Frustration and Delay

Abstract. A decision problem where Causal Decision Theory (CDT) declines a free $1,000, with the foreseeable effect that the agent is $1,000 poorer, and in no other way better off, than if she had taken the offer.

1. *Frustration and Delay*. Here are two boxes, A and B. One contains $1M. The other contains nothing. And here is $1K. Right now you have three options.

- You can take Box A now.
- You can take Box B now.
- You can take this $1K now, no strings attached, and choose between A and B in five minutes.

Five minutes ago we conducted a brain-scan to determine which box you would ultimately (i.e. now or later) take on receiving these instructions. The scan detects a short-lived electrical signature that always precedes decisions of this type. There are two types of signature. The *A-signature* indicates a brain state that makes it likely (with chance $c \approx 1$) that you will ultimately take Box A. The *B-signature* indicates a brain state that makes it likely (with chance $c \approx 1$) that you will ultimately take Box B. If we detected the A-signature then we put $1M in Box B and nothing in Box A. If we detected the B-signature then we put $1M in Box A and nothing in Box B.

You know that taking the $1K up front makes *no* difference, causal or stochastic, to your signature or to your choice. To be completely clear about this, we should emphasize three ways in which it makes no difference.

First, subjects with the A-signature take the $1K with the same relative frequency as subjects with the B-signature. So taking the $1K tells you nothing about whether you have the A-signature or the B-signature.

Second, subjects with the A-signature who *take* the $1K are as likely to take (have the same chance of taking) Box A as are subjects with the A-signature who *turn it down*. The same goes for subjects with the B-signature. So taking the $1K makes no causal difference to which box you will ultimately take.

Third, taking the $1K makes no difference to the whereabouts of the $1M. This has already been fixed and there is no way that anyone can change it.

So it really is a free $1K: taking it makes no conceivable difference to anything else that happens in this problem.

Do you take Box A now, Box B now, or the $1K up front?

2. Obviously, you should take the $1K up front. Taking it makes no difference to your chance of making $1M. The only effect of taking it would be to make you $1K better off. So of course you should take it; but Causal Decision Theory (CDT) advises you to turn it down.[1]

Here's why. Let your present options be A, B, X. We can treat them as propositions describing your possible behaviour. A is the proposition that you now take Box A. B is the proposition that you now take Box B. X is the proposition that you now take the $1K.

Next, we specify all the possible ways in which the outcome (your final wealth) depends causally on which option you now take. These possible ways are called dependency hypotheses. In this problem there are two dependency hypotheses.

[1] My argument here uses Lewis's (1981) formulation of CDT.

The first dependency hypothesis is H_A. H_A is the proposition that the A-signature was present five minutes ago. If H_A is true then the pattern of causal dependence, between what you do and what you get, is as follows:

(1) If you were to choose A then you'd get nothing.

(2) If you were to choose B then you'd get \$1M.

(3) If you were to choose X then you'd have a high chance c of getting \$1K and a low chance $1 - c$ of getting \$(1M+1K).

(1) is true because if the A-signature was present five minutes ago then there is now nothing in Box A and \$1M in Box B. So if you were *now* to take Box A, forgoing the \$1K up front, you would get nothing. (2) is true for the same reason. (3) is true because if you choose to take the \$1K, this makes no difference to your chance of ultimately taking Box A. But if the A-signature was present then your chance of ultimately taking Box A is c. So if you choose option X, there is still a chance c that you will take Box A (which is empty) and a chance $1 - c$ that you will take Box B (with \$1M). And by choosing option X you ensure that you already have \$1K. So (3) is true.

The second dependency hypothesis is H_B. H_B is the proposition that the B-signature was present five minutes ago. If H_B is true then the pattern of causal dependence, between what you do and what you get, is as follows:

(4) If you were to choose A then you'd get \$1M.

(5) If you were to choose B then you'd get nothing.

(6) If you were to choose X then you'd have a high chance c of getting \$1K and a low chance $1 - c$ of getting \$(1M+1K).

The arguments for (4)-(6) exactly parallel the arguments for (1)-(3). Note that (3) and (6) say the same.

Now that we have the options and the dependency hypotheses, we can determine the causal utility U of each option. We calculate this as follows. For each option, take the weighted sum of the values of the prizes that it yields under each dependency hypothesis, where the weight of each summand is your credence in the corresponding dependency hypothesis.

Thus consider option A. The value of A under dependency hypothesis H_A is the value of \$0. The value of option A under option dependency hypothesis H_B is the value of \$1M. Let Cr symbolize your credence function, so that your credences in each of these dependency hypotheses is $Cr(H_A)$ and $Cr(H_B)$ respectively, where $Cr(H_A) + Cr(H_B) = 1$. Then assuming linear value for money, the causal utility of taking Box A at the outset is:

(7) $U(A) = 0.Cr(H_A) + M.Cr(H_B) = MCr(H_B).$

And by an exactly parallel argument the causal utility of taking Box B at the outset is given by:

(8) $U(B) = M.Cr(H_A) + 0.Cr(H_B) = MCr(H_A).$

What about the causal utility of taking the $1K at the outset? By (3) and (6), the value of this option is the same on either dependency hypothesis: it is a lottery that has a chance c of yielding a $1K prize and a chance $1 - c$ of yielding a $1M + $1K prize. On von Neumann and Morgenstern's (1953: 16-27) assumptions concerning the values of lotteries, and still assuming linear value for money, the causal utility of X is:

(9) $U(X) = cK + (1-c)(M+K) = K + M(1-c)$

CDT recommends the option that maximizes causal utility. It is easy to see from (7), (8) and (9) that if c is close to 1 then this is never X. For suppose:

(10) $U(X) \geq U(A)$

(11) $U(X) \geq U(B)$

Then adding (10) and (11):

(12) $2U(X) \geq U(A) + U(B)$; so by (7)-(9):

(13) $2U(X) = 2\big(K + M(1-c)\big) \geq MCr(H_B) + MCr(H_A) = M$ i.e.:

(14) $\frac{K}{M} + \frac{1}{2} \geq c$

But this contradicts the assumption that $c \approx 1$. So either $U(A) > U(X)$ or $U(B) > U(X)$.

So *whatever* your present credences about the presence, five minutes ago, of the A-signature or of the B-signature, *either* the causal utility of option A exceeds the causal utility of option X, *or* the causal utility of option B exceeds the causal utility of option X. CDT therefore recommends throwing away $1K to no good effect.

3. But how can *Causal* Decision Theory recommend turning down a free $1K that makes no *causal* difference to whether you ultimately get $1M? I must have misunderstood what CDT is saying.

More precisely. Think about your beliefs when you are facing the problem, first at the first stage and then, on the assumption that you take option X, at the second stage. Let us ask: would your choosing X make any difference to your opinion about your signature?

Answer: no. This follows from our assumptions about the case. The instructions for the problem emphasized that taking option X makes *no* conceivable difference to what you will ultimately do, what your signature is or where the $1M is.

Now suppose that for whatever reason you start out confident that the $1M is in Box B, with $Cr(H_A) > 0.5$. So CDT prefers taking Box B now to taking Box A now. But we know that $Cr(H_A|X) = Cr(H_A) > 0.5$: your taking the $1K makes no difference to what you will think that your signature is, or to any other belief that is relevant to the choice between Box A and Box B. So CDT will give the same recommendation as before, namely taking Box B. The effect of taking the $1K now can therefore only be to make you $1K richer. Therefore if CDT *now* prefers taking Box B to taking Box A, then CDT now prefers taking the $1K first to taking Box B now. CDT therefore recommends option X, contrary to my argument. A similar argument applies on the assumption that you start out confident that the $1M is in Box A.

Or suppose you start out maximally agnostic about the whereabouts of the $1M, so that $Cr(H_A) = Cr(H_B) = 0.5$. In that case CDT is indifferent between taking Box A now and taking Box B now. But again, if your credences in having the A-signature or the B-signature do not shift, and if you do not expect them to shift as a result of taking the $1K, then CDT's evaluation of the options of taking Box A and of taking Box B will go as initially, and the utilities of each option before and after taking the $1K will agree. CDT's assessment of X at the outset will therefore give it a causal utility that is K more than the corresponding assessments of A or B, so again CDT recommends X after all, contrary to the argument.

4. In reply, let us first write $Cr(H_A) = p, Cr(H_B) = 1 - p$.

You have a present level of confidence p that the $1M is in Box A and $1 - p$ that the $1M is in Box B. CDT therefore regards taking Box A now as a lottery with an expected chance p of winning $1M. And it regards taking Box B now as a lottery with an expected chance $1 - p$ of winning $1M. In a choice between *these two* options, CDT prefers taking Box A now if $p > 1 - p$. It prefers taking Box B now if $1 - p > p$, and it is indifferent if $p = 0.5 = 1 - p$.

None of this would change if you were to take the $1K up front. After you have done that, CDT will still value the option of taking Box A at the same rate as a lottery in which the expected chance of winning $1M is p, and it will still value the option of taking Box A at the same rate as a lottery in which the expected chance of winning $1M is p.

But right now, at the first stage, you also know that there is in fact, now and at any later stage, a high chance c that you will in fact pick the empty box. After all, you know that if H_A is true then the $1M is in Box B and there is a chance c that you will ultimately pick Box A. And you know that if H_B is true then the $1M is in Box A and there is a chance c that you will ultimately pick Box A. And taking the $1K upfront is not going to affect which Box you will take. So taking the \$1K upfront is effectively taking a \$1K payment to enter a lottery for \$1M that you have only a $1 - c$ chance of winning.

But either taking Box A *now* looks better than that, or taking Box B now looks better than that, from the perspective of CDT. After all, your confidence that there is now \$1M in Box A is p. So your confidence that the taking Box A now would net \$1M is also p. Similarly, your confidence that the taking Box B now would net \$1M is $1 - p$. And at least one of these two quantities, p and $1 - p$, exceeds $1 - c$ by enough to make it worth your while to take one of the boxes now, even at the cost of \$1K. At any rate this holds by the lights of CDT, as we saw in §1 when we argued that either $U(A) > U(X)$ or $U(B) > U(X)$.

But if CDT prefers (say) taking Box B now to taking Box A now, then shouldn't it prefer *taking Box B after taking the \$1K* to both? Well yes, but *that is not a present option*. Your present options are: (A) take Box A now, (B) take Box B now, and (X) take the \$1K *and choose later* between the boxes. But if your present self chooses to defer to your future self the decision about which box to take, your present self is not choosing which box to take. There is no option to bind your future self.

The objector of §2 is right that the way CDT *now* evaluates the options of taking Box A and Box B *now* is the same as the way CDT *later* evaluates the options of taking Box A and Box B *then*. But the way CDT *now* evaluates the options of taking Box A and Box B *now* is *not* the same as the way CDT *now* evaluates the option of deferring that choice to *then*. It is of the essence of CDT to evaluate presently contemplated options very differently from the way that it evaluates those same options when they are causally downstream of presently contemplated options.[2] And this means that right now, CDT prefers taking one of the boxes

[2] 'Viewed in prospect, acts in future decisions are treated not as current options, but as potential outcomes lying causally downstream of your current choice. As with anything not under your current control, CDT assesses future acts using their current news values' (Joyce 2018: 146). What Joyce calls the current news value of my

now to taking the \$1K up front, even though it knows that taking the \$1K upfront makes no causal difference, *either* to which box you do take, *or* to whether there is money in it.

5. So far I've been discussing Lewis's version of CDT, which relies on a partition of the set of possible worlds into dependency hypotheses, each of which specifies a pattern of counterfactual dependence between options and chance distributions over outcomes, the partition in this example being $\{H_A, H_B\}$.

But the argument goes for every other version of CDT in which the problem can be stated, including those of Gibbard and Harper[3], Skyrms[4], Sobel[5], the early Joyce[6] and, as I now illustrate, Joyce's more recent theory.

I have in mind Joyce's *deliberational* CDT, as endorsed in a 2012 paper following work of Arntzenius and Skyrms. In a more recent discussion he has characterized it as directing the agent not directly to maximize causal utility (U) but rather to maximize this quantity *after taking into account all readily available information about what her acts may cause*. This means that we need to consider the U-values of the options when *the options to which you attach positive probability are all U-maximizing*. For if you attach positive probability to some option O that is *not* U-maximizing, the fact that other available options are U-superior to it should be evidence for you (as a U-maximizer) that you will not realize O; and so taking this evidence into account should depress your confidence in O. If this process stops anywhere

taking Box A (Box B) *after delaying* is about \$1K, because news that I will do this strongly indicates that the \$1M is in Box B (A).

[3] Gibbard and Harper 1978: 345f.

[4] Skyrms 1980: 133.

[5] Sobel 1989: 73.

[6] Joyce 1999: 161.

then it stops at an equilibrium in which the options to which you attach positive probability are all U-maximizing.[7]

What equilibria, if any, are there at the first stage of *Frustration and Delay*? Clearly there are none at which $Cr(X) > 0$. We already saw that in *any* state of belief, $U(A) > U(X)$ or $U(B) > U(X)$. For by (7)-(9), if you are more confident in H_A than in H_B then CDT will prefer B to X; if you are more confident in H_B than in H_A then CDT will prefer A to X; and if you are equally confident in both then CDT will prefer *both* A and B to X.

So the only equilibria that neo-Joycean CDT permits are states at which you are certain that you will not take the \$1K upfront. Such equilibria do exist: most obviously the state in which $Cr(H_A) = Cr(H_B)$ and $U(A) = U(B)$, i.e. you are indifferent between taking Box A now and taking Box B now. In fact it is easy to see that on natural assumptions you must be indifferent between these options in *any* equilibrium.[8] But what matters here is not the full range of solutions that neo-Joycean CDT permits, but rather that it, like every other version of CDT, does not permit any solution in which you take the \$1K upfront.[9]

6. In connection with Newcomb's Problem, which was the example that motivated CDT in the first place, one often hears defenders of CDT make the following points: that taking the \$1K in that case would never *make* any difference to whether one also made \$1M; that declining it would therefore amount to giving up a free \$1K; and that everyone knew all this in

[7] Joyce 2018: 148ff.; see also Joyce 2012; Arntzenius 2008; Skyrms 1990.

[8] The assumption is that (i) $Cr(H_A|A) > Cr(H_B|A)$ if $Cr(A) > 0$ (ii) $Cr(H_B|B) > Cr(H_A|B)$ if $Cr(B) > 0$. Suppose that $U(A) > U(B)$ in equilibrium. Therefore A is uniquely optimal (according to CDT), so $Cr(A) = 1$ since we are at an equilibrium. By (7) and (8), $U(A) > U(B)$ implies $Cr(H_B) > Cr(H_A)$. But this contradicts (i). Similarly, the assumption that $U(B) > U(A)$ contradicts (ii). Therefore $U(A) = U(B)$.

[9] The same goes for Armendt's recent endorsement of what he calls 'unadorned' CDT, according to which you should, at *off*-equilibrium points, do what CDT recommends as applied to your credences and utilities at that point, rather than (as per Joyce) what it would recommend in some as-yet-unreached equilibrium (Armendt 2019). Even off equilibrium $Cr(H_A) + Cr(H_B) = 1$; it then follows from (7), (8) and (9) that either $U(A) > U(X)$ or $U(B) > U(X)$, as long as c is big enough.

advance.[10] It is ironic that these points all apply to the $1K that CDT declines in *Frustration and Delay*.[11]

[10] See Nozick 1969: 208 for the classic statement of Newcomb's Problem.

[11] I distinguish the case studied in this paper from three other alleged counterexamples to CDT.

In a mildly asymmetric *Death in Damascus* due to Richter that Levinstein and Soares recently discussed (Richter 1984; Levinstein and Soares forthcoming), the arrangement resembles *Frustration and Delay* except that instead of the option to delay, there is (you know) an additional $1000 in Box B. Here it might seem obvious that taking Box B is the only rational choice; but CDT (if it advises anything) appears also to endorse taking Box A. At least, if you are slightly more confident that the $1M is in Box A than that it is in Box B, which in a deliberational equilibrium (see §4) you will be, CDT is indifferent between these options. But arguably if those are your credences, then taking Box A *ought* to look as good as taking Box B: since you are slightly more confident that Box A is worth $1M, you should be willing to take it, even if that means foregoing the $1K associated with Box B.

Similarly, *Dicing with Death* (Ahmed 2014) modifies *Frustration and Delay* by replacing the option to wait with a third option, which in present terms amounts to paying a small amount to use a randomizing device that chooses unpredictably between Box A and Box B. Ahmed objects to CDT that it rejects this option. Joyce's response is that if you are 50-50 about the whereabouts of the $1M, you have 50% confidence that taking Box A directly gives you a 100% chance of winning $1M and 50% confidence that it gives you a 100% chance of winning nothing; and the same goes for the option of taking box B directly. As for randomization, you have 100% confidence that it gives you a 50% chance of winning $1M and a 50% chance of winning nothing. Setting aside Ellsberg-type preferences ('ambiguity aversion'), these gambles *should* all look equally good. 'In terms of your subjective estimates of [probabilities of winning $1M], all three acts offer the same thing. So, paying to [randomize] would be paying for what you already take yourself to have' (Joyce 2018: 156).

But whatever their merits against *Dicing with Death* or asymmetric *Death in Damascus*, these responses are ineffective against *Frustration and Delay*. As in *Dicing with Death*, you are (let's suppose) 50-50 about the whereabouts of the $1M, you have 50% confidence that taking Box A (or B) directly would give you a 100% chance of winning $1M, and you have 50% confidence that it would give you a 100% chance of winning nothing. But you *also* have 50% confidence that taking Box A (B) *after taking the $1K* would give you a 100% chance of winning $1M and 50% confidence that it would give you a 100% chance of winning nothing; and the same goes for the option of taking Box B after taking the $1K. Taking the $1K now gives you a later choice between gambles that are just as good, by Joyce's own standards, as the ones available now. Taking Box A or Box B directly, as CDT recommends, would in effect be giving up a $1K bonus for a gamble that you would have got anyway.

Finally, the case owes something to a version of *Dicing with Death* due to Spencer and Wells, which they (2019: 34) call *The Frustrater*. Here there are two opaque boxes, A and B, and an envelope. The agent can take A, B, or the envelope. The envelope contains $40. One box contains $100. Which one it is depends on the reliable prediction of a 'Randomizing Frustrater'. If he predicted that the agent takes A, he put $100 in B. If he predicted that the agent takes B, he put $100 in A. If he predicted that the agent takes the envelope, he put $100 in A or B based on the toss of a coin. Spencer has since (see Spencer forthcoming) developed a two-stage extension of the case, which he calls *Two Rooms*. The second stage of *Two Rooms* is *The Frustrater*. At the first stage, the agent has the take-it-or-leave it option to pay $5 to bind herself to take the envelope. Spencer argues that an agent who knows she will follow CDT will pay $5 at the first stage of *Two Rooms*, and that this violates what he calls the Guaranteed Principle, which says that if $m > n$ then a rational (and money-seeking) agent will always prefer a decision that makes $\$m$ available to one that forces an outcome of $\$n$.

Frustration and Delay resembles *Two Rooms* in that both cases involve two stages, and both involve a *Frustrater*-style arrangement. My argument resembles Spencer's in that both rely on the fact that CDT takes a different attitude towards present and future options (see Spencer forthcoming: 8-11). But the difference ends there. Spencer makes the descriptive assumption that the agent knows at the first stage that she will follow CDT at the second, whereas my argument does not. And Spencer makes the normative assumption that the Guaranteed Principle is true. Whereas my argument makes a different normative assumption: that it would be irrational to give up $1K when the *only* effect of taking it would be to make you that much richer. The argument from *Two Rooms*, powerful as (in my view) it is, is not at all the same as the argument from *Frustration and Delay*.

References

Ahmed, A. 2014. Dicing with death. *Analysis* 74: 587-92.

Armendt, B. 2019. Causal Decision Theory and decision instability. *Journal of Philosophy* 116: 263-77.

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

Gibbard, A. and W. Harper. 1978. Counterfactuals and two kinds of expected utility. In Hooker, C., J. Leach and E. McClennen (ed.), *Foundations and Applications of Decision Theory*. Dordrecht: Riedel: 125-62. Reprinted in Gärdenfors, P. and N.-E. Sahlin (ed.), *Decision, Probability and Utility* (1988). Cambridge: CUP.

Joyce, J. 1999. *Foundations of Causal Decision Theory*. Cambridge: CUP.

———. 2012. Regret and instability in Causal Decision Theory. *Synthese* 187: 123-45.

———. 2018. Deliberation and stability in Newcomb problems and pseudo-Newcomb problems. In Ahmed, A. (ed.), *Newcomb's Problem*. Cambridge: CUP: 138-59.

Levinstein, B, and N. Soares. Forthcoming. Cheating death in Damascus. *Journal of Philosophy*.

Lewis, D. K. 1981. Causal Decision Theory. *Australasian Journal of Philosophy* 59: 5-30. Reprinted in his *Philosophical Papers Vol. II*. Oxford: OUP 1986: 305-339.

Nozick, R. 1969. Newcomb's problem and two principles of choice. In N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*. Dordrecht: D. Reidel: 114-46. Reprinted in P. Moser (ed.), *Rationality in Action: Contemporary Approaches*. Cambridge: CUP 1990: 207-34.

Richter, R. 1984. Rationality revisited. *Australasian Journal of Philosophy* 62: 392-403.

Skyrms, B. 1980. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven: Yale UP.

———. 1990. *Dynamics of Rational Deliberation*. Cambridge, MA: Harvard UP.

Sobel, J. H. 1989. Partition theorems for causal decision theories. *Philosophy of Science* 56: 71-93.

Spencer, J. Forthcoming. An argument against Causal Decision Theory. *Analysis*.

——— and I. Wells. 2019. Why take both boxes? *Philosophy and Phenomenological Research* 99: 27-48.

Von Neumann, J. and O. Morgenstern. 1953. *Theory of Games and Economic Behaviour*. Third ed. Princeton: Princeton UP.