Philosophers' Imprint

VOLUME 24, NO. 6
MARCH 2024

EVIDENTIAL DECISION THEORY AND THE OSTRICH

Yoaav Isaacs and
Benjamin A. Levinstein

Baylor University
University of Illinois Urbana-Champaign

doi.org/10.3998/phimp.3176

1. Introduction

Evidential Decision Theory is flawed, but its flaws are not fully understood. David Lewis (1981) famously charged that EDT recommends an irrational policy of managing the news and "commends the ostrich as rational". Lewis was right, but the case he appealed to—NEWCOMB—does not demonstrate his conclusion. Indeed, decision theories other than EDT, such as Cohesive Decision Theory and Functional Decision Theory, agree with EDT's verdicts in NEWCOMB, but their flaws, whatever they may be, do not stem from any ostrich-like recommendations.

We offer a new case which shows that EDT mismanages the news, thus vindicating Lewis's original charge. We argue that this case reveals a flaw in the "Why ain'cha rich?" defense of EDT. We argue further that this case is an advance on extant putative counterexamples to EDT.

2. EDT v. CDT

Both Evidential and Causal Decision theory agree you should maximize expected utility. The difference between them arises from how they calculate expected utility. The standard informal way to cash out this difference is as follows: According to EDT, you should evaluate acts based on the extent to which they *indicate* good outcomes, whereas according to CDT, you should evaluate acts based on the extent to which they *cause* good outcomes.[1]

To illustrate their differences, we begin with the familiar:

NEWCOMB You are confronted with two boxes, one transparent and one opaque. You can choose either to take the contents of both boxes or to take only the contents of the opaque box. The transparent box contains $1,000. The opaque box contains either nothing or $1,000,000, depending on a past prediction about what choice you would make. If it was predicted that you would take the contents of both boxes, then the opaque box contains nothing. If it was predicted that you would take the contents of only the opaque box, then the opaque box contains $1,000,000. This predictor is

1. For a more precise characterization which differentiates causation from causal dependence, see Hedden (2023).

known to be highly reliable. Should you take one box or two?

The Evidential Decision Theorist tells you to one-box. One-boxing is strong evidence you'll get $1M, whereas two-boxing is strong evidence you'll only get $1,000.

The Causal Decision Theorist says you should take both boxes. Either the money is in the opaque box or it isn't. It's too late to do anything about that now. And either way, you cause a better result by taking both.

Before diagnosing whether EDT's verdict stems from an irrational news management policy, it's worth exploring the difference between EDT and CDT more carefully.

For simplicity, we'll formulate EDT and CDT with the same framework. EDT and CDT both appeal to a set of acts $\mathcal{A}$, states $\mathcal{S}$, and outcomes $\mathcal{O}$. An act and a state jointly result in a unique outcome. Outcomes are objects of ultimate concern for an agent. If the agent would prefer world w_1 over w_2, then w_1 and w_2 are elements of distinct outcomes. We measure the desirability of an outcome with a real-valued function u unique up to positive affine transformation. The agent also comes equipped with a probability function Pr that measures her uncertainty over $\mathcal{A}$ and $\mathcal{S}$.[2]

To capture the difference between the two theories, we follow Gallow (2020). We can divide up a given state into factors that are causally downstream and causally upstream of your acts.[3] The downstream factors are exactly those over which you exert causal influence in a given state of the world. Call the upstream factors K and the downstream factors C. Then we can distinguish EDT and CDT as follows:

2. Some formulations of EDT dispense with the division of acts, states, and outcomes, and some formulations of CDT avoid probabilities over acts. Neither of these finer points makes a substantive difference to our discussion below. See Jeffrey (1983) for more on the finer points about EDT and see Hájek (2016) for more on the finer points about CDT.
3. By 'upstream', we mean not downstream.

$$\mathcal{V}(A) = \sum_K \Pr(K \mid A) \sum_C \Pr(C \mid KA) u(KCA) \tag{EDT}$$

$$\mathcal{U}(A) = \sum_K \Pr(K) \sum_C \Pr(C \mid KA) u(KCA) \tag{CDT}$$

This formulation of EDT and CDT brings out the fundamental difference between the two theories. EDT thinks you should consider how likely your act renders upstream factors ($\Pr(K \mid A)$), whereas CDT thinks you should only consider the unconditional probability of those factors ($\Pr(K)$). EDT favors maximizing the expected value of the information that you perform your action. For EDT, an act's expected value derives *both* from its causal contributions to what you value and from the evidence it provides that the underlying state of the world conduces to what you value. In contrast, for CDT an act's expected value derives *solely* from its causal contributions to what you value.

In Newcomb's problem, EDT doesn't care whether the presence or absence of $1M is upstream or downstream of your act, so it considers $\Pr(1M \mid 1B)$, $\Pr(1M \mid 2B)$, etc., when calculating $\mathcal{V}(1B)$ and $\mathcal{V}(2B)$. Thus EDT recommends one-boxing. But since whether there's money in the box is upstream of your act, CDT considers $\Pr(1M)$ and $\Pr(\neg 1M)$ when calculating $\mathcal{U}(2B)$ and $\mathcal{U}(1B)$. Thus, CDT recommends two-boxing.

This divergence famously led David Lewis (1981) to charge that EDT recommends an irrational policy of managing the news, alleging that it "commends the ostrich as rational". But the case that EDT is irrational and ostrich-like is questionable.

Admittedly, one can get oneself into the mood where it seems strange to consider $\Pr(K \mid A)$ when A is downstream of K. After all, A can't affect K! But on the other hand, one can get oneself into the mood where it doesn't. After all, if you're trying to determine how much utility you'd get from performing A, you only want to consider worlds where A is true. How likely K is in those worlds is just $\Pr(K \mid A)$. This is, in effect, just to articulate the different

fundamental intuitions behind EDT and CDT. EDT tells you to perform the act that gives you the best distribution over outcomes. CDT tells you to perform the act that gives you the best distribution over outcomes holding things outside of your control fixed. Put this way, it's far from clear that EDT's policy is irrational.

While there's much more to say theoretically, we don't think that EDT's verdict in NEWCOMB is enough to show that EDT mismanages the news. For one, it remains controversial what the right answer in NEWCOMB is.[4] Second, there are other decision theories that *don't* manage the news the way EDT does and that still recommend one-boxing. Functional decision theory, for instance, appeals to the decision procedure the agent uses.[5] According to FDT, one should consider what would happen if your decision procedure were to output different acts in the act space. FDT thinks of these procedures as abstract objects (like computer programs) that are not local to your own mind. If another agent is using or simulating the same procedure, then, on FDT's counterfactuals, the output of your decision procedure will vary for that agent too. According to functional decision theorists, moreover, you can control what your procedure outputs.

In NEWCOMB, FDT claims that if the predictor is accurate, then her choice is affected by the output of your decision procedure (even if you haven't yet decided). If your decision procedure were to output *one-box* when you run it, then it also would have output *one-box* when the predictor ran it.[6]

Structurally, FDT is very close to CDT, with two basic changes.[7] Whereas CDT divides states into factors that are upstream and downstream of the *act* itself, FDT divides states into factors that are upstream or downstream of your *decision procedure*. Since both the predictor's and your choice are influenced by the output of your decision procedure in NEWCOMB, the predictor's choice is downstream of your procedure but upstream of the physical action of selecting one or two boxes. Second, whereas CDT considers only *causal* influence, FDT has a broader notion of influence. Even though the predictor's choice is not *causally* influenced by anything you do, it is still influenced by something you have control over, namely, the output of your decision procedure.

Whatever the merits or demerits of FDT, it does not 'manage' the news in the way EDT does. The equation for FDT's notion of expected utility looks just like equation (CDT) above. The only difference is that what counts as an upstream factor (K) is different for FDT than it is for CDT.[8]

Therefore, causal decision theorists cannot charge FDT with mismanaging the news. They will charge that it delivers the wrong verdicts and appeals to the wrong counterfactuals and perhaps even that it has bad metaphysics. But the one-boxing of FDT is not ostrich-like, and so one-boxing is not automatically ostrich-like.[9]

This does not mean that EDT is not objectionably ostrich-like, or that EDT does not prescribe one-boxing for objectionably ostrich-like reasons. But it does mean that Newcomb's problem makes a poor diagnostic case for being objectionably ostrich-like. Dialectically, the case against EDT would be stronger if there were a case in which EDT gave a prescription which was more straightforwardly unreasonable and which other standard decision theories did not share.

4. For defenses of one-boxing, see Spohn (2012); Ahmed (2014); Horwich (1987); Horgan (1981); Levinstein and Soares (2020); Yudkowsky and Soares (2017).
5. See Levinstein and Soares (2020); Yudkowsky and Soares (2017).
6. If the predictor runs a simulation of your decision procedure, then the simulation still would have likely output *one-box* according to FDT. Note that the important thing is that the predictor's choice is somehow influenced by the output of the procedure you use to decide, even if the predictor herself doesn't 'run' it.
7. There are actually many different versions of FDT, but those differences need not matter to us. See Yudkowsky and Soares (2017).
8. Of course, one could criticize FDT for giving the wrong recommendations based on the news it does get, but that doesn't make it ostrich-like. The crux of Lewis' charge is that EDT wrongly recommends actions based not on causal effects, but instead on epistemic upshots.
9. Cohesive Decision Theory (Meacham, 2010) also prescribes one-boxing for reasons unrelated to news-mismanagement. Although the exact technical details are rather involved, the rough idea is that COHDT tells you to do whatever you would have wanted to bind yourself to do at the beginning of your life (and before any predictions were made). In NEWCOMB, you would have wanted to bind yourself to one-box before any predictions were made. In that way, whenever a prediction actually ends up being made, it's highly likely there will be money in the opaque box. So, COHDT tells you to one-box because one-boxing conforms to a hypothetical prior plan, not because one-boxing is good news.

3. A New Case

To show that invoking $\Pr(K \mid A)$ instead of just $\Pr(K)$ when calculating expected utility is irrational, we provide a new case.

Consider:

Torture John has been abducted by a fiendish organization. His captors flip a fair coin in private. If the coin lands *H*eads, John will eventually be set free unharmed. If it lands *T*ails, he'll be brutally tortured. Before John learns his fate, his captors place him in a cell and subject him to two rounds of the following decision problems. In round 1, if the coin lands Heads, John will see a Green light flash with 90% probability and a Red light flash with 10% probability. If it lands Tails, he'll see a Red light flash with 90% probability and a Green light flash with 10% probability. If he sees a Green light, he has no decision to make. If he sees a Red light he'll then be offered a choice to pay $1 to rig the lighting device so that he'll be sure to see a Red light in any future round. (So, if John sees Red in round 1, and John pays, then he'll see Red in round 2. If he sees Red in round 2 and pays, then he simply loses the dollar.) After making this decision, his memory will be erased. John is certain he will always decide the same way whenever he sees a Red light. John is in his cell and sees a Red light. John cares a little bit about money, but much more about not being tortured. What should he do?

John obviously shouldn't pay. However, EDT mandates that he does pay.

If John doesn't pay, then he'll believe to degree .9 that the coin landed tails, and he'll be tortured.

If John pays, then he knows the sequence he observes over the two rounds is (or will be) either *RR* or *GR*. If the sequence is *RR*, then there's a 90% chance he'll be tortured, since in the first round the probability of *T* given that the light was Red is .9, but the second round's reading was meaningless. If the sequence he sees is *GR*, then there's only a 50% chance he'll be tortured since he saw one *G* and one *R* that are equally well correlated with *H* and *T* respectively. So, assuming he's not certain he's in round 1, then upon seeing Red, his credence will be somewhere strictly between .5 and .9 that he'll be tortured. So, by paying he lowers the probability of being tortured. (We assume this difference is big enough on his utility function to trump the small amount of money he loses.)

By paying, John is playing the ostrich. He's merely changing the *information* he gets from the Red signal, but he's not actually doing anything about the possibility of upcoming torture.[10] In other words, he's merely managing the news, and he's paying $1 (or $2) for the privilege.[11]

CDT and FDT agree that John shouldn't pay. According to CDT, John can't do anything to change how the coin landed, so he may as well save his money. According to FDT, John's decision procedure that tells him to pay or not to pay has no effect (causal or otherwise) on whether the coin lands Heads. Both theories—though wildly different in orientation—agree that paying is dominated by not paying. The only value that paying has is news value.

Note that this is a different sort of case from others where EDT will pay to avoid information to protect against future decisions. For example:

Optional Newcomb As in Newcomb, you are confronted with two boxes. The transparent box has $1,000. The opaque box contains either nothing or $1,000,000 depending on a past prediction about which choices you will make. At t_1, the experimenters tell you they will reveal whether the money is in the opaque box unless you pay them $1. At t_2, you'll get to decide whether to one-box or two-box. The predictor is highly reliable both at determining whether you will pay not to know what's in the opaque box

10. As Ahmed (2021) notes, there are multiple senses in which a decision theory could be deemed ostrich-like. Some are senses in which all standard decision theories are ostrich-like and some are senses in which even EDT is not ostrich-like. Ahmed favors a definition according to which an ostrich-like decision theory would recommend manipulating one's beliefs directly (such as by taking a pill to make you think that everything is fantastic). We agree with Ahmed that such direct manipulations are foolish and that EDT does not recommend them. We also don't want to get into a debate about what the definition of ostrich-like is. But our core point is that EDT's verdict in Torture shows that EDT is flawed and that this flaw is due to news mismanagement.

11. Note that a ratifiability requirement would plausibly alter EDT's verdict in this case. We're skeptical of ratifiability requirements for standard reasons (particularly that they sometimes forbid all actions, see Egan (2007) for more). And in any case, our intended topic is classic EDT.

and whether you'll one-box or two-box at t_2. If it was predicted that you'd ultimately take only the contents of the opaque box, then the opaque box contains $1,000,000. Otherwise, it contains nothing. What should you do?

Suppose you know at t_1 that you'll follow EDT at both t_1 and t_2. Then you know that if at t_2, you are certain there's nothing in the opaque box, you'll two-box. And you know that if you're certain there's a million in the opaque box, you'll also two-box. So, given that you know the contents of the box, you'll two-box no matter what at t_2. However, the predictor is very reliable, so at t_1, you think that if you decide not to pay the experimenters, it's highly likely you'll learn there's nothing in the opaque box. On the other hand, if you aren't certain what's in the opaque box at t_2, EDT will recommend one-boxing. In that case, you're very likely to find $1,000,000 in the opaque box. So, EDT tells you to pay not to know at t_1.

This case *may* be troublesome for EDT, but we don't think it's as troublesome as TORTURE. In OPTIONAL NEWCOMB, you pay not to know at t_1 to stop yourself from choosing an act at a different time that you now foresee as sub-optimal. If you had your druthers at t_1, you'd avoid paying and commit your t_2-self to one-boxing no matter what. But you don't have that option. Instead, it's worth a small fee to avoid letting your later self decide differently from how you'd like.[12] In TORTURE, no future decisions ride on whether John pays to rig the device. He buys himself nothing. All that's avoided is bad news.

A further virtue of this case—although inessential for the main point of news-management—is that it does not involve any strange prediction, as in NEWCOMB. John does in some sense predict himself, but it's the sort of prediction that is entirely mundane: he knows that he would behave in a particular way in a given situation. While we here assume he knows this with certainty for simplicity, the case also works if one relaxes this assumption.[13,14]

4. Reexamination

Our reasoning that John would think himself less likely to be tortured conditional on paying the $1 than conditional on not paying the $1 is plausible, but not beyond criticism. John's situation involves possible memory loss and attendant self-locating uncertainty, just as Adam Elga's (2000) Sleeping Beauty Problem does. And our TORTURE case is subject to some of the same same controversies as the Sleeping Beauty Problem. While it is uncontroversial that John's credence that he will be tortured should be a mixture of his credence that he will be tortured conditional on it being round 1 and his credence that he will be tortured conditional on it being round 2, it is controversial what his credences in it being round 1 or round 2 should be. (That's why our argument did not employ any particular probabilities for those possibilities, but only assumed intermediate credences for each.) Moreover, even our natural-seeming claim that—conditional on not paying—John should have credence .9 that he will be tortured is not beyond doubt. Some advocate what Titelbaum (2008) terms the "Relevance Limiting Thesis", according to which credences about uncentered propositions should only be affected by uncentered evidence. Given that thesis, seeing a red light would rule out the sequence *GG*, but would not favor *RR* over either *RG* or *GR*, and as a result John's credence in torture

12. See Arntzenius (2008) and Ahmed and Price (2012) for discussion of OPTIONAL NEWCOMB-like cases.

13. This case involves the possibility of memory loss, which some consider to be a rational failing. We don't share this view, but those who do may consider a variant of the case in which John has a twin, and both twins are sure that they will make the same choices. In this variant, the relevant issues are reproduced without the possibility of memory loss.

14. Soares and Fallenstein (2015) present a case called XOR BLACKMAIL where EDT comes apart from both FDT and CDT. We believe that XOR BLACKMAIL also supports the accusation that EDT is ostrich-like, but that TORTURE supports it even more strongly. The most important advantages for TORTURE are that it doesn't involve an exotic predictor but only appeals to self-prediction, and it shows that an EDT agent will *directly manipulate* a signal in order to receive auspicious news. See also Conitzer (2015) for another case that, like ours, involves *de se* credences.

would be less than .9.[15]

The Relevance Limiting Thesis does not merely muddy the waters; it invalidates our reasoning. Given the Relevance Limiting Thesis, John's credence that he will avoid torture conditional on paying the $1 is no greater than his credence that he will avoid torture conditional on his not paying.

4.1 Calculation

Let's look at the details of why the Relevance Limiting Thesis invalidates our reasoning. In our case, we have a sequence of states of the world: s_0 is either H or T, s_1 and s_2 are either red lights or green lights. We will index R and G accordingly, so HR_1G_2 is the world where the coin lands Heads, a red light blinks first, and a green light blinks second.

The agent has uncertainty both over which world is actual and over which center he occupies. So, we'll write $\Pr(s \text{ in } s_0s_1s_2 \mid E)$ for his subjective probability of the world being $s_0s_1s_2$ and him currently occupying center s given E. For instance, $\Pr(R_1 \text{ in } TR_1R_2)$ is his probability that he's seeing the red light flash for the first time in the world where the coin lands tails and the light flashes red both times.

According to the Relevance Limiting Thesis, upon seeing red, the agent only rules out the G_1G_2-worlds. It provides him with no *further* evidence that he is in an R_1R_2-world relative to an R_1G_2- or G_1R_2-world. Put differently: the agent takes seeing red *now* to be equivalent to learning the uncentered proposition that he sees red *at least once*, that is, the set of worlds with some red flashes.

One way to make this concrete is to appeal to the most common form of the Relevance Limiting Thesis, known as Compartmentalized Conditionalization (CC).

According to CC, $\Pr(s \text{ in } s_0s_1s_2 \mid E)$ should be equal to $\Pr(s_0s_1s_2 \mid E) \cdot 1/\#(E, s_0s_1s_2)$, where $\#(E, s_0s_1s_2)$ is the number of times the agent has total evidence E in the world $s_0s_1s_2$. For instance, if 'red' refers to the evidence the agent has when he has observed a red light, $\#(\text{red}, HR_1R_2) = 2$.

To see that paying is sub-optimal, we need only calculate John's subjective probabilities for being tortured (equivalently, for the coin landing tails) conditional on paying or not paying given that he observes red.

First, consider the policy of not paying, which we abbreviate $\bar{p}$. John's subjective probability here is:

$$\begin{aligned}\Pr(T \mid \bar{p}, \text{red}) &= \Pr(R_1 \text{ in } TR_1R_2 \mid \bar{p}, \text{red}) + \Pr(R_2 \text{ in } TR_1R_2 \mid \bar{p}, \text{red}) \qquad (1)\\ &\quad + \Pr(R_1 \text{ in } TR_1G_2 \mid \bar{p}, \text{red}) + \Pr(R_2 \text{ in } TG_1R_2 \mid \bar{p}, \text{red})\\ &= \Pr(TR_1R_2 \mid \bar{p}, \text{red}) + \Pr(TR_1G_2 \mid \bar{p}, \text{red}) + \Pr(TG_1R_2 \mid \bar{p}, \text{red}) \qquad (2)\end{aligned}$$

The second line follows given the Relevance Limiting Thesis in general (and from CC in particular). Note that $\Pr(TR_1R_2 \mid \bar{p}, \text{red}) = \Pr(TR_1R_2, \text{red} \mid \bar{p}) \cdot \frac{1}{Pr(\text{red} \mid \bar{p})}$ and similarly for the other terms in (2).

Furthermore, we can verify $\Pr(\text{red} \mid \bar{p}) = \Pr(\text{red} \mid p)$, so $\Pr(\text{red} \mid \bar{p}) = \Pr(\text{red})$. To see why, note:

$$\begin{aligned}\Pr(\text{red} \mid \bar{p}) &= 1 - (\Pr(HG_1G_2 \mid \bar{p}) + \Pr(TG_1G_2 \mid \bar{p}))\\ &= 1 - \left(\frac{1}{2} \cdot .9^2 + \frac{1}{2} \cdot .1^2\right)\\ &= 1 - (\Pr(HG_1G_2 \mid p) + \Pr(TG_1G_2 \mid p)\\ &= \Pr(\text{red} \mid p)\\ &= .59\end{aligned}$$

15. One could revise the procedure, replacing the single Green light flash with a sequence of a Green light, a Green light, and a Red light and replacing the single Red light flash with a sequence of a Green light, a Red light, and a Red light. It's natural to think that seeing a Green light is evidence that the coin landed Heads and that seeing a Red light is evidence that the coin landed Tails. But since it's certain that John will see at least one Green light and at least one Red light, according to the Relevance Limiting Thesis the flashes give him no evidence at all. This peculiar consequence is often taken as an argument against the Relevance Limiting Thesis. See also Weintraub (2004), Bostrom (2002), Titelbaum (2008), Briggs (2010), and Dorr (ms) for more.

Putting this all together, we have:

$$\begin{aligned}\Pr(T \mid \text{red}, \bar{p}) &= \frac{1}{Pr(\text{red})}\left[\Pr(TR_1R_2 \mid \bar{p}) + \Pr(TR_1G_2 \mid \bar{p}) + \Pr(TG_1R_2 \mid \bar{p})\right] \\ &= \frac{1}{\Pr(\text{red})}\left(\frac{1}{2}\cdot .9^2 + \frac{1}{2}\cdot(.09) + \frac{1}{2}\cdot(.09)\right) \\ &= \frac{.495}{\Pr(\text{red})}\end{aligned}$$

To calculate the conditional probability of torture given John pays upon seeing red, we use the same derivation to see that:

$$\begin{aligned}\Pr(T \mid \text{red}, p) &= \frac{1}{\Pr(\text{red})}\left[P(TR_1R_2 \mid p) + \Pr(TR_1G_2 \mid p) + \Pr(TG_1R_2 \mid p)\right] \\ &= \frac{1}{\Pr(\text{red})}\left(\frac{1}{2}\cdot .9 + 0 + \frac{1}{2}\cdot .09\right) \\ &= \frac{.495}{\Pr(\text{red})}\end{aligned}$$

So, if John follows both the Relevance Limiting Thesis and EDT, he won't pay.

4.2 A Variant

Our analysis of TORTURE only holds if the Relevance Limiting Thesis is false. And, admittedly, the general consensus is that the Relevance Limiting Thesis is false—most of the controversy regarding the epistemology of self-locating belief concerns *how* self-locating evidence affects credences in uncentered propositions, not *whether* it does. So it would not be the end of the world if our argument had to assume that the Relevance Limiting Thesis was false. But we don't. Happily, it's possible to modify TORTURE slightly so that the Relevance Limiting Thesis loses its relevance.

The Relevance Limiting Thesis matters for our initial statement of TORTURE only because of the possibility of duplicate experiences. But it's easy to adapt Michael Titelbaum's (2008) "technicolor" trick and thereby avoid that possibility. Let's suppose that there's another fair coin that's tossed, and it will affect the brightness of the red / green lights that John is shown. If this coin lands Heads then the light he sees at t_1 will be bright and the light he sees at t_2 will be dim, and if the coin lands Tails then the light he sees at t_1 will be dim and the light he sees at t_2 will be bright. Since the brightness of the light is guaranteed to vary across times, even cases in which John sees two red lights or two green lights will not contain duplicate experiences, and thus the Relevance Limiting Thesis will not apply. Whatever sort of light John sees, he can rule out worlds in which he never sees that sort of light and renormalize his credences in the worlds in which he does see that sort of light.

Most problems in the epistemology of self-locating belief are not so easily avoided. As we mentioned, the main controversies involve how self-locating evidence affects credences in uncentered propositions. And the crux of the controversies is how confirmation works between worlds that contain different numbers of agents (or different quantities of experience for some agent).[16] But TORTURE involves the same quantity of experiences for John no matter what. Thus although John is uncertain whether he's at t_1 or t_2, this self-locating uncertainty is entirely pedestrian—like not being sure exactly what time it is under ordinary circumstances. All major views regarding the epistemology of self-locating belief will validate the following calculations.[17]

4.3 The Details

To see why the technicolor trick works, we'll assume without loss of generality that John sees a dim red light, which we abbreviate dr.

We'll write the results of the first coin toss (which determines whether John gets tortured) as either H_1 or T_1 and the second coin toss as H_2 or T_2 and use upper and lower case letters to denote bright or dim lights, respectively. So $H_1H_2R_1g_2$ denotes the fact that both coins landed heads, the first light was bright red, and the second light was dim green.

16. In the framework of time-slice epistemology these amount to the same thing. See Hedden (2015) for more.
17. For discussions of how to update in the face of centered evidence, see Bostrom (2002) and Titelbaum (2012). For a proof that the major theories of self-locating belief all agree in pedestrian circumstances, see Isaacs et al. (2022).

Suppose John will pay upon seeing a red light (dim or not). Then:

$$\begin{aligned}\Pr(T_1 \mid \mathrm{dr}, p) &= \Pr(T_1H_2R_1r_2 \mid \mathrm{dr}, p) + \Pr(T_1H_2G_1r_2 \mid \mathrm{dr}, p) \\ &\quad + \Pr(T_1T_2r_1R_2 \mid \mathrm{dr}, p) + \Pr(T_1H_2r_1G_2 \mid \mathrm{dr}, p) \\ &= \frac{1}{\Pr(\mathrm{dr} \mid p)}\left[\Pr(T_1H_2R_1r_2 \mid p) + \Pr(T_1H_2G_1r_2 \mid p)\right. \\ &\quad \left. + \Pr(T_1H_2r_1R_2 \mid p) + \Pr(T_1H_2r_1G_2 \mid p)\right] \\ &= \frac{1}{\Pr(\mathrm{dr} \mid p)}\left(\frac{1}{4}\cdot.9 + \frac{1}{4}\cdot.09 + \frac{1}{4}\cdot.9 + 0\right) \\ &= \frac{1}{\Pr(\mathrm{dr} \mid p)}\cdot.4725\end{aligned}$$

The second equality follows by the definition of conditional probability and the fact that observing a dim red light is guaranteed in each of the worlds considered.

Next we calculate:

$$\begin{aligned}\Pr(\mathrm{dr} \mid p) &= \Pr(H_1H_2R_1r_2 \mid p) + \Pr(H_1H_2G_1r_2 \mid p) + \Pr(H_1T_2r_1R_2 \mid p) \\ &\quad + \Pr(T_1H_2R_1r_2 \mid p) + \Pr(T_1H_2G_1r_2 \mid p) + \Pr(T_1T_2r_1R_2 \mid p) \\ &= \frac{1}{4}\left[.1 + .09 + .1 + .9 + .09 + .9\right] \\ &= .545\end{aligned}$$

So,

$$\begin{aligned}\Pr(T_1 \mid p) &= .4725/.545 \\ &\approx .867\end{aligned}$$

On the other hand, if John doesn't pay, a similar calculation reveals that:

$$\begin{aligned}\Pr(T_1 \mid \mathrm{dr}, \bar{p}) &= \frac{1}{\Pr(\mathrm{dr} \mid \bar{p})}\left[\Pr(T_1H_2R_1r_2 \mid \bar{p}) + \Pr(T_1H_2G_1r_2 \mid \bar{p})\right. \\ &\quad \left. + \Pr(T_1H_2r_1R_2 \mid \bar{p}) + \Pr(T_1H_2r_1G_2 \mid \bar{p})\right] \\ &= \frac{1}{\Pr(\mathrm{dr} \mid \bar{p})}\cdot.45\end{aligned}$$

A tedious calculation shows that $P(\mathrm{dr} \mid \bar{p}) = .5$. So:

$$\begin{aligned}\Pr(T_1 \mid \bar{p}) &= \frac{.45}{.5} \\ &= .9\end{aligned}$$

Given that the cost of payment is trivial, John prefers paying to not paying if he follows EDT.

(The astute reader may notice that John has more possible options now, such as paying if the red light is dim but not when it's bright or vice versa. We won't go through the calculations here, but John will prefer paying upon seeing *any* red light to these more complicated options.)

5. Why Ain'cha Rich?

A traditional motivation for EDT is that its followers tend to do better than followers of CDT. In NEWCOMB, for instance, one-boxers tend to end up richer than two-boxers. So one-boxers can challenge two-boxers by saying, "If you're so smart, why ain'cha rich?"[18]

The causal decision theorist can of course retort that the evidentialist is looking at the wrong reference classes. Of people who walk into a room with only a thousand dollars, causalists do better. And of people who walk into a room with a million and a thousand dollars, causalists also do better. From the CDT point of view, the fact that evidentialists tend to walk into better rooms is irrelevant.

In this case, though, there's no good sense in which EDT outperforms CDT.

18. See Lewis (1981).

Evidentialists get tortured just as often as causalists. People who choose to pay get tortured just as often as people who choose not to pay.[19] What's different is the ratio of instances of torture to red-seeing time-slices—evidentialists have fewer instances of torture per red-seeing time-slice than causalists do. But that's merely because evidentialists stupidly produce extra red-seeing time-slices; they make themselves get bad news more often so as to dilute the significance of the bad news. This plainly is an irrational manipulation of the news.

In Newcomb, "why ain'cha rich" reasoning militates in favor of EDT's verdict. But in Torture, "why ain'cha rich" reasoning militates against EDT's verdict. So EDT is not supported by "why ain'cha rich" reasoning. In fact, that reasoning cuts against EDT.[20]

Consider an analogous situation: You suffer from infrequent but very painful migraine headaches. There's a biotech company that can predict when you'll get migraines, and it notifies you about upcoming migraines the day before they happen. But if you pay them extra, they'll also randomly tell you that you're going to get a migraine even when you won't. That way the news value of being told that you're going to get a migraine won't be as bad. It's obviously irrational to pay to get bad news more often in order to make each instance of bad news less bad. By paying you don't get to have any fewer headaches, so it's not worth anything. And indeed, EDT would not recommend paying extra; the risk of migraines is the same either way. However, it's good news to learn that in the past you *had* paid to make it more likely that you'd get fallacious notifications of future migraines. In effect, EDT recommends paying in the present so as to get evidence that you paid in the past. This is obviously foolish. It's good news not worth paying for.

6. The Larger Dialectic

It is a common view that the correct decision theory mandates the maximization of expected utility.[21] Yet there are deep disagreements about how expected utilities should be calculated—in effect, about what expected utilities are. EDT and CDT are the most prominent positions (though there are others). The standard methodology is to come up with cases where these decision theories disagree and pump intuitions about which verdict is right. But intuitions differ, and any verdict is liable to be justifiable in some fairly natural sense.[22] EDT will maximize evidential expected utility and fail to maximize causal expected utility, while CDT will maximize causal expected utility and fail to maximize evidential expected utility. Any sensible decision theory will be optimal relative to its own sense of optimality. So the strongest argument against a sensible decision theory is one that makes its sense of optimality seem foolish.

Several such arguments have been attempted regarding EDT. Newcomb was meant to show that followers of EDT foolishly reject free money. But followers of EDT (unlike those who diverge from EDT in Newcomb) tend to wind up rich. That doesn't seem straightforwardly foolish.

Arntzenius (2008) offers a case in which, given a predictor who predicts whether an agent will win or lose their bets, a follower of EDT will tend to lose money in the long-run. But it's both odd to have predictions about whether or not bets will win and it's odd that the argument only applies to long-run tendencies.[23]

Wells (2019) offers a complicated case involving multiple decisions, predictions, and coin tosses in which a follower of EDT is guaranteed to end up poorer than a follower of CDT. But Wells' case crucially relies on the follower

19. To see why, note that this is essentially the same question as the proponent of Compartmentalized Conditionalization asks: those who see red at least once and pay are tortured just as often as those who see red at least once and don't pay.
20. Ahmed and Price (2012) unpack "why ain'cha rich" reasoning, deploying it to support EDT. But one can formulate an argument parallel to theirs which opposes EDT:

    (1) The average return of being a non-payer exceeds that of being a payer.
    (2) Everyone can see that (1) is true.
    (3) Therefore not paying foreseeably does better than paying.
    (4) Therefore EDT is committed to the foreseeably worse option for anyone facing Torture.

    Torture shows that—by the very lights of EDT's defenders—EDT is flawed.

21. For recent alternatives to expected utility theory, see Buchak (2013) and Rinard (2015).
22. For more on this point, see Horgan (2017) and Bales (2018).
23. See Ahmed and Price (2012) for an extended critique of Arntzenius' argument on these two points.

of EDT and the follower of CDT having different credences (about what they expect to do), and thus the agents Wells compares do not actually face the same decision problem.[24]

A further advantage of TORTURE is that it is straightforwardly unaffected by the tickle defense. Ellery Eells (1981, 1982) argues that EDT doesn't actually recommend one-boxing in NEWCOMB, and thus that Lewis' accusation against EDT on the basis of that recommendation is misguided. Eells contends that both the predictor's prediction and the agent's action are based on the agent's beliefs and desires, and further that the agent can feel the pull of these beliefs and desires—the tickle—prior to action. Detecting the character of one's tickle will screen off the correlation between prediction and action, thus removing any incentive to one-box. It's unclear whether the tickle defense works in NEWCOMB or in the Arntzenius and Wells cases. But in TORTURE, it is obvious that what matters is what the agent actually chooses, and not any sort of doxastic or bouletic tickle. There's no way to screen off the relevant correlation, and thus no way to claim that EDT avoids making a foolish recommendation.

The most prominent problem cases for EDT do not make it clear that EDT has a problem. The case presented in this paper is simpler, more straightforward, and does show what's wrong with EDT. Lewis' famous charge that EDT irrationally manages the news is vindicated.[25]

References

Ahmed, A. (2014). *Evidence, Decision and Causality*. Cambridge University Press.

Ahmed, A. (2020). Equal opportunities in Newcomb's problem and elsewhere. *Mind 129*(515), 867–886.

Ahmed, A. (2021). *Evidential Decision Theory*. Cambridge University Press.

Ahmed, A. and H. Price (2012). Arntzenius on 'why ain'cha rich?'. *Erkenntnis 77*(1), 15–30.

Arntzenius, F. (2008). No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis 68*, 277–297.

Bales, A. (2018). Decision-theoretic pluralism. *Philosophical Quarterly 68*(273), 801–818.

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge.

Briggs, R. (2010). Putting a value on beauty. In T. S. Gendler and J. Hawthorne (Eds.), *Oxford Studies in Epistemology, Volume 3*, pp. 3–34. Oxford University Press.

Buchak, L. (2013). *Risk and Rationality*. Oxford University Press.

Conitzer, V. (2015). A dutch book against sleeping beauties who are evidential decision theorists. *Synthese 192*(9), 2887–2899.

Dorr, C. (ms). A challenge for halfers.

Eells, E. (1981). Causality, utility, and decision. *Synthese 48*(2), 295–329.

Eells, E. (1982). *Rational Decision and Causality*. Cambridge University Press.

Egan, A. (2007). Some counterexamples to causal decision theory. *Philosophical Review 116*(1), 93–114.

Elga, A. (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis 60*(2), 143–147.

Gallow, J. D. (2020). The causal decision theorist's guide to managing the news. *The Journal of Philosophy 117*(3), 117–149.

Hájek, A. (2016). Deliberation welcomes prediction. *Episteme 13*(4), 507–528.

Hedden, B. (2015). Time-slice rationality. *Mind 124*(494), 449–491.

Hedden, B. (2023). Counterfactual decision theory. *Mind 132*, 730–761.

Horgan, T. (1981). Counterfactuals and newcomb's problem. *The Journal of Philosophy 78*(6), 331–356.

Horgan, T. (2017). *Essays on Paradoxes*. Oxford, England: Oxford University Press USA.

Horwich, P. (1987). *Asymmetries in Time: Problems in the Philosophy of Sciences*. Series Bradford Books.

Isaacs, Y., J. Hawthorne, and J. Sanford Russell (2022). Multiple universes and self-locating evidence. *Philosophical Review 131*(3), 241–294.

24. For more on this point see Ahmed (2020).

25. Thanks to John Hawthorne, Vince Conitzer, and an audience at the Formal Rationality Forum at Northeastern University. Special thanks to Caspar Oesterheld who provided insightful comments and pushed us on the Relevance Limiting Thesis. Ben Levinstein's research was partly supported by Mellon New Directions grant 1905-06835.

Jeffrey, R. C. (1983). *The Logic of Decision* (2nd ed.). University of Chicago Press.

Levinstein, B. A. and N. Soares (2020). Cheating death in damascus. *The Journal of Philosophy 117*(5), 237–266.

Lewis, D. K. (1981). Why ain'cha rich? *Noûs 15*(3), 377–380.

Meacham, C. J. (2010). Binding and its consequences. *Philosophical Studies 149*(1), 49–71.

Rinard, S. (2015, February). A decision theory for imprecise probabilities. *Philosohers' Imprint 15*(7), 1–16.

Soares, N. and B. Fallenstein (2015). Toward idealized decision theory. *arXiv Pre-print arXiv: 1507.01986*.

Spohn, W. (2012). Reversing 30 years of discussion: Why causal decision theorists should one-box. *Synthese 187*(1), 95–122.

Titelbaum, M. (2008). The relevance of self locating beliefs. *Philosophical Review 117*(4), 555–606.

Titelbaum, M. G. (2012). *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. Oxford University Press.

Weintraub, R. (2004). Sleeping beauty: A simple solution. *Analysis 64*(1), 8–10.

Wells, I. (2019). Equal opportunity and Newcomb's problem. *Mind 128*(510), 429–457.

Yudkowsky, E. and N. Soares (2017). Functional decision theory: A new theory of instrumental rationality.