Some Counterexamples to Causal Decision Theory[1]

Andy Egan
Australian National University/University of Michigan

Introduction

Many philosophers have been converted to causal decision theory by something like the following line of argument: Evidential decision theory endorses irrational courses of action in a range of examples, and endorses "an irrational policy of managing the news".[2] These are fatal problems for evidential decision theory. Causal decision theory delivers the right results in the troublesome examples, and does not endorse this kind of irrational news-managing. So we should give up evidential decision theory, and be causal decision theorists instead.

Unfortunately, causal decision theory has its own family of problematic examples for which it endorses irrational courses of action, and its own irrational policy that it is committed to endorsing. These are, I think, fatal problems for causal decision theory.

1. The Case against Evidential Decision Theory

Evidential decision theory says that the action that it's rational to perform is the one (ignoring the possibility of ties) with the greatest expected utility – the one such that your expectations for how well things will turn out, conditional on your performing it, are greater than the expectations conditional on performing any other action. So the action that it's rational to perform will also be the one that you (or a friend with your interests in

[1] I am particularly grateful to David Braddon-Mitchell for the series of conversations that led to this paper, and for many further conversations as it was in progress. Thanks also to Adam Elga, Ned Hall, James Pryor, Brian Weatherson, Karen Bennett, Daniel Stoljar, James Joyce, Vann McGee, Allan Gibbard, Alan Hajek, Anil Gupta, and audiences at MIT, the University of Pittsburgh, Oxford University, Bowling Green State University, and the ANU Philosophical Society for very helpful questions, comments, suggestions, and objections.

[2] Lewis 1981.

mind, and with the same ideas about where your interests lie as you have) would be happiest to learn that you had performed. The case against evidential decision theory is based upon examples like the following:

The Smoking Lesion

> Susan is debating whether or not to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause – a condition that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and prefers smoking with cancer to not smoking with cancer. Should Susan smoke? Is seems clear that she should. (Set aside your theoretical commitments and put yourself in Susan's situation. Would *you* smoke? Would you take yourself to be irrational for doing so?)[3]

Evidential decision theory (at least in its initial form) wrongly condemns smoking as irrational, and endorses refraining as rational, in *The Smoking Lesion*. Causal decision theory distinguishes itself from evidential decision theory by delivering the right results in this case, and others like it. The difference between the two theories is in how they compute the relative value of actions. Roughly: evidential decision theory says to do the

[3] This example is a standard medical Newcomb problem, representative of the many to be found in the literature. The original Newcomb's problem is from Nozick 1969. For some excellent discussions of medical (and other) Newcomb problems, see (among many others) Gibbard and Harper 1976, Eells 1982, Lewis 1979 and Lewis 1981.

thing you'd be happiest to learn that you'd done, and causal decision theory says to do the thing most likely to bring about good results.

Evidential decision theory tells Susan not to smoke, roughly because it treats the fact that her smoking is evidence that she has the lesion, and therefore is evidence that she is likely to get cancer, as a reason not to smoke. Causal decision theory tells her to smoke, roughly because it does not treat this sort of common-cause based evidential connection between an action and a bad outcome as a reason not to perform the action. Let's look at how the differences between the formal theories deliver these results:

Following Lewis, let a *dependency hypothesis* be a proposition which is maximally specific about how things that the agent cares about depend causally on what the agent does. Also following Lewis, let us think of such propositions as long conjunctions of subjunctive conditionals (of the appropriate, non-backtracking kind) of the form, *if I were to do A, then P*. (Written, from now on, "$A\Box\!\!\rightarrow P$".)

The difference between causal and evidential decision theory is that causal decision theory privileges the agent's unconditional assignment of credences to dependency hypotheses in determining the relative values of actions.

If the H's form a partition of the worlds that the agent assigns non-zero credence, the value assigned to an action A by evidential decision theory (henceforth EDT) is given by:

$$VAL_{EDT} = \Sigma_H c(H|A)v(HA)$$

(Note a harmless ambiguity: I'm using 'A' to name both an action and the proposition that the agent performs that action.)

In particular, in the case of the partition of dependency hypotheses (let these be the Ks), the value assigned by EDT is given by:

$$VAL_{EDT} = \Sigma_K c(K|A)v(KA)$$

The important thing to notice about this formula is that it's the agent's *conditional* credences in dependency hypotheses that figure in it.

The value assigned by causal decision theory (henceforth CDT) is given by:

$$VAL_{CDT} = \Sigma_K c(K)v(KA)$$

The crucial difference is that now the assignments of values to actions are sensitive only to the agent's *unconditional* credences in dependency hypotheses, not her credences conditional on her performing A. The effect of this is to hold fixed the agent's beliefs about the causal structure of the world, and force us to use the same beliefs about the causal order of things in determining the choiceworthiness of each candidate action. Rather than the expected payoffs of smoking being determined by reference to how Susan thinks the causal structure of the world is likely to be, *conditional on her smoking*, and the expected payoffs of not smoking determined by reference to how she thinks the causal structure of the world is likely to be, *conditional on her not smoking*, the expected payoffs of both smoking and not smoking are determined by reference to Susan's *unconditional* beliefs about how the causal structure of the world is likely to be.

Cases like *The Smoking Lesion* motivate the move from EDT to CDT. In *The Smoking Lesion¸* there is a strong correlation between smoking and getting cancer, despite the fact that smoking has no tendency to *cause* cancer, due to the fact that smoking and cancer have a common cause. Still, since Susan's c(CANCER|SMOKE) is much higher than her c(CANCER|NOT SMOKE), EDT assigns not smoking a higher value than smoking. And this seems wrong.

So we have an argument against EDT: The correct theory of rational decision won't endorse any irrational actions or policies. In *The Smoking Lesion*, EDT endorses an irrational course of action: it's irrational for Susan not to smoke, and EDT endorses not smoking. EDT also endorses an irrational policy: it endorses a policy of performing the action with the greatest *evidential* value, rather than the action with the best expected causal upshot. So EDT isn't the correct theory of rational decision.

CDT, on the other hand, uses the agent's *unconditional* credences in dependency hypotheses to assign values to actions. The effect of this is to make our assignments of values to actions blind to the sort of common-cause correlations that make EDT's value assignments in *The Smoking Lesion* go bad.

Causal decision theory now looks very attractive. It gets the cases that made trouble for EDT right, and it seems to get them right for the right reasons – by assigning the agent's causal beliefs a special role.

3. The Case against Causal Decision Theory

Causal decision theory is supposed to be a formal way of cashing out the slogan, "do what you expect will bring about the best results". The way of implementing this sound advice is to hold fixed the agent's unconditional credences in dependency hypotheses. The resulting theory enjoins us to *do whatever has the best expected outcome, holding fixed our initial views about the likely causal structure of the world.*

The following examples show that these two principles come apart, and that where they do, causal decision theory endorses irrational courses of action.

(Obviously I think that each of the cases succeeds in showing this. But it's not important that you agree with me about both cases. For my purposes, all I need is one successful case.)

The Murder Lesion

> Mary is debating whether to shoot her rival Alfred. If she shoots and hits, things will be very good for her. If she shoots and misses, things will be very bad. (Alfred always finds out about unsuccessful assassination attempts, and he is sensitive about such things.) If she doesn't shoot, things will go on in the usual, okay-but-not-great kind of way. Though Mary is fairly confident that she will not actually shoot, she has, just to keep her options open, been preparing for this moment by honing her skills at the shooting range. Her rifle is accurate and well-maintained. In view of this, she thinks that it is very likely that, if she were to shoot, then she would hit. So far, so good. But Mary also knows that there is a certain sort of brain lesion that tends to cause both murder attempts and bad aim at the critical moment. If she has this lesion, all of her training will do her no good – her hand is almost certain to shake as she squeezes the trigger. Happily for most of us, but not so happily for Mary, most shooters have this lesion, and so most shooters miss. Should Mary shoot? (Set aside your theoretical commitments and put yourself in Mary's situation. Would *you* shoot? Would you take yourself to be irrational for not doing so?)

The Psychopath Button[4]

> 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 very strongly prefers living in a world *with* psychopaths to dying. Should Paul press the button? (Set aside your theoretical commitments and put yourself in Paul's situation. Would *you* press the button? Would you take yourself to be irrational for not doing so?)

It's irrational for Mary to shoot. It's irrational for Paul to press.[5] In general, when you are faced with a choice of two options, it's irrational to choose the one that you confidently expect will cause the worse outcome.[6] Causal decision theory endorses shooting and pressing. In general, causal decision theory endorses, in these kinds of cases, an irrational policy of performing the action which one confidently expects will cause the worse outcome. The correct theory of rational decision will not endorse irrational actions or policies. So causal decision theory is not the correct theory of rational decision.

[4] This case was suggested by David Braddon-Mitchell.

[5] Some people lack the clear intuition of irrationality for the *Murder Lesion* case. Pretty much everyone seems to have the requisite intuition for *The Psychopath Button*, however. That's enough for my purposes. Personally, I think both cases work as counterexamples to causal decision theory. But all I need is that at least one of them does.

[6] Whether it's irrational in a particular case depends, of course, on just what the payoffs are. It can be worth doing something that's more likely than not to cause a bad outcome if the low-probability good outcome is good enough. But in the cases above (and as spelled out below), it's better not to do the thing that you expect will cause the worse outcome. See below for some sample numbers.

Let's pause for a moment to address two natural thoughts to have at this stage. First, the reader is likely to have noticed the similarity between these cases and Gibbard and Harper's (1978) *Death in Damascus* case, and may be thinking thoughts about unratifiability. The cases are, however, importantly different – we have (or at least my informants and I have) clear intuitions that it's irrational to shoot or to press, and rational to refrain, in *The Murder Lesion* and *The Psychopath Button*, while we lack any such asymmetric intuitions about the *Death in Damascus* case. (The cases I'm concerned with here are much more like versions of *Death in Damascus* in which the road to Damascus is more pleasant than the road to Aleppo.) I'll discuss ratifiability-based responses in detail in section 4.

Second, it's natural to be concerned about the part of the setup of *The Murder Lesion* when I said that "she thinks that it is very likely that, if she were to shoot, then she would hit". How, one might wonder, could such confidence survive in the face of her confidence that most shooters have the lesion? Answer: Mary can have high credence that, were she to shoot, she would hit, so long as she has high credence that she does not have the lesion. And she can have high credence that she does not have the lesion so long as she does not have high credence that she will shoot. (A similar worry arises for *The Psychopath Button* – for CDT to endorse pressing, Paul must start off with a sufficiently low credence that he is a psychopath. But there's no problem with Paul having such a low credence, so long as he does not have a very high credence that he will press the button.)[7]

[7] The worry may persist. Why shouldn't we instead think that, since Mary knows that most shooters have the lesion, and so most shooters miss, she will (indeed, she *must*) think that it is very likely that, if she were to shoot, then she would miss? After all, she has no extra information about herself that indicates that she would be unusual among shooters. We can stipulate that the bulk of Mary's credence is distributed to no-

Back to the main thread, and on to some analysis of what's gone wrong. What's generating the problem here is that the very same mechanism that allows causal decision theory to deliver the *right* results in cases like *The Smoking Lesion* leads it to deliver the *wrong* results for cases like *The Murder Lesion* and *The Psychopath Button*. Let's look at what happens in *The Murder Lesion*. (The analysis of *The Psychopath Button* will be relevantly similar.) Let S be the proposition that Mary shoots, and H the proposition that Mary hits. The relevant partition of dependency hypotheses is $\{S\Box\!\!\rightarrow H, S\Box\!\!\rightarrow \neg H\}$.

Some constraints on Mary's credences:

$c(S\Box\!\!\rightarrow H) > .5$.

(Because she's been going to the shooting range, the gun is well-maintained, accurate and reliable, Alfred is a large, slow-moving target, etc.)

$c(S\Box\!\!\rightarrow H|S) < .5$

(Because if she shoots, it's very likely because she has the lesion, and if she has the lesion, she's very likely to have bad aim when push comes to shove.)[8,9]

lesion, non-shooting worlds. The question, then, is this: what happens in the shooting worlds closest to such no-lesion, non-shooting worlds?
Answer: She hits. In evaluating the sorts of non-backtracking, causal counterfactuals that are relevant to CDT, we hold Mary's past – including the presence of the lesion, her time at the shooting range, etc. – fixed. In worlds where Mary shoots, lacking the lesion, and with her actual training in marksmanship, she hits. (Better – in *almost every* such world, she hits. The no-lesion, no-shooting worlds in which, were she to shoot, she would hit, receive a much higher proportion of her credence than the no-lesion, no-shooting worlds in which something *else* peculiar is going on, such that if she were to shoot, she would miss, or the gun would explode, or…)

[8] Another reason: We know that Mary's $c(H|S) < .5$, since shooting is such good evidence for having the lesion, and her credence that she hits conditional on both shooting and having the lesion is very low. Given that, we can prove that $c(S\Box\!\!\rightarrow H|S) < .5$:
By the definition of conditional probability,
$c(S\Box\!\!\rightarrow H|S) = c(S \,\&\, S\Box\!\!\rightarrow H)/c(S)$
Since every world in which both S and $S\Box\!\!\rightarrow H$ are true is a world in which H is true as well,
$c(S \,\&\, S\Box\!\!\rightarrow H) \leq c(SH)$.

Mary's value assignments:

$v(S.H) = 10$

$v(S.\neg H) = -10$

$v(\neg S) = 0$

If Mary is a causal decision theorist, she must use $c(S\Box\!\!\rightarrow H)$, not $c(S\Box\!\!\rightarrow H|S)$, when she's determining the relative values of shooting and refraining. (Since it's unconditional credences in dependency hypotheses that feature in CDT's formula for determining the choiceworthiness of actions.) So shooting is going to come out looking better than not shooting.[10]

But that's the wrong result. It's irrational for Mary to shoot. Unfortunately, if that's right, then causal decision theory is wrong.

The same phenomenon occurs in a particularly striking way in time travel cases. Suppose that you have a time machine, and you are convinced that time travel works in the single-timeline, no-branching way outlined by Lewis (1976). You want to use your time machine to preserve some document, thought to be lost in the fire at the library of

So we know that:
$c(S\Box\!\!\rightarrow H|S) \leq c(S\&H)/c(S)$.
Again by the definition of conditional probability, $c(S\&H)/c(S) = c(H|S)$.
So $c(S\Box\!\!\rightarrow H|S) \leq c(H|S) < .5$.

[9] Note, for future reference, that $c(S)$ must be $< .5$ for these credences to be coherent.

[10] Because CDT says that Mary should determine the value of smoking by computing:
$\Sigma_K c(K)v(KA)$, which in this case gives us:
$VAL_{CDT}(S) = c(S\Box\!\!\rightarrow H)v(S\Box\!\!\rightarrow H \ \& \ S) + c(S\Box\!\!\rightarrow \neg H)v(S\Box\!\!\rightarrow \neg H \ \& \ S)$
Assuming that Mary doesn't care about dependency hypotheses for their own sakes, $v(S\Box\!\!\rightarrow H \ \& \ S) = v(S.H)$, and $v(S\Box\!\!\rightarrow \neg H \ \& \ S) = v(S.\neg H)$. (The value of shooting while in a SHOOT $\Box\!\!\rightarrow$ HIT world is the value of shooting and hitting; the value of shooting while in a SHOOT $\Box\!\!\rightarrow$ MISS world is the value of shooting and missing.) So we get:
$VAL_{CDT}(S) = c(S\Box\!\!\rightarrow H)v(S.H) + c(S\Box\!\!\rightarrow \neg H)v(S.\neg H)$
And since $c(S\Box\!\!\rightarrow H) > c(S\Box\!\!\rightarrow \neg H)$, it will turn out that $VAL_{CDT}(S) > 0$, and so $VAL_{CDT}(S) > V(\neg S)$.

Alexandria. One option is to attempt to surreptitiously spirit the document out of the library before the fire. Another is to attempt to prevent the fire from ever happening. If you don't have a firm opinion about which course you'll actually pursue, you're likely to be confident that, if you *were* to attempt to prevent the fire, you would succeed. (After all, you're competent and knowledgeable, you have many willing and able accomplices, access to excellent equipment, plenty of time to plan and train, etc.)

But you know that the fire really did happen. So you know that any attempt you make to go back and prevent it will fail.[11] It's irrational to pursue this sort of doomed plan – a plan that you already know will fail, and the failure of which you take to be worse than the expected result of some alternative plan – and so it's irrational to try to prevent the fire.[12] (Similarly, when you go back in time to set up a holding company that will, when the investments mature, pay a large lump sum into your bank account, you should arrange for the cash to be deposited in your account *after* the last time you checked your balance and saw that there hadn't been any large deposits.) But CDT doesn't deliver these results. Determining the relative choiceworthiness of actions using only your *unconditional* credences in dependency hypotheses makes your ranking of actions insensitive to your knowledge – knowledge to which your decision-making *should* be sensitive – that the past-changing plans are sure to fail.

Oracle cases are relevantly similar. It's irrational to try to avoid the fate that the (infallible) oracle predicts for you. The thing to do, faced with an unpleasant oracular prediction, is to try to ensure that the predicted fate comes about in the best possible way.

[11] There are complications. Some of these are discussed in Braddon-Mitchell and Egan (MS).

[12] Calling the plan 'doomed' is, of course, a provocative way of putting the point. But what's doing the work in making the case for the plan's irrationality isn't some objectionable fatalism – it's the fact that you confidently expect the plan to fail and, in failing, to bring about a bad outcome. (Not just a bad outcome – a worse outcome than the one you would expect some alternative plan to bring about.)

If the oracle predicts that you'll be bitten by a rabid dog, the thing to do is to get vaccinated and wear thick clothes so that the bite won't do much harm, not to poison your neighbors' dogs in hopes of avoiding the predicted bite.

(It's worth pointing out that neither the oracle nor the time-travel cases rely on absolute certainty. What's really going on is that, the more reliable you take the oracle, or your information about the past, to be, the worse an idea it is to try to avert the predicted fate, or change the apparent past.)

I include the time travel and oracle cases because (a) they provide particularly stark examples of cases where CDT endorses performing an action that one confidently expects will bring about a worse outcome than some alternative, and (b) they may serve to make clearer just what's gone wrong in the other cases. In these cases, just as in cases like *The Murder Lesion* and *The Psychopath Button*, the fact that CDT forces us to use only the agent's unconditional credences in dependency hypotheses in determining the choiceworthiness of actions makes its verdicts blind to features of the agent's beliefs to which it should be sensitive – namely, the agent's confidence that a particular course of action, if undertaken, is doomed to fail, and bring about a worse outcome than the alternative.

I don't want to put very much argumentative weight on the time travel and oracle cases, since it's not completely obvious how big a problem it is for CDT to give the intuitively wrong results in these peculiar sorts of situations – perhaps it's okay to just bite the bullet here, and say that our intuitions about these sorts of cases ought not to be taken seriously. (Either because the cases are "don't cares", where it's not important for our theory to deliver the right results, or because they're cases where we ought to allow our theory to trump our intuitive judgments about which results are the right ones). In fact, I don't think that this sort of response is very attractive, but it doesn't really matter. Even if it is okay to dismiss our intuitive judgments about time travel and oracle cases, it's not okay to dismiss our intuitive judgments about *The Murder Lesion* and *The Psychopath Button*. Or at least – and this is enough for my purposes – it's not okay for the causal decision theorist to dismiss our judgments about *The Murder Lesion* and *The Psychopath Button* if it's not okay for the evidential decision theorist to dismiss our judgments about *The Smoking Lesion*.

Here is the moral that I think we should draw from all of this: Evidential decision theory told us to perform the action with the best expected outcome. Examples like *The Smoking Lesion* show us that having the best expected outcome comes apart from having the best expected causal impact on how things are, and that rationality tracks the latter rather than the former. So, they show us that evidential decision theory is mistaken. Causal decision theory told us to perform the action which, *holding fixed our current views about the causal structure of the world*, has the best expected outcome. Examples like *The Murder Lesion* and *The Psychopath Button* show us that this too comes apart from having the best expected causal impact on how things are. So, they show us that causal decision theory is mistaken.

4. Objections, Responses, and Further Problems

There are some responses available to the causal decision theorist. Unfortunately, I don't think that any of them work. In fact, the most promising response fails in a way that shows us that the problem is actually quite a bit worse than I've suggested so far, and that advocates of evidential decision theory ought to take no comfort in the difficulties for CDT.

Are the cases too science-fictional and/or morally loaded to make good counterexamples?

One might be concerned that the cases I've used against CDT – *The Murder Lesion* and *The Psychopath Button* – are either too science-fictional or too morally loaded to make good counterexamples, perhaps because our intuitions about such cases are not to be trusted. I'm inclined to insist on the legitimacy of the cases as given. I'm particularly inclined to insist in the case of the 'too science fictional' objection, because all that's needed is a case where the subject *believes* that there are the relevant sorts of lesions, buttons, oracles, or what have you – the actual presence of the science fictional apparatus is not important. But it's not important that you agree with me about the cases as given. Once you know where to look, there are many more such cases to be found, and many of them are much less exotic, and less fraught with potentially distracting moral issues.

For example, it's easy to modify *The Smoking Lesion* in order to make it a counterexample to CDT rather than EDT. We just have to change the case in the following way: Rather than letting Susan believe that the lesion (a) causes one to smoke, and (b) causes one to get cancer, let her believe that the lesion (a) causes one to smoke, and (b) causes one's lungs to be vulnerable to cigarette smoke, such that smoking causes cancer in those with the lesion, but not in those without.

In this sort of situation, it is irrational to smoke. But CDT still endorses smoking, so long as one's initial credence that one has the lesion is sufficiently low. Further, this modified smoking lesion case is certainly not objectionably morally loaded. Nor is it objectionably science-fictional. At least, it's not objectionably science-fictional unless the original *Smoking Lesion* case is objectionably science-fictional. So as long as *The Smoking Lesion* succeeds as a counterexample to EDT, the modified smoking lesion case will succeed as a counterexample to CDT.

This is an instance of a quite general recipe for generating counterexamples to CDT: Start with a counterexample to EDT in which some condition is (believed to be) a common cause both of some action A and of some undesirable outcome O. Change the case so that, rather than directly causing O, the condition puts in place an enabling condition which allows A to cause O. Finally, point out to your audience that our intuitions about what one ought to do switch when we change the causal background in this way, while CDT's recommendations remain the same. (Note: CDT's recommendations don't stay the same in every version of the case – the agent's credences and values have to be right. In particular, the agent's unconditional credence that the troublemaking condition (having the lesion, etc.) obtains must be fairly low.)[13]

These anti-CDT examples will be no more science-fictional or morally loaded than the original anti-EDT examples we started with. If the original examples were unacceptable, then CDT is unmotivated – we don't have a counterexample to EDT. If the original examples were acceptable, then the modified examples are as well, and CDT is subject to counterexamples. Neither outcome is a good one for the advocate of CDT.

Do the cases put unacceptable constraints on the agents' credences regarding their own actions?

[13] Thanks to Martin Smith for extracting the general recipe from the cases.

Notice that, in order for CDT to endorse shooting in *The Murder Lesion,* Mary must start off confident that, if she were to shoot, she would hit. For her to be confident of this, she must also start off confident that she does not have the lesion. And so, it seems, she must start off confident that she will not shoot. Similarly, for CDT to endorse pressing in *The Psychopath Button*, Paul to start off confident that, if he were to press the button, he would live. For him to be confident of this, he must start off confident that he is not a psychopath. And so, it seems, he must start off confident that he will not press the button. So in order for my cases to work, the agents' credences about what they are likely to do must be a certain fairly specific way. Is this a problem?

No. The cases do indeed place some constraints on the agents' credences regarding their own future actions. But so too do the examples, like *The Smoking Lesion*, that motivate CDT over EDT. For those cases to succeed, the agents mustn't be certain of what they're going to choose. So if the fact that a case places *any* constraints on the agent's credences about their own future actions renders it ineligible to serve as a counterexample, then the counterexamples to EDT will be ruled out along with the counterexamples to CDT, and CDT loses its motivation..

But perhaps it's not the fact that a case places *some* constraints on the agent's credences that rules it out as a counterexample, but the fact that it places a certain, objectionable *sort* of constraint on the agent's credences, that rules it out as a counterexample. And while the counterexamples to EDT impose only innocent constraints, those imposed by the would-be counterexamples to CDT are objectionable.

I don't think that there is any plausible way to cash out the distinction between innocent and objectionable constraints that will deliver this result. Certainly neither the counterexamples to EDT nor the counterexamples to CDT require the agents to have credences that violate the constraints of Bayesian rationality. And it's unclear where else we might non-arbitrarily draw the line.

We might also be concerned that the putative counterexamples are illegitimate because they force agents to *have* credences about their own actions, and that this is unacceptable – agents don't, or ought not to, have any credences at all in propositions about which actions they will freely perform. If A is a proposition stating which action I will perform, c(A) should not be defined. (Or, alternatively, should not take any value other than 0 or 1.)[14]

But in fact, we don't ever need to appeal directly to Mary's or Paul's credences about which action they're going to perform when calculating the values that CDT assigns to the candidate actions in *The Murder Lesion* or *The Psychopath Button*. What we *do* need to appeal to is the agents' conditional credences of the form c(P|A), where A is a proposition stating which action they will perform. (For example, Mary's credence that she has the lesion conditional on her shooting, and Paul's credence that he is a psychopath conditional on his pressing.) For the cases to work, Mary's and Paul's conditional credences of this sort do need to meet certain constraints. And those constraints are enough, if we accept the standard formula for conditional probability (that is: c(B|A) = c(AB)/c(A)), to impose constraints on Mary's and Paul's credences about what they're going to do. But if we reject the standard formula for conditional probability, it's available to us to deny that the agents have any credences at all about what they will do, or to let their credences take only extreme values.

[14] See, for example, Levi 1997, Kyburg 1988, Gilboa 1994, and Spohn 1977 for views of this kind.

Proponents of the view that we cannot have well-defined credences (or cannot have well-defined credences other than 0 or 1) in propositions specifying which free actions we will perform ought *not* to deny that we can have well-defined conditional credences for various outcomes, conditional on our various possible choices. What they ought to do is deny that the standard formula for conditional credences is correct. And in fact, taking conditional credences to be primitive, or at least separating them to some extent from the standard formula, is independently well-motivated. (Price (1986), Edgington (1995), and Hajek (2003), for example, are all advocates of separating, to some extent, c(B|A) from c(AB)/c(A).)

So even if we don't want to admit well-defined credences (other than 0 and 1) for propositions about which free actions I'll perform, we can still admit well-defined conditional credences of various outcomes conditional on my various candidate actions. And it's these conditional credences, not the unconditional credences in the performance of the actions, that are actually doing the heavy lifting in the examples. If we *do* allow that Mary and Paul have well-defined unconditional credences for propositions like SHOOT and PRESS, *and* we accept the standard formula as a definition of conditional credence, then we do get some constraints on just what their credences in those propositions can coherently be. But this conditional result should be unobjectionable.

(Personally, I think that it's just fine for agents to have all kinds of (probabilistically coherent) credences about what they're going to do. The upshot of the preceding discussion is just that the outcome of this fight is irrelevant to the legitimacy of cases like *The Murder Lesion* and *The Psychopath Button* as counterexamples to causal decision theory.)

Finally, notice two things: First, we also need to appeal to such conditional credences in order to determine EDT's endorsements in the cases (like *The Smoking Lesion*) that are supposed to provide the motivation to abandon EDT in favor of CDT. So if this reliance on well defined conditional credences of outcomes on actions undermines my counterexamples to CDT, it undermines the CDTer's counterexamples to EDT as well. So this is a bad defense for the advocate of CDT to appeal to: if it succeeds, CDT is unmotivated. If it fails, CDT is subject to counterexamples.

Second, giving up even these *conditional* credences really does seem like it will lead to very serious trouble. A theory according to which we're not allowed to have any views at all about what's likely or unlikely, conditional our choosing one thing rather than another, cannot be correct. Certainly it cannot underwrite a theory of rational decision.

Can we fix everything by going ratificationist?

Consider Paul's situation as he deliberates about whether or not to press the 'kill all psychopaths' button. Suppose that Paul is an orthodox causal decision theorist. Pressing will, at the beginning of his deliberations, look better than refraining. Paul becomes convinced that pressing is the thing to do, and so he becomes convinced that he will, at the end of his deliberations, choose to press. But as Paul becomes more and more convinced that he's going to choose to press, he becomes more and more confident that he's a psychopath. And as he becomes more and more confident that he's a psychopath, pressing starts to look like less and less of a good idea. At a certain point, as Paul becomes increasingly convinced that he's going to press, CDT will stop telling him to press, and start telling him to refrain.

Pressing the psychopath button is *unratifiable* by the lights of CDT: it's impossible for Paul both to be convinced that he will press the button, and also to rationally endorse doing so. It's tempting to think that we can exploit this fact in order to save (a version of) CDT from the apparent counterexamples, by imposing a ratifiability requirement on rational actions.

Perhaps the simplest way to impose a ratifiability requirement is just to add the following *Maxim of Ratifiability*[15] to our original version of EDT:

> **Maxim of Ratifiability**. An agent can rationally perform act *A* only if *A* is *ratifiable* in the sense that there is no alternative *B* such that $\mathrm{VAL}_{\mathrm{CDT}}(\mathrm{B})$ exceeds $\mathrm{VAL}_{\mathrm{CDT}}(\mathrm{A})$ on the supposition that *A* is decided upon.

The resulting theory tells us that it's rational to perform an action A iff:

1) A is ratifiable, and

2) There is no other ratifiable option with greater (present) $\mathrm{VAL}_{\mathrm{CDT}}$ than A.

(Another way to implement a ratificationist version of CDT is just to say that it's rational to perform A iff A is ratifiable, in the sense specified in the Maxim of Ratifiability above. For our purposes here, we needn't decide which of these ratificationist theories is better, as the differences between them will not be relevant to the objections that I'll make below. The crucial feature that they have in common is that on both accounts, it's never rational to perform an unratifiable action – being unratifiable is sufficient for being ruled out as a rational option.)

[15] This statement of the maxim is lifted from Joyce (forthcoming).

A version of CDT that includes a ratifiability requirement will not endorse shooting in *The Murder Lesion*, or pressing in *The Psychopath Button*. When Mary becomes convinced that she will choose to shoot, shooting will look bad to her – $\mathrm{VAL}_{\mathrm{CDT}}$(SHOOT) will be less than $\mathrm{VAL}_{\mathrm{CDT}}$(NOT SHOOT). When Paul becomes convinced that he will choose to press, pressing will look bad to him – $\mathrm{VAL}_{\mathrm{CDT}}$(PRESS) will be less than $\mathrm{VAL}_{\mathrm{CDT}}$(NOT PRESS). So a theory that counts all unratifiable actions as irrational will not deliver the bad endorsements that we got from the version of CDT that we considered above, which did not include a ratifiability requirement. Problem solved?

Unfortunately, no. There are two reasons why this response fails. The first is that, if successful, it does too much: If an appeal to ratifiability succeeds here, then the EDTer's appeal to ratifiability in the cases that were supposed to motivate the move to CDT will succeed as well. (Not surprisingly, since the appeal to ratifiability was originally a move in defense of EDT in the face of just such examples - see Jeffrey 1983.) In *The Smoking Lesion*, not smoking is unratifiable: once Susan becomes convinced that she will *choose* not to smoke, her smoking or not ceases to be evidence one way or the other for her having the lesion, and smoking looks better, by EDT's lights, than refraining. So again, we have a situation in which, if the CDTer's defense works, it works for the EDTer as well, and CDT loses its motivation.[16]

The second difficulty with this response is that it doesn't do enough. Here are two constraints on any adequate theory of rational decision:

[16] But see Joyce (forthcoming) for an argument that the appeal to ratifiability is, in fact, *only* available to causal decision theorists.

SOUNDNESS: If it's irrational to φ, the correct theory of rational decision will not endorse φing.

COMPLETENESS: If it's rational to φ, the correct theory of rational decision will endorse φing.

While the imposition of a ratifiability requirement prevents CDT from falling afoul of the SOUNDNESS requirement, the resulting theory still fails to satisfy COMPLETENESS.

In *The Psychopath Button*, it's irrational for Paul to press. It's rational for Paul to refrain from pressing. Neither action, however, is ratifiable. (When Paul becomes convinced that he will choose to refrain, he will become quite confident that he is not a psychopath, and pressing will look better than refraining.)

It's rational for Paul to refrain. So the correct theory of rational decision will endorse refraining. Refraining is not ratifiable. So no theory that imposes a ratifiability requirement will endorse refraining. So no theory that imposes a ratifiability requirement is the correct theory of rational decision.[17]

This shows us that imposing a ratifiability requirement will not help us to save CDT. It also shows us that fans of EDT should take no comfort in the difficulties for CDT – what we have here is definitely *not* an argument for a return to evidential decision theory. These cases are all counterexamples to versions of EDT that impose ratifiability

[17] This demonstrates the important difference between cases like *The Murder Lesion* and *The Psychopath Button* and cases like Gibbard and Harper's (1978) *Death in Damascus*, in which it's also the case that neither option is ratifiable. In the cases we're concerned with, unlike in *Death in Damascus*, we still have clear intuitions about which action it's rational to perform.

requirements as well, and these seem to be the only versions of EDT with the resources to deal with cases like *The Smoking Lesion*.

So things are actually worse than I've been making them out to be – these cases are trouble not just for CDT, but also for any version of EDT with the resources to avoid refutation at the hands of common-cause based counterexamples like *The Smoking Lesion*.

In fact, there are cases where imposing a ratifiability requirement makes things *worse*, particularly for CDT. Consider the following modification of the original Newcomb's problem:

Newcomb's Firebomb

> There are two boxes before you. Box A definitely contains $1,000,000. Box B definitely contains $1,000. You have two choices: take only box A (call this *one-boxing*), or take both boxes (call this *two-boxing*). You will signal your choice by pressing one of two appropriately labeled buttons. There is, as usual, an uncannily reliable predictor on the scene. If the predictor has predicted that you will two-box, he has planted an incendiary bomb in box A, wired to the *two-box* button, so that pressing the *two-box* button will cause the bomb to detonate, burning up the $1,000,000. If the predictor has predicted that you will one-box, no bomb has been planted – nothing untoward will happen, whichever button you press. The predictor, again, is uncannily accurate.

It is, I submit, rational to one-box, and irrational to two-box, in *Newcomb's Firebomb*. (You should expect that, if you press the two-box button, you will be causing the incineration of your $1,000,000, which is certainly sitting there in Box A just waiting for you to carry it off to the bank. Crucially, it is your choice will cause its incineration – this is the key difference between *Newcomb's Firebomb* and the original Newcomb's problem.)

But neither option is ratifiable. A ratificationist theory will not endorse two-boxing, but it won't endorse one-boxing either. So if we adopt a ratificationist theory, we will be forced to say that there is no rational option in this case. And this seems wrong – one boxing is pretty clearly the rational thing to do here.

The imposition of a ratifiability requirement makes things *worse* in this case, because versions of CDT that do *not* include a ratifiability requirement deliver, on almost every way of spelling out the case, the verdict that it's rational to one-box. (The exceptions are cases in which one starts off *extremely* confident that one is going to choose one-boxing, and so starts off *extremely* confident that there is no bomb in box A.) Holding fixed any but the most extreme credences about whether or not there's a bomb in box A, we get the result that one-boxing has greater VAL_{CDT} than two-boxing. It is only in the cases where one assigns a very, very low unconditional credence to the presence of the firebomb that CDT will tell us that the possibility of gaining the extra $1,000 is worth the risk of setting fire to the $1,000,000. So CDT without a ratifiability requirement *almost* always tells us, in accordance with our intuitions about the case, that it is rational to one-box, and irrational to two-box, in *Newcomb's Firebomb*.

Ratificationist versions of CDT, however, can *never* endorse one-boxing in *Newcomb's Firebomb*. This is still more bad news, I think, for the ratificationist defense of CDT. Evidentialists cannot rejoice in this, however – the news for the ratificationist defense of EDT is equally bad, as ratificationist EDT also fails to endorse the unratifiable option of one-boxing in *Newcomb's Firebomb*.

5. An Instructive Failure

What about a fancier version of ratificationism? The trouble with COMPLETENESS was generated by the fact that standard ratificationist proposals say that being unratifiable is sufficient for being irrational – that it's *never* rational to perform an unratifiable action. We can take ratifiability to be important, though, without going quite this far. Suppose we said this:

LEXICAL RATIFICATIONISM: It is rational to perform an action A iff:

1) A is ratifiable, and there is no other ratifiable option with higher VAL_{EDT} than A, *or*

2) There are no ratifiable options, and no other (unratifiable) option has higher VAL_{EDT} than A.

This is equivalent to a view according to which actions are ordered by choiceworthiness, in the following, two-step manner. Step one: order by ratifiability – that is, if A is ratifiable and B is unratifiable, then A is to be preferred over B. Step two: within each of the two groups, order by VAL_{EDT}. This imposes a *lexical* ordering, on which ratifiable actions are always to be preferred over unratifiable ones, but within the categories, the action with greater VAL_{EDT} is to be preferred.

This seems, in fact, to deliver the right verdicts for all of the cases we've discussed so far. It endorses smoking in *The Smoking Lesion*, since smoking is ratifiable while refraining is not. It endorses not shooting in *The Murder Lesion*, since neither shooting nor not shooting is ratifiable, and not shooting has higher $\mathrm{VAL}_{\mathrm{EDT}}$. For the same reason, it endorses not pressing in *The Psychopath Button*, and one-boxing in *Newcomb's Firebomb*. In both cases, all of the agent's options are unratifiable, but not pressing and one-boxing have higher $\mathrm{VAL}_{\mathrm{EDT}}$s than their respective competitors.

(Notice that it's important that it's evidentialist value – $\mathrm{VAL}_{\mathrm{EDT}}$ – that imposes the second part of the ordering, rather than $\mathrm{VAL}_{\mathrm{CDT}}$. A theory that appealed to $\mathrm{VAL}_{\mathrm{CDT}}$ at the second stage would endorse shooting in *The Murder Lesion*, and pressing in *The Psychopath Button*, as shooting and pressing have higher $\mathrm{VAL}_{\mathrm{CDT}}$ than refraining.)

This was going to be my tentative positive proposal until Anil Gupta presented me with what I take to be a completely decisive counterexample, which he has kindly allowed me to reproduce here:

The Three-Option Smoking Lesion

> Samantha is deciding whether to smoke. But her situation is slightly more complicated than Susan's. Samantha has three options: Smoke cigars, smoke cigarettes, or refrain from smoking altogether. Call these options CIGAR, CIGARETTE, and NO SMOKE. Due to the ways that various lesions tend to be distributed, it turns out that cigar smokers tend to be worse off than they would be if they were smoking cigarettes, but better off than they would be if they refrained from smoking altogether. Similarly, cigarette smokers tend to be worse off than they would be smoking cigars, but better off than they would be refraining from smoking altogether. Finally, non-smokers tend to be best off refraining from smoking.

So: CIGAR in unratifiable, because choosing to smoke cigars is very good evidence that you'd be better off smoking cigarettes. CIGARETTE is unratifiable, because smoking cigarettes is very good evidence that you'd be better off smoking cigars. NO SMOKE, however, *is* ratifiable, because not smoking is very good evidence that you'd be best off not smoking. So LEXICAL RATIFICATIONISM endorses NO SMOKE, since it's the only ratifiable option. But this is wrong. If you find yourself deciding to smoke cigars, one thing you know for sure is that NO SMOKE is *not* the way to go. You've got good reason to think that you'd be better off smoking cigarettes, but you've got equally good reason to think that you'd be *worse* off refraining from smoking altogether.

The important structural feature of the case is this: We have three options. Option 1 is unratifiable because, conditional on choosing option 1, option 2 looks better than option 1. Option 2 is unratifiable because, conditional on choosing option 2, action 1 looks better than option 2. Option 3 is ratifiable because, conditional on choosing option 3, option 3 looks better than either 1 or 2. However, conditional on choosing either of options 1 or 2, option 3 looks very bad. In this sort of case, we can understand someone who finds herself deciding on option 1 or 2 rethinking and doing some vacillating *between options 1 and 2*. What seems clearly *irrational* is for the person who finds herself deciding on either 1 or 2 to perform action 3 on grounds of its ratifiability. If she finds herself deciding on 1 or 2, she has excellent reason to think that 3 would be the *worst* thing to choose.

I think that this kind of case is fatal for the lexical ratificationist strategy. Lexical ratificationism gets the right results in *The Murder Lesion* and *The Psychopath Button*, but it goes disastrously wrong here. More importantly, though, this kind of case is fatal for ratificationism in general. *No* ratificationist account will be able to deliver the right results in the sorts of three-option cases that Gupta has pointed out. The real importance of the Gupta cases is not that they refute *lexical* ratificationism – it's that they refute *every* form of ratificationism.

Conclusion

If all of the above is correct, causal decision theory is in a bad way. Either it's subject to counterexamples, or there's no reason to prefer it to EDT. That's what I hope to have shown above, and that is what I'm primarily concerned to emphasize in this concluding section. I will close, though, with some speculation about what's gone wrong and the how to fix it.

What conclusions should we draw from all this? I take cases like *The Smoking Lesion* to show that EDT is informed by the wrong principle of rational decision. It's informed by the principle (roughly), *do the thing which would give you the best evidence that the best things are happening.* Where the advice of this principle comes apart from that of the principle, *do what's most likely to bring about the best results*, it delivers advice that it's irrational to follow. Enter causal decision theory, which aims to give a satisfactory formal characterization of the correct, causal principle. What I take cases like *The Murder Lesion* and *The Psychopath Button* to show is that Lewisian CDT's formal characterization of the informal principle isn't satisfactory. The principle that Lewisian CDT actually endorses, *do what has the best expected outcome, holding fixed* *your current views about the causal structure of the world*, isn't quite the right way of understanding the original principle, *do what's most likely to bring about the best results*.

My hope, then, is that there will be an alternative formal theory which provides a better understanding of the appealing principle. I regret that I do not have such a theory to offer.

References:

Braddon-Mitchell, David and Egan, Andy. MS. How Ignorance Empowers Time Travelers.

Edgington, Dorothy. 1995. On Conditionals. *Mind* 104:235-329.

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

Gibbard, Allan and Harper, William. 1978. Counterfactuals and Two Kinds of Expected Utility. In *Foundations and Applications of Decision Theory*, v.1, edited by Clifford Hooker, J.J. Leach, and Edward McClennen. Dordrecht: Riedel.

Gilboa, Itzhak. 1994. Can Free Choice Be Known?, in *The Logic of Strategy*, edited by Cristina Bicchieri, Richard Jeffrey, and Brian Skyrms. Oxford: Oxford University Press.

Hajek, Alan. 2003. What Conditional Probability Could Not Be. *Synthese* 137:273-323.

Jeffrey, Richard. 1983. *The Logic of Decision*, 2nd Edition. New York: McGraw-Hill.

Joyce, James. Forthcoming. Are Newcomb Problems Really Decisions?

Kyburg, Henry. 1988. Powers, in *Causation in Decision, Belief Change, and Statistics*, edited by William Harper and Brian Skyrms. Dordrecht: Kluwer.

Levi, Isaac. 1997. *The Covenant of Reason: Rationality and the Commitments of Thought*. Cambridge: Cambridge University Press.

Lewis, David. 1976. The Paradoxes of Time Travel. *American Philosophical Quarterly* 13:145-52.

Lewis, David. 1979. Prisoners' Dilemma Is a Newcomb Problem. *Philosophy and Public Affairs* 8:239-49.

Lewis, David. 1981 Causal Decision Theory. *Australasian Journal of Philosophy* 59:5-30.

Nozick, Robert. 1969. Newcomb's Problem and Two Principles of Choice. In *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher. Dordrecht: Riedel.

Spohn, Wolfgang. 1977. Where Luce and Kranz Do Really Generalize Savage's Decision Model. *Erkenntnis* 11: 113-134.