two-boxing present is appealing. Even *more* appealing is the the part that mimics being able to ensure a one-boxing past—that is, Fry's being able to trade his Tuesday box for Bender's. Since Bender one-boxed yesterday, his box, box 335-Tuesday, is 90% certain to contain a million dollars. Indeed, as long as $\Delta < \$400{,}500$, Fry will choose to trade.

Argument: let β be the amount of money that is already in Fry's *Wednesday* box—that is, the box he will get today. No matter what, Fry is not trading away his Wednesday box (rather, if he takes ω he switches *Tuesday* boxes and Wednesday *forms* with Bender.) So the value of β—though it is act dependent—is, again, causally independent from what Fry chooses to do.

If Fry chooses ω, he switches forms with Bender, so he insulates his own payoff from the identity of @; he will get $\beta + (1000 - \Delta)$ today in any case. Fry's payoffs will then further depend on whether Bender's Tuesday box contains $\$1M$ or nothing (let these be propositions $5T1$ and $5T0$, respectively).

If Fry declines the combo deal, opting for $\neg\omega$, he keeps the contents of his own Tuesday box, and his payoffs will then depend on whether his Tuesday box contains $\$1M$ or nothing (let these be propositions $6T1$ and $6T0$, respectively). Moreover, his total payoff does now depend on the identity of @—he will get an extra thousand just in case $@ = 2B$ (row 2, below).

	I. $5T1, p, 6T1$	II. $5T1, p, 6T0$	III. $5T1, \neg p, 6T1$	IV. $5T1, \neg p, 6T0$
ω	$1M + (1000 - \Delta) + \beta$	$1M + (1000 - \Delta) + \beta$	$1M + (1000 - \Delta) + \beta$	$1M + (1000 - \Delta) + \beta$
$\neg\omega$	$1M + 0 + \beta$	$0M + 0 + \beta$	$1M + 1000 + \beta$	$0M + 1000 + \beta$
	V. $5T0, p, 6T1$	VI. $5T0, p, 6T0$	VII. $5T0, \neg p, 6T1$	VIII. $5T0, \neg p, 6T0$
ω	$0M + (1000 - \Delta) + \beta$	$0M + (1000 - \Delta) + \beta$	$0M + (1000 - \Delta) + \beta$	$0M + (1000 - \Delta) + \beta$
$\neg\omega$	$1M + 0 + \beta$	$0M + 0 + \beta$	$1M + 1000 + \beta$	$0M + 1000 + \beta$

Utilities.

What of the appropriate causalist credences? Knowing Bender one-boxed yesterday, Fry is 90% confident in $5T1$, the proposition that the predictor put $\$1M$ in Bender's box on Tuesday. Hence he has credence 90% across columns I-IV and credence 10% across columns V-VIII. Within each of these possibilities, he is 90% confident the predictor predicted his own move (#) correctly yesterday, so he expects a correlation of .9 between $\neg p$ (which entails that # is one-boxing) and $6T1$, the proposition that the predictor put $\$1M$ in 336-Tuesday. Finally, his overall confidence in p is .5. That yields:

	I. $5T1, p, 6T1$	II. $5T1, p, 6T0$	III. $5T1, \neg p, 6T1$	IV. $5T1, \neg p, 6T0$
ω	0.045	0.405	0.405	0.045
$\neg\omega$	0.045	0.405	0.405	0.045
	V. $5T0, p, 6T1$	VI. $5T0, p, 6T0$	VII. $5T0, \neg p, 6T1$	VIII. $5T0, \neg p, 6T0$
ω	0.005	0.045	0.045	0.005
$\neg\omega$	0.005	0.045	0.045	0.005

Credences.

What of the appropriate causalist credences? Knowing Bender one-boxed yesterday, Fry is 90% confident in $5T1$, the proposition that the predictor put $\$1M$ in Bender's box on Tuesday. Hence he has credence 90% across columns I-IV and credence 10% across columns V-VIII. Within each of these possibilities, he is 90% confident the predictor predicted his own move (#) correctly yesterday, so he expects a correlation of .9 between $\neg p$ (which entails that # is one-boxing) and $6T1$, the proposition that the predictor put $\$1M$ in 336-Tuesday. Finally, his overall confidence in p is .5. That yields:

A calculation shows that $CEU(\omega)$ = 900,000 + (1000 - Δ) over the baseline β, while $CEU(\neg\omega)$ = \$500,500 in excess of the same baseline. Hence Fry will take the deal as long as Δ <\$400,500.

3.2 Karma Foretold

From the Lewisian point of view, time-biased trading at the Newcomb Hospital seems to be motivated. Of course the causal decision theorist wants the *past* of a one-boxer—this is an excellent indicator that she is already a millionaire! And of course she wants to two-box *today*—that guarantees that she walks away with \$1,000 more than she would have gotten otherwise![21]

[21] I here use the counterfactual locution Lewis favored in describing causalist reasoning; see, for example, (Lewis 1981b).

These sentiments are articulated from the point of view of a particular time—for example, the time labeled "*NOW*" in Figure 3. But at the point in time marked "after"—when both rounds lie in the past—taking the combo deal will seem, by the CDTer's own lights, like bad news. More to the point, I fear, it looks like an *irrational* decision. Suppose Fry wakes up amnesiac on *Thursday*, and is told that Bender offered him the combo deal on Wednesday. On Wednesday, Fry could have kept his @ form for free, ensuring that the action he performed on Wednesday was different in kind than the action he performed on Tuesday—hence, that he was in *one* of the scenarios in Figure 3, without knowing which. Now, for Fry to learn whether he took Bender's combo deal is to learn either that:

> (a) In the last two days, he two-boxed once and one-boxed once, in some unknown order, or
>
> (b) In the last two days, he one-boxed once and two-boxed once, and willingly gave Bender $400,000 to ensure that the first happened before the second.

Though I am a CDTer, the idea of learning that *I* secured (b) over (a) in such a situation makes me uncomfortable. It is hard to justify the claim that one way of ordering my acts in time is worth (more than) $400,000 more than the other.[22]

4. CDT's response

The oddity in *Newcomb Hospital Fusion* arises because the tie between one-boxing and being a millionaire is invisible to Fry's deliberations at the moment of choice—only to reappear with a vengeance later, when he assesses what he believes is in yesterday's lockboxes. This makes it impossible to stably assume that Fry does, or doesn't, take the fact that an agent is a one-boxer to be *good reason* to increase one's confidence that that agent is a millionaire.

In most ordinary situations, there does seem to be a stable fact. In *Mood Candles*, for example, the causalist's intuition is that Bob should light a candle even though candle-lighting is—at least, in the general population—correlated with

[22] It is worth noting that, while the post-facto news value of (a) exceeds that of (b), *Newcomb Hospital Fusion* is not a Dutch Book: choosing (b) over (a) does not *guarantee* that you are poorer in every possible world (though it does make it likely). A world *w* will make the agent better off, where

- 335-Tuesday contains $0 in *w*,
- 336-Tuesday contains $1 million in *w*, and
- @ is one-boxing.

For a diachronic Newcomb problem for the two-boxer that *does* make a standard CDTer poorer in every possible world, and the extra assumptions this involves, see Ahmed's "Newcomb Insurance" cases (Ahmed 2017, 2014 pg. 202 ff.).

depression. It also seems like, if he does light up, he *fails* to acquire a good reason to increase his confidence that he is depressed. At least, that is what I suggested above: there is something odd about thinking, with Lewis, that if Bob does light up, he would thereby acquire good evidence for thinking he was depressed after all. He *knows* why he's contemplating lighting up—he is in pursuit of the extra utile, which the act will secure him—and it's not *because* he is depressed.

4.1 Time Bias and EDT

As indicated in the introduction, my first defensive maneuver is to argue that EDT is *also* susecptible to time bias. It will be up to the reader to determine whether the EDTer's form of bias is as unflattering as Fry's.

Recall that in the Pearl-inspired example *Hurrying to Class* (§2.1), hurrying on any given day was generally correlated with lateness on that day:

$$Cr^t(\text{late} \mid \text{hurry}) > Cr^t(\text{late})$$

Nonetheless, relative to a specific time—like 9:46am—that correlation failed (in fact, was reversed), so that relative to the agent's full available information hurrying was *not* after all contraindicated, by either decision theory. For cases where the agent lacked a wristwatch or other timekeeping device, I extended the screening-off argument by appeal to introspection—to the agent's ability to "lock on" to her current time by means of some demonstrative μ, and use Z = *the time is* μ to screen off the support lent to lateness by hurrying.

This ability, which extends only to the present moment, gives rise to time-bias. Suppose an EDTer, who has just woken at t unable to remember what she has done for the last month (the interval of time "T^-"), finds herself in an iterated version of *Hurrying to Class*. If she cares about her track record of arriving on time to class, she may well:

(i) prefer to hurry at t, and

(ii) prefer to learn that she *didn't* hurry at any past time $t' \in T^-$.

The EDTer in *Hurrying to Class* should have attitude (i) for the reasons explained above: overturning Pearl's argument depends on her leveraging her (difficult-to-express) knowledge *that it is now the current time*. But this power is limited to t: with regard to her past, it is rational to continue to maintain attitude (ii), *despite* taking up attitude (i). Though something similar to the thought involving μ was true in the past—after all, for any morning in the agent's past *it was that time then* as much as *it is this time now*—the proposition expressed by the former thought does not have the force, at t, that the latter thought does. EDT's appeal to screening off in many putative NC problems will commit the view to time bias in cases where, as in *Newcomb Hospital*, those NC problems are iterated.

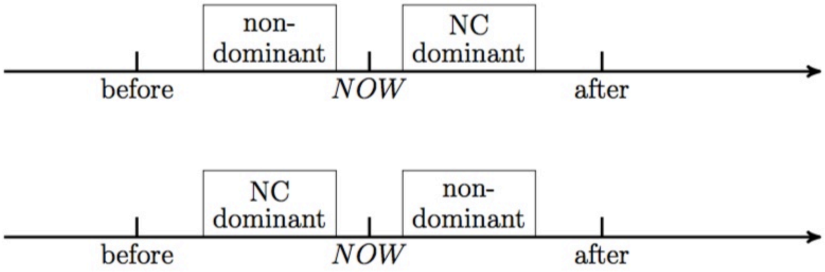


Figure 4: A Shared Structure (top preferred to bottom).

It is worth thinking about why this is so, and what this means for Lewis's views on updating. At issue in Lewis's Package is whether an agent—for example, in an NC Problem—should *plan* to conditionalize on the act she performs. One way of thinking about this is to ask whether, in an NC problem, your attitude towards the acts $a' \in \mathcal{A}$ you are currently contemplating should be the same as the preferences you would have if you woke up tomorrow with amnesia and were about to be *told* which $a' \in \mathcal{A}$ you performed.

But introspection's power to create and discover propositions concerning the time t of choice is sharply different across the two situations. In *Hurrying to Class*, I can think to myself, *the time is* μ. Oddly, at the time of choice, I seem to gain some knowledge by this exercise—knowledge that affects my rational calculations. Later, post-amnesia, I will not have this power. It is patently uninformative to be told, "you reflected *then* that the time was ν, where by ν you meant: the time it in fact was."

A comparison between this situation and a CDTer's reasoning in NC Problems is telling. In an NC problem, I can tell myself, "I will now choose the dominant option because of the (small) benefit I get from it". I am introspecting my motives, which give me a reason to perform a^* *without* giving me an extra reason to think that I won't get the big prize. If I am subsequently subjected to amnesia, later I will only be able to learn that I either performed the dominant option or I didn't. Any extra knowledge I had access to at the earlier time will be wiped out: it feels uninformative to be told "you performed the dominant option because it was the dominant option". While not *vacuous*, this thought seems, again, *obvious*—like the thought that if you hurried, you hurried at some particular time.

5. Conclusion

I began by arguing that CDT is time-biased, and going on to show that this leads to odd choices in the Newcomb Hospital. In response to this bias, I leveled a *tu quoque* against EDT. Time-bias may be a pervasive feature of decision-making, not a weakness in any particular theory of it.

It would be best to close by returning to Newcomb itself. EDTers may feel that, in shifting the focus to an agent's the act-prior viewpoint, and the epistemic reasons that can be articulated from it, I have quietly sidestepped the real bite of the original problem. For in that puzzle, whether it is possible, *before you act*, to come to an evidentially supported credence in the $\mathscr{K}$-partition {*Predictor predicted you'd one-box, Predictor predicted you'd two-box*} (={$P1$, $P2$}) is precisely what is at issue. There is a widespread tendency in the literature to assume

> (P) In Newcomb's puzzle, the only evidence available with respect to distributing credence over the partition {$P1$, $P2$} is how the player him/herself reasons about the expected utility of one-boxing.

This would apparently render full reflection, and any attempt at screening off, impossible. Ahmed, for example, presses the point that "[a]nyone facing Newcomb's problem has *no* evidence that relevantly distinguishes him now from anyone else whom the statistical generalization [$Pr(P1|1B)$ and $Pr(P2|2B)$ are high] covers; that is, all other persons who ever face this problem" (Ahmed 2014 pg. 191). This suggests (P)'s epistemological moral—that there is just no way, in Newcomb, to even *try* to take your full evidence in to account before you act, since nothing *counts* as relevant evidence other than the act itself.

But while (P) might characterize *Newcomb's Problem*, it *isn't* supported, either by deductive or abductive considerations, by the basic mechanics of an NC problem— where the latter is described simply in terms of causal dependency hypotheses, credences, and outcome states, and where it is the latter that usefully diagnoses the split between causal and evidential decision theory.

6. Appendix: *Newcomb Present*: Long argument

As in the original Newcomb problem, Fry's estimate how much money is already in his own box—the quantity called "β" in the main text—is act-dependent. It depends, in part, on how likely he thinks it is that he *will* take the trade. Fry knows that if he chooses to trade, then he will certainly two-box. The predictor will have predicted this with her usual accuracy, so that $Cr^t(P1|\text{trade}) = .1$ and $Cr^t(P2|\text{trade}) = .9$. However, if Fry declines to trade, sticking with the wildcard move @, then the probability that the predictor predicted $P1$ is equal to the probability that the predictor predicted $P2$ (the high correlations of $P1$ with $1B$ and $P2$ with $2B$ are now balanced by the fact that Fry has a 50% chance of doing $1B$ and a 50% chance of doing $2B$.)

$\mathcal{K} = \{(P1 \wedge p), (P1 \wedge \neg p), (P2 \wedge p), (P2 \wedge \neg p)\}$ is a dependency hypothesis partition.

	$P1 \wedge p$	$P1 \wedge \neg p$	$P2 \wedge p$	$P2 \wedge \neg p$
τ	$1M + (1000 - \Delta)$	$1M + (1000 - \Delta)$	$0M + (1000 - \Delta)$	$0M + (1000 - \Delta)$
$\neg\tau$	$1M + 0$	$1M + 1000$	$0M + 0$	$0M + 1000$

Utilities.

We have:

$$\begin{aligned}
Cr(k_1) = Cr(P1 \wedge p) \quad &= Cr(p)Cr_p(P1) \\
&= Cr(p)[Cr_p(P1|\tau)Cr_p(\tau) + Cr_p(P1|\neg\tau)Cr_p(\neg\tau)] \\
&= Cr(p)[.1Cr_p(\tau) + .9Cr_p(\neg\tau)] \\
&= .5[.1Cr_p(\tau) + .9Cr_p(\neg\tau)] \\
&= .5[.1Cr(\tau) + .9Cr(\neg\tau)]
\end{aligned}$$

$$\begin{aligned}
&Cr(k_2) = Cr(P1 \wedge \neg p) \\
&= Cr(\neg p)Cr_{\neg p}(P1) \\
&= Cr(\neg p)[Cr_{\neg p}(P1|\tau)Cr_{\neg p}(\tau) + Cr_{\neg p}(P1|\neg\tau)Cr_{\neg p}(\neg\tau)] \\
&= Cr(p)[.1Cr_{\neg p}(\tau) + .1Cr_{\neg p}(\neg\tau)] \\
&= .5[.1Cr(\tau) + .1Cr(\neg\tau)] \\
&= .5(.1) = .05
\end{aligned}$$

$$\begin{aligned}
Cr(k_3) = Cr(P2 \wedge p) \quad &= Cr(p)Cr_p(P2) \\
&= Cr(p)[Cr_p(P2|\tau)Cr_p(\tau) + Cr_p(P2|\neg\tau)Cr_p(\neg\tau)] \\
&= Cr(p)[.9Cr_p(\tau) + .1Cr_p(\neg\tau)] \\
&= .5[.9Cr_p(\tau) + .1Cr_p(\neg\tau)] \\
&= .5[.9Cr(\tau) + .1Cr(\neg\tau)]
\end{aligned}$$

$$\begin{aligned}
&Cr(k_4) = Cr(P2 \wedge \neg p) \\
&= Cr(\neg p)Cr_{\neg p}(P2) \\
&= Cr(\neg p)[Cr_{\neg p}(P2|\tau)Cr_{\neg p}(\tau) + Cr_{\neg p}(P2|\neg\tau)Cr_{\neg p}(\neg\tau)] \\
&= Cr(p)[.9Cr_{\neg p}(\tau) + .9Cr_{\neg p}(\neg\tau)] \\
&= .5[.9Cr(\tau) + .9Cr(\neg\tau)] \\
&= .5(.9) = .45
\end{aligned}$$

For shorthand, we stipulate that $\mathfrak{t} = Cr(\tau)$. Hence Fry's credences are as follows:

$k1$	$k2$	$k3$	$k4$
$.45 - .4\mathfrak{t}$	$.05$	$.4\mathfrak{t} + .05$	$.45$

Credences, take I.

Letting $z = .4\mathfrak{t}$, this is equivalent to:

$k1$	$k2$	$k3$	$k4$
$.45 - z$	$.05$	$z + .05$	$.45$

Credences, take II.

Calculating causal expected utilities:

$$\begin{aligned}
CEU^t(\tau) &= (.45 - z)(1M + (1000 - \Delta)) + .05(1M + (1000 - \Delta)) \\
&\quad + (z + .05)(0M + (1000 - \Delta)) + .45(0M + (1000 - \Delta)) \\
&= (.5 - z)(1M) + (1000 - \Delta) \\
CEU^t(\neg\tau) &= (.45 - z)(1M + 0) + .05(1M + 1000) + (z + .05)(0M + 0) + .45(0M + 1000) \\
&= (.45 - z)(1M) + (.05)(1M) + (.05)(1000) + (.45)(1000) \\
&= (.45 - z + .05)(1M) + (.05 + .45)(1000) \\
&= (.5 - z)(1M) + 500
\end{aligned}$$

Once again, we have something of the form: $CEU^t(\tau) = X + (1000 - \Delta)$, while $CEU^t(\neg\tau) = X + 500$. Hence $CEU^t(\tau) > CEU^t(\neg\tau)$ iff $\Delta < \$500$. QED.

References

Ahmed, Arif. 2014. *Evidence, Decision and Causality*. Cambridge University Press.

———. 2017. “Exploiting Causal Decision Theory.” Manuscript.

Arntzenius, Frank. 2003. “Some Problems for Conditionalization and Reflection.” *Journal of Philosophy* 100 (7):356–70.

Baratgin, Jean, and Guy Politzer. 2010. “Updating: A Psychologically Basic Situation of Probability Revision.” *Thinking & Reasoning* 16 (4):253–87.

Briggs, R. A. 2010. “Decision-Theoretic Paradoxes as Voting Paradoxes.” *Philosophical Review* 119 (1):1–30.

Byrne, Alex, and Alan Hájek. 1997. “David Hume, David Lewis, and Decision Theory.” *Mind* 106 (423).

Eells, Ellery. 1982. *Rational Decision and Causality*. Cambridge University Press.

Good, I. J. 1967. “On the Principle of Total Evidence.” *The British Journal for the Philosophy of Science* 17 (4):319–21.

Hare, Caspar, and Brian Hedden. 2016. “Self-Reinforcing and Self-Frustrating Decisions.” *Noûs* 50 (3):604–28.

Hedden, Brian. 2015. “Time-Slice Rationality.” *Mind* 124 (494):449–91.

Hesslow, Grant. 1976. “Discussion: Two Notes on the Probabilistic Approach to Causality.” *Philosophy of Science* 43 (2):290–92.

Joyce, James. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press.

———. 2007. "Are Newcomb Problems Really Decisions?" *Synthese* 156:537–62.

———. 2012. "Regret and Instability in Causal Decision Theory." *Synthese* 187:123–45.

Katsuno, Hirofumi, and Alberto O. Mendelzon. 1991. "On the Difference Between Updating a Knowledge Base and Revising It." In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*. Vol. 2.

Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.

———. 1976. "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review* 85.

———. 1981a. "Causal Decision Theory." *Australasian Journal of Philosophy* 59 (1):5–30.

———. 1981b. "Why Ain'cha Rich?" *Noûs* 15:377–80.

———. 1999. "Why Conditionalize?" In *Papers in Metaphysics and Epistemology*. Cambridge University Press.

Moss, Sarah. 2012. "Updating as Communication." *Philosophy and Phenomenological Research* 85 (2):225–48.

———. 2013. "Epistemology Formalized." *Philosophical Review* 122 (1):1–43.

———. forthcoming. *Probabilistic Knowledge*. Oxford University Press.

Nozick, Robert. 1981. *Philosophical Explanations*. Harvard University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Ramsey, Frank. 1990. "Truth and Probability." In *Philosophical Papers*, edited by D. H. Mellor. Cambridge University Press.

Savage, Leonard. 1972. *The Foundations of Statistics*. Dover.

Skyrms, Brian. 1990. "The Value of Knowledge." *Minnesota Studies in the Philosophy of Science* 14 (245-266).

Teller, Paul. 1973. "Conditionalization and Observation." *Synthese* 26:218–58.

Yudkowsky, Eliezer. 2010. "Timeless Decision Theory." http://www.intelligence.org.