Risk-averse Causalists aren't so easily frustrated

Toby C.P. Solomon

January 1, 2025

Abstract

Jack Spencer and Ian Wells have recently argued that Causal Decision Theory faces special difficulty in cases of decision-instability where a play-it-safe option is present. They argue that CDT recommends taking a risky option, while the rational thing to do is to play it safe. In this paper I will show that CDT only recommends the risky option if we assume risk neutrality—a risk-averse CDT can play it safe. This opens two lines of response to Causalists: They can embrace a risk-averse CDT. Or they can reject the intuition to play it safe on the general grounds that risk-aversion is irrational. I will also generalise this argument to several other examples involve decision-instability. Of course, risk-aversion cannot explain all CDT's problems and I will bolster the case for risk-aversion playing a special role in these cases by showing it cannot help in all such cases.

Keywords: Causal Decision Theory; Risk-Aversion; Decision-Instability; Frustrater Paradox

1 Introduction

Counterexamples to Causal Decision Theory (CDT) based on decision-instability are now commonplace.[1] Decision-instability occurs when deciding on option A produces a change in credence which makes a different option B look more attractive and vice versa. Jack Spencer and Ian Wells (Spencer and Wells 2019; Spencer 2021) have recently presented a new example of this kind—The Frustrater—which seems to make things worse for CDT by offering the decision-maker an attractive third, play-it-safe, option C. Fortunately for defenders of Causal Decision Theory (Causalists), we can explain the intuition in favour of this third option as an instance of risk-aversion. And counterexamples involving risk-aversion are even older than Causal Decision Theory itself.[2] Causalists can, therefore, simply transfer their favoured approach to risk-aversion to The Frustrater. If they are happy to embrace risk-aversion, they can appeal to a version of CDT that allows for risk-aversion. Or, if they take risk-aversion to be irrational, they can debunk Spencer and Wells' intuitions with whatever story they apply to debunk risk-averse intuitions in existing cases. The Frustrater is, hence, no new threat to CDT.

In the main part of this paper I argue for all this. I will also argue that paying attention to risk-aversion might reduce the sting for Causalists of some other cases of decision-instability, using Egan's (2007) Psychopath Button as a test case. I will also consider an example due to Arif Ahmed (2014), which is superficially similar to The Frustrater and which shows that risk-aversion cannot be the entire story with these kinds of cases; I take this to be a virtue and not a vice since we should be suspicious of theories which overgeneralise. Finally, I will consider Spencer's (2021) diachronic elaboration of The Frustrater, Two Doors, and show it poses no extra problem for the risk-averse approach. Just how Causalists should respond to pure decision-instability is an open problem, but along the

[1] An early example—Death in Damascus—being considered in (Gibbard and Harper 1981). The recent literature following mostly from (Egan 2007).

[2] Most notably from the Allais (1953) and Ellsberg (1961) paradoxes.

way I will consider how this appeal to risk-aversion interacts with the promising *dynamic deliberation* approach.

Throughout this paper I assume that you are familiar with Causal Decision Theory (see, for example, (Lewis 1981)). Before getting to the main discussion I will briefly introduce the problems of decision-instability and risk-aversion. I will also introduce, respectively, the dynamic deliberation approach to decision-instability (as championed by Arntzenius (2008), Joyce (2012), and Armendt (2019)) and Buchak's (2016) Risk-Weighted Expected Utility Theory[3]. You can safely skip the next two sections and proceed directly to the main discussion in section 4 if you are also familiar with these.

2 Instability and Dynamic Deliberation

The *locus classicus* of decision-instability is Death in Damascus, which goes something like this (following (Gibbard and Harper 1981, 186)):

> Death has an appointment book in which is written the name of each person and the time and place of their death. Death is always at the indicated place at the indicated time—the named person dies if and only if they are also there at that time. Death's appointment book is never wrong; though his appointments are written in it many weeks before they occur. Say that you are quite confident of all this. Then, one day, you bump into Death while shopping in the market at Damascus. Death is quite surprised and says "I didn't expect to see you here, I have an appointment with you tomorrow". Knowing the old story about the man who immediately fled to Aleppo, this gives you pause: you know that you should try to be in whichever city Death does not have written in his appointment book. But because Death's appointment book is never wrong, you

[3]Which is explicitly intended as a version of Causal Decision Theory (Buchak 2016, 88).

> are quite sure that whichever city you choose to be in will be the city that death has written in it.

Now, let us suppose, you have only two options: you can remain in Damascus or you can flee to Aleppo. And there are two relevant states of the world: Death's appointment book has Damascus written in it or it has Aleppo written in it. Since your actions now will not change what is written in Death's appointment book—it is a very accurate appointment book, but it is not magic—these states are causally independent of your choice. CDT then advises you to calculate expected utility using your *unconditional* credence that Death will be in Damascus and that he will be in Aleppo. Assuming that your utility for living is 100 and for dying is 0, CDT will tell us that the relevant expected utilities are simply:

$$EU_{CDT}(\text{Stay in Damascus}) = 100 \times Cr(\text{Death will be in Aleppo})$$
$$EU_{CDT}(\text{Flee to Aleppo}) = 100 \times Cr(\text{Death will be in Damascus})$$

Hence, CDT will advise you to stay in Damascus if your credence that Death will be in Aleppo is less than 0.5 (and vice versa). Assume that you are initially confident that Death will be in Aleppo given his surprise about seeing you. The *problem* is that once you decide (or as you incline towards deciding) to stay in Damascus you gain new evidence—that you will likely be in Damascus tomorrow—which suggests that Death will be in Damascus. But as your credence in Death being in Damascus goes up, fleeing to Aleppo will look more and more attractive. Hence, deciding to stay in Damascus will make fleeing to Aleppo look like the better option. And deciding to flee to Aleppo will make staying in Damascus look like the better option. Whatever you decide, you will regret your decision.

Evidential Decision Theory does not face this problem because it asks you to calculate expected utility on the basis of credence in each state conditional on performing each option. These conditional credences automatically take into account the information about

Death's probable location tomorrow that deciding either way would give you. Assuming you take Death's book to be infallible, this will lead you to be indifferent between the options since you will be certain that wherever you go Death will be there and so both options have an expected utility of 0. You will regret either choice equally, but you know that if you changed your choice you would immediately regret doing that as well, so Evidential Decision Theory's advice to be indifferent between your (bad) options seems sensible.

There has, of course, been extensive discussion on how CDT can deal with such cases. The purpose of this paper is not to rehearse that debate but to show that Spencer and Wells' Frustrater offers no new problem above and beyond the existing ones of decision-instability and risk-aversion. I will, however, very briefly introduce my favoured solution to decision-instability—the dynamic deliberation approach—so that we can see how it interacts with risk-aversion later. Here I will only introduce the idea in informal and intuitive terms, since the full formal details would take us too far afield.[4]

The core of the dynamic deliberation approach is the claim that a rational decision-maker will not settle on a course of action until they have taken into account all the freely available information that is relevant to their decision. And in cases of decision-instability this includes the information they receive about how the world is (where Death will be, for example) by paying attention to which decision theory are inclining towards (which currently has highest expected utility).[5] Rational decision-makers will feed this information back into their deliberation until they reach a point where doing so will not make any further difference to their choice.[6] That is, they will deliberate in a way modelled by the following iterative process:

1. Calculate the expected utility of all options using current credences.

[4]See any of (Arntzenius 2008), (Joyce 2012), or (Armendt 2019) for relatively short introductions. See (Skyrms 1990) for the most comprehensive treatment.

[5]As this makes clear, the dynamic deliberation approach *requires* assigning credences to your own actions—but Joyce has elegantly defended this at length in (Joyce 2002, 2016).

[6]On the assumption that nothing external will force a choice to be made and that continuing deliberation is cost free. For more on these complications see (Skyrms 1990).

2. Raise your credence in the option which currently has highest expected utility (and correspondingly lower credence in other options).

3. Update, via conditional probabilities, your credence in the relevant states of the world to reflect step 2.

4. Repeat 1–3 until no further change in credence occurs—that is, until an equilibrium is reached.

Such an equilibrium point can only occur when the expected utilities of all live options—i.e. those which the decision-maker is not certain they will not perform—are equal. If not, the decision-maker will become more certain that they will perform the action with current highest expected utility when step 1 is repeated and hence we cannot be at an equilibrium.

In Death in Damascus, assuming you are initially uncertain where Death will be, the only such equilibrium[7] is when you have 0.5 credence that you will stay in Damascus (and the same for fleeing to Aleppo) and hence 0.5 credence that Death will be in Aleppo (and the same for Damascus). With these credences you will have $EU_{CDT} = 50$ for both options and be indifferent between them. And this, proponents of the dynamic deliberation approach claim, is the right result: whatever you decide to do you will regret it, and you will regret it equally, so you should simply be indifferent across all the (bad) options you have. (And this is, of course, the same advice that Evidential Decision Theory gives in the case.) Things will be more slightly more complicated in cases of *asymmetric* decision-instability—I will return to the classic Psychopath Button as an example of this below.

Now we have some idea of how decision-instability affects CDT and one potential solution. Decision-instability is one ingredient of The Frustrater, the other ingredient, we will see, is risk-aversion. In the next section I introduce risk-aversion. In the following section I introduce The Frustrater and show it generates no new problem for Causalists.

[7]Given some relatively light assumptions; most notably that the dynamic process is continuous and that you take Death's book to be perfectly accurate or to have equal chance of error either way.

	1	2–11	12–100
L_1	0	2	1
L_2	1	1	1

Table 1: First pair of Allais options.

3 Risk-Aversion

You would, I assume, refuse to pay €100,000 to make a bet on a (fair) coin flip that pays out €200,000 if heads comes up and nothing otherwise. I'm sure that you would like to have an extra €100,000, but you probably can't afford to accept a 50% risk of losing €100,000. You thus (I assume) display *risk-aversion* with respect to monetary bets. It is uncontroversial that this kind of risk-aversion is rationally permissible. CDT (or any other expected utility theory) can accommodate it simply by allowing utility functions which give monetary amounts diminishing marginal value: a loss of €100,000 has a utility of, say, -100,000, whereas a gain of €100,000 has a positive utility of only, say, 1,000.

What is more controversial is whether you can be risk-averse[8] with respect to utilities themselves. That is, whether it is rationally permissible to be risk-averse even once the diminishing marginal value of all physical goods, pleasures, and so on is taken into account. The classic example of preferences that require just such risk-aversion are the Allais Preferences—that is, the joint claims that option L_2 is preferable to option L_1 for the decision in Table 1, and option R_1 is preferable to option R_2 for the decision inTable 2.[9] In each choice you must choose which of two fair lotteries to enter (for free). You will be given a ticket with a number between 1 and 100 and you will receive a prize 1 or 2 units of utility or nothing, according to the number on your ticket.

[8]Or even risk seeking, but I will ignore that possibility throughout this paper.

[9]Diminishing marginal utility for money allows us to make either L_1 and R_1 rational or L_2 and R_2, but not L_1 and R_2 or L_2 and R_1.

	1	2–11	12–100
R_1	0	2	0
R_2	1	1	0

Table 2: Second pair of Allais options.

Standard, that is risk-neutral, CDT assigns the following expected utilities:

$$EU_{CDT}(L_1) = 0 \times 0.01 + 2 \times 0.1 + 1 \times 0.89 = 1.09$$

$$EU_{CDT}(L_2) = 1 \times 0.01 + 1 \times 0.1 + 1 \times 0.89 = 1$$

$$EU_{CDT}(R_1) = 0 \times 0.01 + 2 \times 0.1 + 0 \times 0.89 = 0.2$$

$$EU_{CDT}(R_2) = 1 \times 0.01 + 1 \times 0.1 + 0 \times 0.89 = 0.11$$

Hence, standard CDT recommends that you take both L_1 and R_1—in both choices only what happens for tickets 1–11 is different between the options, and for those tickets L_1 and R_1 give a much better chance of getting 2 for only a small increase in risk of getting 0. Many, however, feel that it is reasonable to reverse these preferences in the first choice on the basis that L_2 is a sure thing—it guarantees a gain of 1, where L_1 carries a risk of ending up with nothing. While in the second choice you are risking ending up with 0 either way, so you may as well take the significantly increased chance of getting 2 rather than 1. That is, they are risk-averse.

Now, I will here remain neutral on whether such general risk-aversion is in fact rational (below I show that CDT has options either way). However, Lara Buchak has recently made a major advance in this debate by showing how to build a general decision theory which is compatible with a causal approach and can capture such risk-averse preferences. Buchak's Risk-Weighted Expected Utility works like this:

- First, for each option in a decision we order its possible outcomes—i.e. act-state

pairs—from worst to best. That is, we construct an ordered set for each option $\langle O_1, O_2, ... \rangle$, where O_1 is the worst outcome that option might lead to, O_2 is the next to worst outcome, and so on.

- We calculate the expected utility of each option according to the following equation:

$$EU_{RCDT}(A) = \sum_{j=1}^{n} \left(\left(U(O_j) - U(O_{j-1}) \right) r\left(\sum_{i=j}^{n} p(O_i) \right) \right)$$

More or less simply: an option's expected utility is the utility of its worst outcome, plus the difference between the utility of the worst and next to worst outcomes weighted by the probability *transformed by your risk attitude* that you will get at least that much utility, and so on. If we did not add in a risk-weighting here we would get the same result as standard CDT. The risk-weighting allows us to reduce (with respect to CDT) the probability by which better outcomes are weighted to represent risk-aversion. Buchak's theory thus takes five inputs: A set of options, a set of states, a utility function, a credence function, and a risk function. To ensure a causal interpretation we simply require that the states are causally independent of the options (and none of the examples we will consider here troubles that assumption).

To get a feeling for how this works, let's apply it to the Allais preferences. We will assume you have the risk function $r(p) = p^2$. Notice first that for L_1 the ordered outcome set is $\langle 0, 1, 2 \rangle$ while for L_2 it is simply $\langle 1 \rangle$, while for R_1 it is $\langle 0, 2 \rangle$ and for R_2 it is $\langle 0, 1 \rangle$.

$$EU_{RCDT}(L_1) = 0 + (1 - 0) \times (0.1 + 0.89)^2 + (2 - 1) \times (0.1)^2 = 0.9901$$

$$EU_{RCDT}(L_2) = 1$$

$$EU_{RCDT}(R_1) = 0 + (2 - 0) \times (0.1)^2 = 0.02$$

$$EU_{RCDT}(R_2) = 0 + (1 - 0) \times (0.11)^2 = 0.0121$$

Hence, given this level of risk-aversion, you can rationally prefer L_2 to L_1—because L_2 is a safe thing—while preferring R_1 to R_2—because if you have to take a risk you may as well, so to speak, go for the higher expected pay off.

Of course, there is another way that Causalists might respond to the Allais preferences—namely, they might argue that they are irrational and offer some debunking story for our tendency towards risk-aversion. For example, we might appeal to the rationality of risk-aversion with respect to money and the difficulty of reasoning about abstract utilities (rather than about concrete things like money) to explain the appeal of the risk-averse preferences in the Allais choices without admitting their rationality. In the following I will simply seek to show that the intuitions that trouble CDT in The Frustrater can be explained as a further instance of risk-aversion; I will remain neutral on whether that is further evidence that we should embrace Risk-Weighted Expected Utility Theory or is simply another example of the same irrationality as the Allais preferences. It is time now to make good on my promise and show that The Frustrater poses no new problem for CDT but simply combines the old problems of decision-instability and risk-aversion.

4 The Frustrater

Here is Spencer and Wells' new example, The Frustrater:

> There is an envelope and two opaque boxes, A and B. The agent has three options: she can take A, B or the envelope (a_A, a_B or a_E). The envelope contains \$40. The two boxes together contain \$100. How the money is distributed between the boxes depends on a prediction made yesterday by the Frustrater, a reliable predictor who seeks to frustrate. If the Frustrater predicted that the agent would take A, then B contains \$100. If the Frustrater predicted that the agent would take B, then A contains \$100. If the Frustrater predicted that the

	p_A	p_B	p_E
a_A	0	100	50
a_B	100	0	50
a_E	40	40	40

Table 3: Decision matrix for The Frustrater.

> agent would take the envelope, each box contains \$50. The agent knows all of this (Spencer and Wells 2019, 34).

Spencer and Wells' maintain that the rational option is to take the envelope. They also maintain that CDT cannot secure this result. I will remain neutral on whether you should take the envelope. But I will argue that risk-averse CDT—in both static and dynamic deliberation forms—can secure this result. A decision matrix for this problem, using p_X to indicate predictions in the obvious way and assuming utility = dollars, is shown in Table 3.

First, consider what risk-neutral CDT would say about the case: Since the states are by assumption causally independent of your choice, a_E must have the minimum expected utility no matter what your credences in each prediction (state). Why? Well, assume that $\frac{p_A}{p_A+p_B} > 0.5$—that is, that you assign p_A more than half of the credence you assign to p_A and p_B jointly—then it is simple to see that $EU_{CDT}(a_B) > 50$, and so will be preferred to a_E whose expected utility is a fixed 40. And similarly, when $\frac{p_A}{p_A+p_B} < 0.5$ then $EU_{CDT}(a_A) > 50$. So, whatever your credences are, you will prefer one of a_A and a_B to a_E. So, at first glance CDT will recommend avoiding a_E.

Now, The Frustrater has an element of decision-instability about it: The more you incline to choosing each option the more likely it will seem that that option has been predicted. And the more likely it seems that an option has been predicted the less attractive it will seem—each option is worst on the assumption that it has been predicted. Can a dynamic CDT help us to capture Spencer and Wells' intuition? Unfortunately not. While inclining towards a_A will make a_B seem more attractive, and vice versa, inclining towards

a_E only makes either of a_A and a_B seem more attractive. A dynamic CDT then continuously lower your credence that you will take the envelope until it reaches 0. Hence, dynamic CDT alone cannot vindicate Spencer and Wells' intuition.

However, a risk-averse CDT *can* endorse a_E. The outcome orderings for the options are: $a_A = \langle 0, 50, 100 \rangle$, $a_B = \langle 0, 50, 100 \rangle$ and $a_E = \langle 40 \rangle$. And hence a risk-weighted CDT will give the following general equations for expected utility:

$$\begin{aligned} EU_{RCDT}(a_A) &= 0 + r(Cr(p_E) + Cr(p_B)) \times (50 - 0) + r(Cr(p_B)) \times (100 - 50) \\ &= (r(Cr(p_E) + Cr(p_B)) + r(Cr(p_B))) \times 50 \\ EU_{RCDT}(a_B) &= 0 + r(Cr(p_E) + Cr(p_A)) \times (50 - 0) + r(Cr(p_A)) \times (100 - 50) \\ &= (r(Cr(p_E) + Cr(p_A)) + r(Cr(p_A))) \times 50 \\ EU_{RCDT}(a_E) &= 40 \end{aligned}$$

Hence, $EU(a_E) > EU(a_A)$ and $EU(a_E) > EU(a_B)$ just so long as $r(Cr(p_E) + Cr(p_B)) + r(Cr(p_B)) < \frac{4}{5}$ and $r(Cr(p_E) + Cr(p_A)) + r(Cr(p_A)) < \frac{4}{5}$. Assume again that $r(p) = p^2$. Then these inequalities are satisfied if, for example, we have $Cr(p_E) = 0.5$, $Cr(P_A) = 0.25$, and $Cr(p_B) = 0.25$. And we will have:

$$\begin{aligned} EU_{RCDT}(a_A) &= ((Cr(p_E) + Cr(p_B))^2 + Cr(p_B)^2) \times 50 \\ &= 31.25 \\ EU_{RCDT}(a_B) &= ((Cr(p_E) + Cr(p_A))^2 + Cr(p_A)^2) \times 50 \\ &= 31.25 \\ EU_{RCDT}(a_E) &= 40 \end{aligned}$$

So a risk-weighted CDT can capture Spencer and Wells' intuition in favour of taking the envelope. What does this tell us? Well, I think it suggests (strongly) that Spencer and Wells'

intuition is really tracking risk-aversion, not any new problem for CDT.

Of course, even a risk-weighted CDT will not universally endorse taking the envelope: there are some risk functions for which a_A or a_B will still be preferable for all credences and similarly for all risk functions[10] there will be some credences for which taking a_A or a_B will still be preferable. In particular, if you are sure enough of p_E then a_A or a_B will always be preferable to a_E. Hence, risk-weighted CDT can capture the intuition that you may take the envelope, but not that taking the envelope is always uniquely rational. Is this a problem? No, at least not for Causalists. The only way to defend taking the envelope regardless of your risk attitude is to appeal to the fact that you should expect anyone who takes a_A or a_B to on average walk away with less than \$40. But doing so relies on reasoning on the basis the high conditional, evidential, probabilities $Cr(p_A|a_A)$ and $Cr(p_B|a_B)$, and Causalists are already committed to rejecting such reasoning. The force of the case in the first place—what ensures it is not just a variation on Newcomb's Problem—is that, even assuming that we are happy to accept reasoning on the basis of independence, there is something worrying about passing up the safe option a_E in favour of the risky options a_A and a_B. It cannot, for Causalists, be an additional factor that those options are even more risky if you pay attention to the conditional probabilities.

A dynamic CDT does not help to capture the intuition in favour of taking the envelope on its own. But since we might want to appeal to such an approach in other cases of decision-instability it is important to check that combining risk-aversion with a dynamic approach can still vindicate the intuition in favour of taking the envelope. Fortunately, a dynamic and risk-weighted CDT can do so. In particular, given risk-aversion (and unlike the risk-neutral case) being more sure that you will *not* take the envelope makes you more certain that taking the envelope is the better option, since a_A and a_B are more risky. This ensures that a risk-weighted dynamic CDT will lead us to the equilibrium of being indif-

[10]Except the extreme risk function with $r(p) = 0$ for all $p \neq 0$—which corresponds to considering on the worst possible outcome of each option.

ferent between all three options—rather than ruling out a_E. This is enough to vindicate the intuition that taking the envelope can be rational (since some agents in such an equilibrium will take the envelope). But Spencer and Wells' can sensibly suggest that taking the envelope should be more than just rationally permissible (without going so far as to make it rationally required in violation of the arguments above). Here it is helpful to consider the credence assigned to the various options in the equilibrium for risk-weighted dynamic CDT. These credence depend on how good you think the prediction is; assume that you have $Cr(p_X|a_X) = 0.9$ and $Cr(p_X|\neg a_X) = 0.05$ for all three options—this represents taking the predictor to be correct 90% of the time and to evenly distribute the incorrect predictions over the other options. And assume, again, that your risk function is $r(p) = p^2$. Then the equilibrium occurs when:[11]

$$Cr(p_A) = Cr(p_B) = \frac{1 - \sqrt{\frac{3}{5}}}{2} \approx 0.11$$
$$Cr(p_E) = \sqrt{\frac{3}{5}} \approx 0.78$$

And hence:

$$Cr(a_A) = \frac{\frac{1-\sqrt{\frac{3}{5}}}{2} - 0.05}{0.85} \approx 0.07$$
$$Cr(a_B) = \frac{\frac{1-\sqrt{\frac{3}{5}}}{2} - 0.05}{0.85} \approx 0.07$$
$$Cr(a_E) = \frac{\sqrt{\frac{3}{5}} - 0.05}{0.85} \approx 0.85$$

That is, the equilibrium occurs when you are relatively certain that you will choose the

[11] You can check this be confirming that these credences for the states give all three options equal risk-weighted expected utility.

safe option—although you are indifferent over all three options. What exactly this tells us depends on how we want to interpret the credences assigned to options at equilibrium in dynamic deliberation. They must represent the agent's views about what they will do—but do they also constrain what the agent will do? One option is to say no: since the agent is indifferent between all options with non-zero credence they should simply pick one of those and these credences play no role. The other option is to interpret the equilibrium as the agent's mixed strategy: so that they will play each of these option with the corresponding probability. The interpretation will get us closer to vindicating the intuition that you should, most of the time, take the envelope. Either way, however, a dynamic risk-weighted CDT vindicates both that (a) taking the envelope can be rational and (b) rational agents are quite likely to take the envelope.

So what have we learnt about The Frustrater? Well, I have shown the following:

- A risk-weighted and non-dynamic CDT will endorse taking the envelope for some credences and some risk functions.

- A risk-weighted and dynamic CDT will endorse being indifferent between all three options and will predict that you are highly likely to take the envelope.

And this makes it very plausible that:

- Intuitions in favour of taking the envelope in The Frustrater are driven by risk-aversion.

And, hence, Causalists have two options for responding to the problem:

- If they are accept the rationality of risk-aversion, then they should simply grant the intuition in favour of taking the envelope and apply a risk-weighted CDT—but this will be nothing new, since the same is required to capture the Allais preferences and other instance of risk-aversion (and Buchak has provided an appropriate decision theory for Causalists to do so).

- If they reject the rationality of risk-aversion (for pure *utility*, rather than for money and other concrete goods), then they should simply reject the intuition to take the envelope as irrational and offer some story to explain why it is, nonetheless, attractive—again, this will be nothing new, since some such story will be required to explain why the Allais preferences and other instances of risk-aversion are attractive despite being irrational.

Either way, there is no *new* problem here; merely the combination of two old problems—risk-aversion and decision-instability. Causalists aren't as easily frustrated as Spencer and Wells' have suggested.

5 Back to Psychopaths

We have seen that Spencer and Wells' new counterexample involves both decision-instability and risk-aversion but poses no new challenge for CDT. It is natural at this point to ask whether risk-aversion might be playing a role in other examples of decision-instability that trouble Causalists. Risk-aversion cannot play a role in cases of *symmetric* decision-instability, such as Death in Damascus, because such cases involve the same risks for all options (otherwise they would not be symmetric). And we have seen that a dynamic deliberation approach to CDT is very plausible in such cases. The dynamic deliberation approach has somewhat more difficulty, however, with cases of asymmetric decision-instability. Fortunately, appeal to risk-aversion can at least reduce the sting of such cases. I will demonstrate this with Egan's classic Psychopath Button:

> Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul

	Psycho	¬Psycho
Press	-1000	100
¬Press	0	0

Table 4: Decision matrix for Psychopath Button.

> very strongly prefers living in a world with psychopaths to dying. Should Paul press the button? (Egan 2007, 97)

Table 4 shows a decision matrix with the utility values. Now, Paul is quite confidence that only a psychopath would press the button—say $Cr(\text{Psycho}|\text{Press}) = 0.9$. For simplicity we will also assume he is quite confidence that all psychopaths would press, so $Cr(\text{Psycho}|\neg\text{Press}) = 0.05$. And, of course, whether Paul is a psychopath is causally independent of whether he presses the button—pressing the button, or being inclined to do so, is merely evidence that he is already a psychopath and not a cause of it. If, as seems plausible, Paul starts out quite confidence he is not a psychopath—say, $Cr(\text{Psycho}) = 0.95$—then it is trivial to calculate that CDT will advise Paul to press the button:

$$\begin{aligned} EU_{CDT}(\text{Press}) &= -1000 \times Cr(\text{Psycho}) + 100 \times Cr(\neg\text{Psycho}) \\ &= -1000 \times 0.05 + 100 \times 0.95 = 45 \\ EU_{CDT}(\neg\text{Press}) &= 0 \end{aligned}$$

Yet many have the intuition that Paul should not press the button.

This is a case of decision-instability—from CDT's perspective at least—because being more likely to press the button increases Paul's credence that he is a psychopath, and hence makes pressing the button look like a worse option. Whereas being more likely not to press the button increases Paul's credence that he is not a psychopath, and hence makes pressing the button look like a better option.

How does risk-aversion affect CDT's verdict in this case? Again assume, for example, that Paul has the risk function $r(p) = p^2$. Then even with the relatively low disutility for dying of -1000, risk-weighted CDT will advise Paul not to press the button:

$$\begin{aligned} EU_{RCDT}(\text{Press}) &= -1000 + 1100 \times r(Cr(\neg\text{Psycho})) \\ &= -1000 + 1100 \times 0.95^2 = -7.25 \\ EU_{RCDT}(\neg\text{Press}) &= 0 \end{aligned}$$

Of course, even a risk-weighted CDT will *sometimes* endorse pressing the button, if Paul's initial credence that he is not a psychopath is high enough or if his risk tolerance is high enough. But this on its own cannot be an objection because *any* plausible decision theory will have to say there is some level of certainty at which Paul should press the button—otherwise we would be endorsing the irrational policy of never taking *any* risk of death in order to potentially achieve something good. Note that even Evidential Decision Theory—which is usually taken to do better with decision-instability and this case in particular—will endorse pressing the button if Paul's *conditional* credence that only a psychopath would press the button is low enough or if the utility of killing all the psychopaths is high enough. Never pressing the button in this case is irrational. A risk-weighted CDT endorses pressing the button for a significantly smaller range of credences than a risk-neutral CDT—which is enough to vindicate the relevant intuitions.

What about dynamic CDT? Once again, a risk-neutral dynamic CDT has an equilibrium where both options are assigned some credence. Given the numbers above a risk-neutral dynamic CDT carries us to an equilibrium where $Cr(\text{Press}) \approx 0.05$ and $Cr(\neg\text{Press}) \approx 0.95$—already quite far from likely to press. However, a risk-weighted dynamic CDT, for the same numbers, will have an equilibrium where $Cr(\neg\text{Press}) = 1$—since the risk-weighted expected utility of pressing is lower than not pressing. In general, a dynamic

CDT makes pressing the button impermissible for a wider range of credences about the correlation between pressing and being a psychopath.

All this suggests, once again, that risk-aversion may be playing a role in driving the intuitions that are supposed to trouble CDT in Psychopath Button. Once again, Causalists can interpret this fact in two ways, depending on their general view of risk-aversion: as further confirmation that a risk-weighted CDT is the right theory, or as further confirmation that irrational risk-aversion drives mistaken intuitions.

6 Dicing with Death

James Joyce (2018, 155) has claimed that Spencer and Wells' Frustrater is equivalent to a case presented by Ahmed (2014). Since Ahmed does not name the case, I will call it *Dicing with Death*. However, once we see the role of risk-aversion in The Frustrater we can see that these cases are not equivalent, and that Causalists will need different resources to deal with them. This should, I think, reassure us about the claim that risk-aversion is involved in The Frustrater (and Psychopath Button) because it shows that appeal to risk-aversion does not overgeneralise to cases which intuitively do not involve it.

Dicing with Death is identical to Death in Damascus *except* that you also have the third option of flipping a magic (fair) coin which will send you to Damascus if it lands heads and to Aleppo if tails. This coin's most important magical property is that Death cannot predict its outcome, nor predict that you will use it. Hence, if you elect to use the coin you will have an objective 50% chance of ending up in the city which Death has predicted you will be in and, hence, a 50% chance of surviving. A decision matrix summarising all this is shown in Table 5.

Ahmed claims that it is obvious that you should use the coin (and be willing to pay to do so, but we will ignore that complication here). However, standard CDT will still

	$D_D\&H$	$D_D\&T$	$D_A\&H$	$D_A\&T$
Aleppo	0	0	-1000	-1000
Damascus	-1000	-1000	0	0
Coin	0	−1000	−1000	0

Table 5: Decision matrix for Dicing with Death.

recommend that you go to whichever city you are initially least confident Death will be in—forgoing the coin.[12] While dynamic CDT will suggest that you should be indifferent between all three options; since they are symmetric with respect to your credence in the states conditional on what you will do.

Risk-averse CDT, of either the static or dynamic kind, will in this case *also* suggest indifference between the three options—because they all have the same worst outcome and same probabilities of leading to the better outcome. That is, risk-aversion makes no difference here. And this makes sense: In The Frustrater you are offered a third *risk free* option which risk-aversion will take account of. Whereas here you are offered a third option that is just as risky—although the risk is of a different kind.

At least in the dynamic equilibrium, what the coin essentially offers is a way to swap an *epistemic* probability of 0.5 that you will survive by going to Aleppo or Damascus for an objective chance of 0.5 that you will survive by flipping the coin. At first this might sound like a good deal, but we need to be careful. Conditional on your choosing to go to Aleppo (Damascus) you have a credence of 0.5 that you will die. But this credence represents the average of two different chance situations you take to be possible: if you choose to go to Aleppo then the objective chance that you will die is either 1—if Death has predicted Aleppo—or 0—if he has predicted Damascus. So what you are being offered with the coin is the opportunity to take the average of these two chances. Is that a good deal? Well, it is if your chance of dying is 1—then you reduce it by 0.5—but it is not if your chance of dying

[12]It will suggest indifference if you initially have exactly 0.5 credence in Death being in each city.

is 0—then you *increase* it by 0.5. It is very plausible here, I would suggest, that you should be indifferent between these options—even if you are risk-averse—because taking the coin is just as likely to make your situation more risky as to make it less risky.

You may or may not find this convincing, and may or may not still find Dicing with Death a problematic example for CDT (risk weighted, dynamic, or not). But I would suggest that seeing that risk-aversion is not a panacea for CDT's problems should make us more, rather than less, confident that risk-aversion is driving intuitions in The Frustrater (and Psychopath Button) and that, therefore, Causalists can respond to those examples in whatever way they want to deal with risk-aversion more generally.

7 Two Rooms and the Guaranteed Principle

Finally, Spencer (2021) has more recently presented a diachronic elaboration of The Frustrater which he takes to be even more problematic for CDT. Will appeal to risk-aversion help in this case as well? First, here is the case, *Two Rooms*:

> An agent must enter either Room #1 or Room #2. If she enters Room #1, she gets $35. If she enters Room #2, she faces The Frustrater. The agent knows all of this (Spencer 2021, 4).

Spencer argues that CDT will endorse entering Room #1 and that this is irrational. In reverse order: Spencer takes it that entering Room #1 is irrational because it violates the following principle (paraphrased from Spencer 2021):

> **The Guaranteed Principle:** Faced with a choice between two courses of action, one of which leads to a guaranteed utility of x and the other of which leads to a choice where at least one option guarantees y, with $y > x$, you should always prefer the course of action which leads to the choice that allows you to get y.

Entering Room #1 violates this principle because Room #1 guarantees you \$35, whereas entering Room #2 would allow you to choose a_E and get a guaranteed \$40.

Now, Spencer argues that CDT endorses entering Room #1 because, from your perspective before entering either room, you expect that if you enter Room #2 you will take either a_A or a_B[13] *but* that the expected utility of doing so is *less than* \$35. How could this be so, given that the expected utilities for a_A and a_B must be greater than \$40 once you enter the room (or else it would not be rational to take them)? It is because CDT calculates expected utility for an *action* differently depending on whether it is being considered as a present *option* or merely a possible future state of the world. The expected utilities of actions as possible future states of the world are calculated the same way as the expected utilities of any other state of the world: with conditional, evidential, probabilities. Only the expected utilities of your present options are calculated instead using unconditional probabilities (or causal probabilities of an appropriate kind). The expected utility of a_A considered as a possible future state is, hence:

$$EU(a_A) = 0 \times Cr(p_A|a_A) + 100 \times Cr(p_B|a_A) + 50 \times Cr(p_E|a_A)$$

Since $Cr(p_A|a_A)$ is high, while $Cr(p_B|a_A)$ and $Cr(p_E|a_A)$ are low—the predictor is accurate—this expected utility will be close to zero (or at least lower than EU(Room #1) = \$35) for most credences. And mutatis mutandis for a_B. Since a standard Causalist expects to take a_A or a_B (or possibly to randomise between them) then they expect entering Room #2 to be worse for them than entering Room #1 and taking the \$35.

I'm sure you can see what comes next: Spencer's whole argument falls down if CDT actually endorses taking taking the envelope—which *always* has expected utility of \$40. In that case the Causalist will expect to get \$40 from entering Room #2 and so will prefer

[13]Given the caveat that you expect to remain rational, expect to be a Causalist, and expect your utilities and probabilities to remain the same—none of which I shall challenge.

doing so to entering Room #1, vindicating Spencer's intuition. And, as we have just seen, a risk-weighted CDT endorses taking the envelope! So, if risk-weighted CDT is the right way to go, Two Rooms is no more threat to CDT than The Frustrater is.

Causalists who reject risk-aversion and the intuition in favour of taking the envelope in The Frustrater cannot appeal to this defence. And, unfortunately, in this case it is not obvious that the intuition against entering Room #1 can be explained by risk-aversion—so they cannot appeal to that as a reason to debunk the intuition. Can they avoid the result or undermine Spencer's intuition against entering Room #1 in some other way? Rothfus (2022) has recently proposed a plan based CDT that may allow Causalists to avoid the result of entering Room #1. On the other hand, cases where CDT has trouble with diachronic inconsistency are not new[14] and Causalists might assimilate Two Doors to those cases rather than to risk-aversion.[15] In the end, if these do not work, Two Rooms might simply be taken as more reason to embrace a risk-weighted CDT.

References

Ahmed, Arif. 2014a. "Dicing with Death." *Analysis* 74 (4): 587–592. https://doi.org/10.1093/analys/anu084.

———. 2014b. *Evidence, Decision and Causality*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139107990.

Allais, Maurice. 1953. "Le Comportement de l'Homme Rationnel Devant Le Risque: Critique Des Postulats et Axiomes de l'Ecole Americaine." *Econometrica* 21 (4): 503. https://doi.org/10.2307/1907921.

[14] See, for example, (Ahmed 2014b, Ch. 8).

[15] Either by arguing that CDT will avoid the problematic results in such cases as Joyce (2016) does, or by arguing that diachronic inconsistency is not such a high cost as Hedden (2015) does.

Armendt, Brad. 2019. “Causal Decision Theory and Decision Instability.” *The Journal of Philosophy* 116 (5): 263–277. https://doi.org/10.5840/jphil2019116517.

Arntzenius, Frank. 2008. “No Regrets, or: Edith Piaf Revamps Decision Theory.” *Erkenntnis* 68 (2): 277–297. https://doi.org/10.2307/40267481. JSTOR: 40267481.

Buchak, Lara. 2016. “Decision Theory.” In *The Oxford Handbook of Probability and Philosophy,* edited by Alan Hájek and Christopher Hitchcock, 769–816. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199607617.013.40.

Egan, Andy. 2007. “Some Counterexamples to Causal Decision Theory.” *Philosophical Review* 116 (1): 93–114. https://doi.org/10.1215/00318108-2006-023.

Ellsberg, Daniel. 1961. “Risk, Ambiguity, and the Savage Axioms.” *The Quarterly Journal of Economics* 75 (4): 643. https://doi.org/10.2307/1884324.

Gibbard, Alan, and William Harper. 1981. “Counterfactuals and Two Kinds of Expected Utilities.” In *Ifs,* edited by William L. Harper, Robert Stalnaker, and Glenn Pearce, 152–191. Dordrecht: Springer Netherlands.

Hedden, Brian. 2015. *Reasons without Persons: Rationality, Identity, and Time.* Oxford: Oxford University Press.

Joyce, James. 2002. “Levi on Causal Decision Theory and the Possibility of Predicting One’s Own Actions.” *Philosophical Studies* 110 (1): 69–102. JSTOR: 4321285.

———. 2012. “Regret and Instability in Causal Decision Theory.” *Synthese* 187 (1): 123–145. https://doi.org/10.1007/s11229-011-0022-6.

———. 2016. “Review of Arif Ahmed, Evidence, Decision and Causality.” *The Journal Of Philosophy* 113 (4): 224–232.

Joyce, James. 2018. “Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems.” In *Newcomb’s Problem,* edited by Arif Ahmed, 138–159. Classic Philosophical Arguments. Cambridge: Cambridge University Press.

Lewis, David. 1981. “Causal Decision Theory.” *Australasian Journal of Philosophy* 59 (1): 5–30. https://doi.org/10.1080/00048408112340011.

Rothfus, Gerard. 2022. “A Plan-Based Causal Decision Theory.” *Analysis* 82 (2): 264–272. https://doi.org/10.1093/analys/anab064.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation.* Cambridge, Mass: Harvard University Press.

Spencer, Jack. 2021. “An Argument against Causal Decision Theory.” *Analysis* 80 (1): 52–61. https://doi.org/10.1093/analys/anaa037.

Spencer, Jack, and Ian Wells. 2019. “Why Take Both Boxes?” *Philosophy and Phenomenological Research* 99 (1): 27–48. https://doi.org/10.1111/phpr.12466.