

Causation and Decision

Arif Ahmed, University of Cambridge

1. (a) *EDT and CDT*. This paper is a comment upon what is called *evidential* decision theory (EDT). Evidential decision theory tells us how to choose between options in situations of measurable uncertainty. Its main rival—causal decision theory (CDT)—does the same thing in a different way and the best way to introduce both theories is by contrasting them.

Evidential decision theory says that your evaluation of a contemplated option should reflect its value to you as news. Hence you should when faced with a choice between options take the most *auspicious* one. Causal decision theory says that your evaluation of the option should reflect the value to you of its effects (including itself). Hence you should when faced with a choice between options take the most *efficacious* one.

More formally and to simplify somewhat: let the variable A take the value 0 or 1 depending on which option actually occurs. Let the variable N take the value 0 or 1 depending on the obtaining or otherwise of some possible state of nature. Let your interests be restricted solely to the values taken by A and N; and let your utility function U reflect this by being a function from specifications of values of A = i and N = j to a number U_{ij}. Let ‘Cr’ denote your subjective probability distribution or credence over possible outcomes. Let ‘E’ denote your relevant background knowledge. And let the subjunctive conditional—here denoted by ‘>’—be *causally* interpreted i.e. so that any difference between the truth values of ‘If X1 were to be the case then Y would be the case’ and ‘If X2 were to be the case then Y would be the case’ reflects a difference in the efficacy of the events described by X1 and X2 to bring about that described by Y. Then EDT and CDT respectively assign to an option A = i the expected utility:

$$(1) V_E(A = i) = \sum_j Cr(N = j / A = i, E) U_{ij}$$

$$(2) V_C(A = i) = \sum_j Cr(A = i > N = j / E) U_{ij}$$

And each theory rules out every option except for those that maximize expected utility according to its proprietary measure of that quantity.

To get an impression of how these theories work let us suppose that one morning I have to decide whether to walk or drive to work. I enjoy walking more than driving but it is very important to me that I not miss the meeting at 9am. Walking to work this morning will probably cause me to miss the meeting. Driving there will almost certainly get me there on time. So driving is more auspicious of my being on time than walking. It is also more efficacious than walking in bringing about that desired result. So both EDT and CDT will recommend that I drive this morning.

In more formal terms let A = 0 and A = 1 represent the option of walking to work and driving respectively. Let N = 0 and N = 1 represent the events that I miss and that I make the meeting at 9am this morning respectively. I prefer walking to driving but I really don’t want to miss the meeting and the following utility function reflects this:

	N = 0: miss meeting	N = 1: make meeting
A = 0: walk to work	U ₀₀ = 1	U ₀₁ = 4
A = 1: drive to work	U ₁₀ = 0	U ₁₁ = 3

Table 1

I know that the walk will take much longer than the drive and in consequence my relevant credences are:

- (3) $\text{Cr}(N = 0 / A = 0, E) = 80\%$
- (4) $\text{Cr}(N = 1 / A = 0, E) = 20\%$
- (5) $\text{Cr}(N = 0 / A = 1, E) = 10\%$
- (6) $\text{Cr}(N = 1 / A = 1, E) = 90\%$
- (7) $\text{Cr}(A = 0 > N = 0 / E) = 80\%$
- (8) $\text{Cr}(A = 0 > N = 1 / E) = 20\%$
- (9) $\text{Cr}(A = 1 > N = 0 / E) = 10\%$
- (10) $\text{Cr}(A = 1 > N = 1 / E) = 90\%$

We may calculate $V_E(A = 0)$ and $V_E(A = 1)$ from (1), (3)-(6) and Table 1 as follows:

- (11) $V_E(A = 0) = 80\%.1 + 20\%.4 = 1.6$
- (12) $V_E(A = 1) = 10\%.0 + 90\%.3 = 2.7$

We may calculate $V_C(A = 0)$ and $V_C(A = 1)$ from (2), (7)-(10) and Table 1 as follows:

- (13) $V_C(A = 0) = 80\%.1 + 20\%.4 = 1.6$
- (14) $V_C(A = 1) = 10\%.0 + 90\%.3 = 2.7$

From (11) and (12) it follows that EDT recommends driving this morning. From (13) and (14) it follows that CDT recommends the same.

In that introductory example we saw that EDT and CDT make the same recommendation. The same thing is true of most decision problems that most people actually ever face or can readily imagine. Why then should we care about distinguishing them? One philosophical (rather than practical) reason for caring is that CDT provokes a question that EDT does not viz: why should it be the *causal* efficacy of an option that determines its value? No binary relation is more poorly understood than the causal one that I am here expressing by '>'; so it would be intolerable mystery-mongering to say without giving any further explanation that we should decide how to act on the basis of its extension in accordance with (2). And giving further explanation inevitably involves appeal to further normative principles of decision theory; the latter therefore occupy a more fundamental place in that theory than CDT itself. EDT faces no such objection. The evidential relation upon which it relies is transparent by comparison with the causal one; and what is mysterious or otherwise objectionable about saying that you should prefer those options that good fortune most typically accompanies—which is just what EDT in its crudest possible formulation *does* say?

I don't expect that point to convince that majority of philosophers and also of philosophers who are decision theorists that EDT is as I think the only game in town. But I do hope that they motivate an interest in defending it against objections that exploit a further and more obvious difference between it and CDT. I turn now to one of these.

(b) *An objection to EDT.* A variety of such objections fall under the label of ‘Newcomb’s Problem’. Of these the most interesting and powerful are not the best known: they are what are called ‘Medical Newcomb Problems’. In the standard example you have to decide between not smoking ($A = 0$) and smoking ($A = 1$). A tendency to smoke is known to be associated with a genetic predisposition to cancer: either you already lack that predisposition ($N = 0$) or you already have it ($N = 1$). You know that you’d enjoy smoking whether or not you have it, but you really want not to have the genetic predisposition. The following utilities reflect these facts:

	N = 0: Gene absent	N = 1: Gene present
A = 0: No smoking	U00 = 3	U01 = 0
A = 1: Smoking	U10 = 4	U11 = 1

Table 2

You are also aware of the association between smoking and the presence of the gene; you also know that since the gene is either already present or already absent it *makes* no difference whether or not you smoke. The following credences reflect your awareness of the association (these figures are no more intended to be plausible than is anything else in the example of which they are details):

- (15) $\text{Cr}(N = 0 / A = 0, E) = 90\%$
- (16) $\text{Cr}(N = 1 / A = 0, E) = 10\%$
- (17) $\text{Cr}(N = 0 / A = 1, E) = 30\%$
- (18) $\text{Cr}(N = 1 / A = 1, E) = 70\%$

The following credences reflect your awareness that smoking makes no difference:

- (19) $\text{Cr}(A = 0 > N = 0 / E) = 80\%$
- (20) $\text{Cr}(A = 0 > N = 1 / E) = 20\%$
- (21) $\text{Cr}(A = 1 > N = 0 / E) = 80\%$
- (22) $\text{Cr}(A = 1 > N = 1 / E) = 20\%$

From (1), (15)-(18) and Table 2 we derive:

- (23) $V_E(A = 0) = 90\%.3 + 10\%.0 = 2.7$
- (24) $V_E(A = 1) = 30\%.4 + 70\%.1 = 1.9$

And from (2), (19)-(22) and Table 2 we derive:

- (25) $V_C(A = 0) = 80\%.3 + 20\%.0 = 2.4$
- (26) $V_C(A = 1) = 80\%.4 + 20\%.1 = 3.4$

From (23) and (24) it follows that EDT recommends $A = 0$ i.e. your *not* smoking. From (25) and (26) it follows that CDT recommends $A = 1$ i.e. your *smoking*.

So EDT and CDT actually give different verdicts in this case. And the difference appears to count against EDT. In the example it is surely true that you are better off smoking precisely because smoking *makes* no difference to whether or not you have a gene. EDT wrongly takes the inauspiciousness of even inefficacious

acts—as e.g. reflected in (15)-(18)—to count against them; and that is why its recommendations will be wrong when they diverge from CDT’s.

(c) *Price’s defence of EDT*. Huw Price and others have sought to defend EDT on the grounds that its recommendations in cases like this do *not* in fact diverge from CDT’s.¹ Price’s argument—on which I shall focus—is best introduced by application to examples like the one that has just been occupying us.

According to Price the impression that EDT recommends not smoking arises from a mistaken application of (1) to the credences at (15)-(18). If you are deliberating over whether or not to smoke you ought not to have had those credences in the first place. This is because they violate a principle that he states² as follows:

(27) In making a probabilistic judgment take into account all the relevant available evidence.

To see how this principle rules out your having the credences (15)-(18)—which are what Price means by ‘probabilistic judgments’—consider that when you are choosing between not smoking and smoking your *total* evidence ought to include those utilities and credences that jointly—and by the definition of rational choice—*produce* whichever option is actual³. And this means that if you adhere to EDT then your evidence ought to include the fact that your utilities are as specified in Table 2 and (crucially) the fact that your credences or probabilistic judgments are as specified at (15)-(18)⁴. But clearly your credences (15)-(18) are *not* formed on a basis that takes that evidence into account because they are not and couldn’t be formed on a basis that includes *themselves*: that is, (15)-(18) don’t belong either individually or jointly to the evidential set E that each of them mentions. So an enlarged evidential basis that *does* recognize (15)-(18) in addition to everything in E is distinct from E: call it E*. But if your evidence is E* then your evidence makes smoking evidentially irrelevant to the presence of the gene: *anyone* whose credences are as described in (15)-(18) will refrain from smoking, whether or not he has the gene. So if formed on the basis of *all* the relevant evidence one’s credences will not be as in (15)-(18) but rather as follows for some *x* between 0 and 1:

- (28) $\text{Cr}(N = 0 / A = 0, E^*) = x$
 (29) $\text{Cr}(N = 1 / A = 0, E^*) = 1 - x$
 (30) $\text{Cr}(N = 0 / A = 1, E^*) = x$
 (31) $\text{Cr}(N = 1 / A = 1, E^*) = 1 - x$

¹ Price 1986: 198-202, 1991; Price and Weslake 2009: 29; Eells 1981; Horgan 1981

² 1986: 199

³ It has been objected (McKay 2007: 396-7) that in most real cases one’s motivational state does *not* screen off one’s acts from prior states because one frequently acts for reasons (or rather from causes) that are opaque to the deliberator and do not operate via the deliberative process. No doubt that is true: but to the extent that such acts are so affected they are not the proper object of decision theory. Decision theory is concerned with which acts one should perform *as a result of rational deliberation*: if in fact no actual bodily movements (e.g.) could ever be such acts then this does not refute Price’s version of that theory but only imply that it ought to take the contemplated to be, not e.g. bodily movements, but something else instead: say, the raising or lowering of the *chance* of some such movement, to whatever extent rational deliberation is able by itself to accomplish *that*.

⁴ The qualification ‘if you adhere to EDT’ is necessary because otherwise your conditional credences (15)-(18) may be both causally and evidentially irrelevant to your choice; this would be so if, for instance, you both adhered to CDT and knew that you did.

Clearly the credences (15)-(18) are *not* consistent with (28)-(31): for the right hand sides of (15) and (17) are distinct as are those of (16) and (18); whereas the right hand sides of (28) and (30) are identical as are those of (29) and (31).⁵

It follows that the anti-smoking credences (15)-(18) are unstable under reflection upon their own existence, a piece of evidence that is always available to a rational agent who is wondering what *he* should now do and always relevant to one who knows of his own adherence to EDT. It follows that those credences violate (27).

By contrast any credences that instantiate (28)-(31) are both *pro-smoking and* consistent with (27). To see the first point note that from (1), (28), (29) and the utilities in Table 2 we have:

$$(32) \quad V_E(A = 0) = 3x + 0. (1 - x) = 3x$$

—whereas from (1), (30), (31) and the utilities in Table 2 we have:

$$(33) \quad V_E(A = 1) = 4x + 1. (1 - x) = 1 + 3x$$

And so EDT recommends $A = 1$ i.e. smoking for any value of x . To see the second point note that if an agent's credences are instances of (28)-(31) and hence (by the first point) pro-smoking, then they will be pro-smoking *whether or not* the agent has got the gene. Hence an enlarged evidential basis E^{**} that took (28)-(31) into account would *still* make smoking evidentially irrelevant to possession of the gene; hence it too would induce credences that instantiated (28)-(31) and which were therefore themselves therefore pro-smoking. So a set of credences C that instantiate (28)-(31) always make the same recommendation as a set of credences C^* based upon an enlarged evidential basis that incorporates C itself. Hence the additional evidence of C 's existence is irrelevant in the sense of not affecting what you should do; so C itself does not violate (27) after all.

Price sometimes writes as though there was some temporally extended process within the decision maker corresponding to the foregoing argument against (15)-(18): anyone who *starts* with those credences will or should on reflection *end up* with ones that instantiate (28)-(31).⁶ But the expectation of such a process would undermine your assumption, crucial to rejecting (15)-(18) in the first place, that any EDT-adherent whose credences are as they describe will refrain from smoking. So somebody who thought that he should go through such a process of reflective revision would thereby lose all reason to do so.⁷

So the point about the kind of reflective equilibrium that Price is here demanding should not be that thinkers will or ought to adjust their beliefs in an attempt to reach it; it is rather that you are *already* in such a (pro-smoking)

⁵ In order for (30) and (31) to make sense it may also be necessary to assume—like other defences of EDT—the ‘trembling hand’ condition that even somebody whose credences and utilities definitely favour not smoking *might* ‘slip up’ and choose cornflakes instead of eggs. Otherwise we should have $\Pr(A = 0 / E^*) = 0$ in which case $\Pr(A = 0. E^*) = 0$ and the left hand side of (30) and (31) are meaningless; we must also assume that the occurrence of such a slip is evidentially irrelevant to the presence of the gene. For an argument that the trembling hand condition is *not* necessary see Joyce 1999: 201-14.

⁶ For instance: ‘[Your] powerful desire to avoid cancer thus turns out to play no part in [your] final [EDT]-guided choice... However, it is crucial earlier, in de-stabilising any judgments on [your] part of a probabilistic relevance of [your] action on [your] prospects of cancer’ (1986: 201).

⁷ Note that position too is unstable: see my 2005.

equilibrium if your credences satisfy (27); and if they do not then your consequent and mistaken abstention from cigarettes must be blamed upon that fact and not in any case upon your adherence to EDT. Thus it is that Price and others have sought to defend EDT from objections based upon what I have called Medical Newcomb Problems.

2. (a) *Problems for this account.* I want now to present two examples that appear to raise a difficulty for those whose position on EDT is the one that I have just described. Both examples raise the same difficulty. They are both cases of decision problems in which *no* set of credences satisfies (27). Hence they are both cases of which somebody who sought to defend EDT in Price’s way must say that EDT does not apply to them at all. But the inapplicability of EDT to these cases is grounds for objecting to EDT.

First example. You are playing a variant of ‘matching pennies’ with an evil demon; we represent the demon’s choice as the state of nature. The payoff matrix looks like this:

	N = 0 (Demon heads)	N = 1 (Demon tails)
A = 0 (You heads)	-1	1
A = 1 (You tails)	1	0

Table 3

This game resembles the standard version of ‘matching pennies’ in that one player (you) has an incentive for the pennies not to match. We may extend the resemblance by imagining that the other player (nature / the demon) has an incentive for the pennies to match⁸. But it differs two respects. First: your payoff in case both pennies show heads is much lower than your payoff in the case that they both show tails (as you can see from the top left and bottom right entries of table 3). Second: you and the demon simultaneously make your choices by raising your left hand (for heads) or your right hand for tails (in the manner of scissors, paper, stone). But the demon has telepathic powers: he can observe your credences as surely as you can. And if the credences that he observes are in favour of your choosing heads or indifferent between heads and tails (according to EDT i.e. when inputted together with table 3 into (1)), he will very probably choose heads; if it is in favour of your choosing tails then he will very probably choose tails.

Let us ask whether your credences could conform to (27). Consider first some arbitrary set of credences that together with table 3 recommends that you choose *tails*. The following will do:

$$(34) \quad \text{Cr}(N = 0 / A = 0, E) = 50\%$$

$$(35) \quad \text{Cr}(N = 1 / A = 0, E) = 50\%$$

$$(36) \quad \text{Cr}(N = 0 / A = 1, E) = 50\%$$

$$(37) \quad \text{Cr}(N = 1 / A = 1, E) = 50\%$$

First check that (34)-(37) recommend heads. From (1), (34), (35) and table 3 we have:

$$(38) \quad V_E(A = 0) = 50\% \cdot (-1) + 50\% \cdot 1 = 0$$

⁸ For instance let his utilities be yours with reversed signs.

And from (1), (36), (37) and table 3 we have:

$$(39) \quad V_E(A = 1) = 50\% \cdot 1 + 50\% \cdot 0 = 0.5$$

So the credences (34)-(37) together with the utilities in table 3 recommend the option $A = 1$ i.e. tails. Now clearly the evidential basis E that (34)-(37) each mention does not take those credences themselves into account. The question is whether or not this convicts (34)-(37) of violating (27). In particular: is the omission of (34)-(37) themselves from the E that they all mention a failure to consider evidence that is *relevant*?

Given that in conjunction with table 3 they recommend tails to an evidential decision theorist the answer must be yes. For the enlarged evidential basis E^* —got by adding those credences to E —implies that the demon, who can also see your credences, will choose tails too. But since that choice of his is based upon your credential state itself, given which it is both causally and evidentially independent of what you then do, E^* must make your act and the demon's evidentially independent of one another. So the credences that are rational upon the basis of E^* will be something like this:

$$(40) \quad Cr(N = 0 / A = 0, E^*) = 10\%$$

$$(41) \quad Cr(N = 1 / A = 0, E^*) = 90\%$$

$$(42) \quad Cr(N = 0 / A = 1, E^*) = 10\%$$

$$(43) \quad Cr(N = 1 / A = 1, E^*) = 90\%$$

The figure of 90% on the right hand side of (41) and (43) reflects the fact that if the demon sees credences that reflect a preference for tails—as do (34)-(37)—he will very probably choose tails. (The figure of 10% on the right hand side of (40) and (42) is simply a consequence of this.) Now from (1), (40), (41) and table 3 we can derive:

$$(44) \quad V_E(A = 0) = 10\% \cdot (-1) + 90\% \cdot 1 = 0.8$$

And from (1), (42) and (43) we can derive:

$$(45) \quad V_E(A = 1) = 10\% \cdot 1 + 90\% \cdot 0 = 0.1$$

So the credences (40)-(43) in conjunction with table 3 recommend $A = 0$ i.e. that you choose *heads*. Now recall from (38) and (39) that this recommendation *differs* from that of the credences (34)-(37). It follows that the absence of (34)-(37) themselves from their own evidential basis E is a violation of (27): the additional evidence that their existence constitutes is *relevant* in the relevant sense i.e. it issues in different advice over *what to do*. And the argument relied upon no feature of (34)-(37) other than the fact that they in conjunction with table 3 induce a recommendation that you choose tails. So the same applies to any coherent set of credences that makes the same recommendation. *All* such sets violate (27).

But the same also applies to any set of credences that makes the *opposite* recommendation (i.e. heads) or is *indifferent*. To see this we need only consider one such set: that consisting of (40)-(43) will do. We have already seen from (44) and (45) that it recommends heads. Now clearly the evidential basis E^* that (40)-(43) each mention does not take those credences themselves into account. The question—

again—is whether or not this convicts (40)-(43) of violating (27). Again and in particular: is the omission of (40)-(43) themselves from the E that they all mention a failure to consider evidence that is *relevant*?

Given that in conjunction with table 3 they recommend heads to an evidential decision theorist the answer must be yes. For the enlarged evidential basis E**—got by adding those credences to E*—implies that the demon will choose heads too (as it also would have implied if those credences had recommended indifference). Now E** must make your act and the demon’s evidentially independent of one another for the same reason that E* does. So the credences that are rational upon the basis of E** will be something like these:

- (46) $\text{Cr}(N = 0 / A = 0, E^{**}) = 90\%$
- (47) $\text{Cr}(N = 1 / A = 0, E^{**}) = 10\%$
- (48) $\text{Cr}(N = 0 / A = 1, E^{**}) = 90\%$
- (49) $\text{Cr}(N = 1 / A = 1, E^{**}) = 10\%$

But from (1), (46), (47) and table 3 we can derive:

$$(50) \quad V_E(A = 0) = 90\% \cdot (-1) + 10\% \cdot 1 = -0.8$$

And from (1), (42) and (43) we can derive:

$$(51) \quad V_E(A = 1) = 90\% \cdot 1 + 10\% \cdot 0 = 0.9$$

So the credences (46)-(49) in conjunction with table 3 recommend $A = 1$ i.e. that you choose *tails*. But by (44) and (45) this recommendation *differs* from that of the credences (40)-(43). It follows that the absence of (40)-(43) themselves from their own evidential basis E *is* a violation of (27): the additional evidence that their existence constitutes is *relevant* in the relevant sense i.e. it issues in different advice over *what to do*. And the argument relied upon no feature of (40)-(43) other than the fact that they in conjunction with table 3 counselled either indifference or heads. So the same applies to any coherent set of credences that makes the same recommendation. *All* such sets violate (27).

So we have seen that any set of credences that in conjunction with table 3 recommends tails violates (27). And we have seen that any set of credences that in conjunction with table recommends heads or indifference violates (27). But these are all the credences that there are. It follows that there are no credences that satisfy (27): whatever you think is going to happen you will—so long as you are both rational and and adherent of EDT—have ignored evidence that is both relevant and available.

It follows that the supplementary requirement that appears to have been forced upon EDT by the medical Newcomb problems—the requirement that EDT only operate upon credences that satisfy (27)—makes EDT *inapplicable* to the present situation. Note that this is not to say that it counsels indifference. That recommendation would at least have been *implementable* (say by means of a randomizing device) if implausible. The problem is that EDT gives *no* advice that a rational adherent of it could follow: not tails, not heads and not indifference either.

Second example. Another variant of ‘Matching Pennies’: the utilities are again as in table 3 but your opponent is even more diabolical. He has *already* chosen heads or tails by raising his left or right arm out of your sight—only this time his doing so is not an *effect* but a *cause* of your relevant credences. More precisely: his nerves are

connected not to a receiver but to a transmitter; and your brain is wired up to a receiver. If your credences are such as (in conjunction with table 3 and via EDT) to favour heads or to counsel indifference then their being in that state was almost certainly *caused* by the demon's *prior* choice of heads. And if your credences are such as to favour tails then their being in *that* state was almost certainly caused by the demon's prior choice of *tails*.

Having been through a similar case we are now in a position to see more easily why in this case no set of credences satisfies (27). So I will go through it relatively quickly. Suppose that your credences are such as to favour tails e.g. those at (34)-(37). Their evidential basis E does not take this into account; an enlarged evidential basis E* that does so will support some such credences as (40)-(43) i.e. ones that—in conjunction with the utilities at table 3—favour *heads* instead of tails; and this means that the original credences (34)-(37) were formed on an evidential basis E that missed out evidence that was both available to you and relevant to the case because it affected your verdict. So the credences at (34)-(37) violate (27); and the same goes, for the same reason, for any credences that—in conjunction with the utilities specified at table 3—recommend tails.

But the same also goes for credences that so recommend heads or indifference. Let (40)-(43) be representative of such credences. Their evidential basis E* does not and could not take account of this recommendation. But an enlarged basis E** that does so will support some such credences as (46)-(49) i.e. ones that recommend tails instead. Again this means that a deliberator whose credences recommended heads or indifference between heads and tails must be ignoring the available and relevant evidence that their very existence constitutes. Putting this result together with that of the previous paragraph we arrive at the same conclusion for this second example as for the first: *all* credences upon which EDT might be expected to operate violate (27); hence again EDT itself has nothing to say in this case.

(b) *Responses to these problems.* One might raise against these examples an objection that has been raised against standard versions of Newcomb's problem. The objection notes that we are being asked to consider fantastic situations that could not really arise. So—the objection continues—the fact that EDT cannot deal with it is no special objection to that doctrine. There could not really be any demon of the sorts that I described in my examples—it might even be physically impossible—so what does it matter that EDT tells us nothing about what to do when confronted with one?

Now it is not entirely clear that the demon that I have described in examples 1 and 2 is physically impossible. Nor is it clear that *if* he is physically impossible then my invocation of him is not any sort of threat to EDT. But set these points aside. It is possible to think—if only in a more schematic way—of examples that have the same causal structures and evidential structure as my first two, that make the same point and which are physically possible.

In the case of the first example one only needs to imagine (i) that it is either true or believed by the deliberator that any combination of credences and utilities favouring heads or indifference has *some* side effect that any combination favouring tails causally excludes (perhaps a distinctive twitch); (ii) that if the values of the variable N correspond to the non-occurrence ($N = 0$) and the occurrence ($N = 1$) of that side-effect then the deliberator's utilities are as specified at table 3 without the headings 'Demon heads' and 'Demon tails'. There is nothing impossible about *that*. But the only additional feature of my first example—the feature that made it

impossible—was the demonic aspect of the side effect. This made the example more picturesque but it did not make it any more damaging to EDT.

On the assumption of determinism we may construct an even more realistic case that shares with the second example everything that is not thus ornamental. On that assumption we may divide the possible initial states of the universe into two sorts: those that cause your credences and utilities now to be such as to favour heads or indifference, and those that causes them to be such as to favour tails. Now take the value of the variable N to indicate which of those states actually obtained. If the universe was initially in a state of the first type then $N = 0$; if it was initially in a state of the second type then $N = 1$. And let us suppose that on this interpretation of ‘ N ’ one’s utilities are as specified in table 3 without the headings ‘Demon heads’ and ‘Demon tails’. There is nothing impossible about that situation. What is *unlikely* about it is the idea that anyone should have the utilities specified in table 3 on that interpretation of ‘ N ’. But whilst a decision theory should perhaps not be required to make a recommendation for every conceivable decision situation, it *is* reasonable to require that it issue a prescription for any distribution of *utilities* over the possible outcomes of a decision situation that is in other respects not at all fantastical.⁹

A further objection to the second example has found oblique expression in the writings of some philosophers (Price and Weslake 2009: 35; Nozick 1990: 232-3). It is that one cannot consider oneself as freely choosing in cases where one knows that one’s choice is determined by causal factors outside of one’s control: so we cannot regard the second example as a *decision* problem at all; hence it is neither a surprise nor an objection to it that EDT makes no recommendation in this case. But ‘one knows that one’s choice is determined by causal factors outside one’s control’ has four interpretations; as applied to the second example these are:

- (52) One knows that either $N = 0$ or $N = 1$ obtains and that one of them will cause $A = 0$ and the other will cause $A = 1$ (via the motivational state that it produces); but one does not know which of $N = 0$ and $N = 1$ obtains and one does not know which of them will cause $A = 0$ and which will cause $A = 1$;
- (53) As in (52), except that one knows which of $N = 0$ and $N = 1$ obtains (but not which causes which act);
- (54) As in (52), except that one knows which of $N = 0$ and $N = 1$ causes which act (but not which one obtains);
- (55) As in (52), except that one knows which of $N = 0$ and $N = 1$ obtains *and* which one causes which act (and so also what one will do).

Of these perhaps (55) is incompatible with deliberation: deliberation might seem pointless if one already knows what one is going to do—at least until one forgets. But

⁹ Note also that the assumption of determinism is not really necessary. As long as one can divide the (epistemically) possible initial states of the universe into two classes A and B such that class A was much more likely than class B given present indifference or favouring of heads; and that class B was much more likely than class A given present favouring of tails, then it is possible to construct a table of utilities suitable for making the same point.

I just don't see what makes (52)-(54) being incompatible with deliberation¹⁰; and (54) describes my second example.

(c) *A different response.* The solution that I want to propose is meant to apply to both examples in virtue of what their structures have in common. So I shall start by specifying what that is.

It is usual when choosing between options to suppose that the variables characterizing them *screens off* every other state of the world that is of present interest from one's present credences and utilities. In other words: when facing a decision those 'inner' states have no evidential bearing upon any outer state of interest except *via* the deliberate acts that they are decisive in bringing about. More formally: let E represent some evidential basis that does *not* include any information about a deliberator's credences and utilities. Let E* represent the evidential basis got by adding to E certain information that exclusively concerns those credences and utilities (for instance and following our the examples, this information might specify the act that those credences and utilities favour when fed into EDT). Let the variable A take the value 0 or 1 depending on which act the deliberator actually chooses. Let the variable N take the value 0 or 1 depending on whether or not a certain external event that happens to be of interest to the deliberator actually occurs. Then that deliberator makes the supposition that I have just described only if his credences satisfy, for any i, j:

$$(56) \quad \text{Cr}(N = i / A = j, E) = \text{Cr}(N = i / A = j, E^*)$$

Put simply (56) says that acts screen off the inner from the outer.

A common feature of the first and second examples that is also the source of the trouble that they cause EDT is their violation of (56). Informally what is happening in both examples is that your credences are evidentially relevant to the state of nature—the demon's act—: in the first case by being its cause and in the second case by being its effect. We can see the formal violation of (56) itself by comparing e.g. (34) with (40) or (43) with (49).

It is also easy to see that this is the source of the trouble if we reflect that no deliberator whose credences conformed to (56)—call him a *regular* deliberator—could ever face it. For it follows from (56) that a regular deliberator who ignored the evidence of his own credences themselves would never be ignoring *relevant* evidence. That evidence would be irrelevant because—and this is what (56) assures us—adding it to his evidential basis would never affect any of the credences that *via* were material *via* (1) to the recommendation of one act over another.

Now consider the first example. When I described it I simply *stipulated* that you were choosing between the option of heads and the option of tails. But with what right did I stipulate that *these* were the options? Was it based upon some theoretical basis for dividing acts from states of nature?—no: it was simply the intuitive thing to say about the causal set-up that I described when I introduced the first example. But given the trouble that it caused I suggest that we say about it not what intuition recommends but what conforms to (56).

¹⁰ Except for general arguments in favour of incompatibilism that are beyond the scope of this paper; in any case these have no force against somebody who is content with a version of example 2 that is compatible with one's credences and utilities arising by chance: see fn. 9 above.

We can make the example conform to (56) if we redescribe the options that you have in it. Consider first an extreme version of the first example in which you are *certain* that the demon will pick heads (tails) if your credences favour heads (tails). Then instead of saying that the options in this situation are heads and tails let us say that one option ($A = 0$) is: you signaling heads *and* the demon signaling heads; and the other option ($A = 1$) is: you signaling tails *and* the demon signaling tails.

It is clear that on this conception of it the decision problem in the first example conforms to (56). For if the demon's signal is part of your act—as though his body were in this respect merely an extension of yours—then your act *does* screen off your credences and utilities from your state of nature. Thus suppose that your credences start out—as on this interpretation they must—as follows:

- (57) $\text{Cr}(N = 0 / A = 0, E) = 100\%$
- (58) $\text{Cr}(N = 1 / A = 0, E) = 0\%$
- (59) $\text{Cr}(N = 0 / A = 1, E) = 100\%$
- (60) $\text{Cr}(N = 1 / A = 1, E) = 0\%$

Clearly these credences in conjunction with table 3 favour the option of: you signaling tails *and* the demon signaling tails; suppose now that adding this information to E gives an enlarged evidential basis E^* . E^* does not change your expectation of the demon's signal. It therefore supports credences that *still* favour the option favoured of (57)-(60); in particular they are:

- (61) $\text{Cr}(N = 0 / A = 0, E^*) = 100\%$
- (62) $\text{Cr}(N = 1 / A = 0, E^*) = 0\%$
- (63) $\text{Cr}(N = 0 / A = 1, E^*) = 100\%$
- (64) $\text{Cr}(N = 1 / A = 1, E^*) = 0\%$

That is: they are the same as before and we have conformity to (56). This also shows that the problem that the first example generated for EDT no longer arises. EDT gives a clear verdict in this case that does not undermine itself: that verdict is tails.

On a less extreme version of the first example your options are not so simple. You are not choosing between *definite* combinations of your signal with the demon's; rather, you are choosing between combinations of your signal with a *lottery* over the demon's signal. For instance: suppose you are 90% confident that the demon will signal heads (tails) if your credences and utilities favour heads (tails). Then your options are: *either* signaling tails whilst taking a gamble on the demon's signal in which you are 90% sure that it will be tails; *or* signaling heads whilst taking another gamble on the demon's signal: one in which you are 90% sure that he will signal heads. Again given these credences and the utilities in table 3, EDT unambiguously favours tails; this is both in line with intuition and impervious to further reflection upon the credences themselves.

The difference in the causal structure of the second example makes no difference to the present proposal as to how EDT should treat it. It will suffice to consider an extreme version of it according to which the value of N *determines* your credences. Conformity to (56) requires that your options are not heads or tails but rather as follows. *Either*: you choose heads and it is the case that the demon has already signaled heads; *or*: you choose tails and it is the case that the demon has already signaled tails. On this conception of the options EDT again favours tails unambiguously; and again this is both in line with intuition and impervious to

reflection upon the credences that forced that recommendation. And we can treat less extreme versions of the second example in a way that is precisely analogous to what I suggested for the first example. I will not give details here.

Instead I should like to conclude by commenting upon the metaphysical significance of the second example. It is quite natural to suppose that the objects of human deliberation inevitably conform to the following temporal asymmetry: we have options over the future but not options over the past. And this supposition finds expression in a *branching* picture of one's life and its modalities: at points in one's life when one chose it was between options that entailed different futures but a common past.

If what I have said is correct then that is all wrong. Decision theory must allow options to exist over the past as well as over the future; and the only reason that we do not usually consider such options (for instance options concerning the initial state of the universe) is that we are not aware of them under descriptions that have any practical interest.