Defending Causal Decision Theory

Brian Weatherson *

November 27, 2008

In "Some Counterexamples to Causal Decision Theory", Andy Egan argues that causal decision theory cannot handle certain cases that I'll call 'asymmetric Death in Damascus' cases. I'll argue that causal decision theory is not undermined by these cases. Egan's arguments all turn on intuitive judgments about cases. While I think these kind of philosophical intuitions are an important part of our philosophical evidence, I also think they are defeasible. And the particular intuitions that Egan relies on lead, in some extensions of his cases, to conclusions that are quite implausible. So I think the intuitions should be rejected, and without them causal decision theory is safe.

1 Death in Damascus

The original Death in Damascus is introduced to the decision theory literature in Allan Gibbard and William Harper's classic paper, "Counterfactuals and Two Kinds of Expected Utility." (Gibbard and Harper 1976)

> Consider the story of the man who met Death in Damascus. Death looked surprised, but then recovered his ghastly composure and said, 'I AM COMING FOR YOU TOMORROW'. The terrified man that night bought a camel and rode to Aleppo. The next day, Death knocked on the door of the room where he was hiding, and said 'I HAVE COME FOR YOU'.
>
> 'But I thought you would be looking for me in Damascus', said the man.
>
> 'NOT AT ALL', said Death 'THAT IS WHY I WAS SURPRISED TO SEE YOU YESTERDAY. I KNEW THAT TODAY I WAS TO FIND YOU IN ALEPPO'.
>
> Now suppose the man knows the following. Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with Death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo...
>
> If... he decides to go to Aleppo, he then has strong grounds for expecting that Aleppo is where Death already expects him to be, and hence it is rational for him to prefer staying in Damascus. Similarly, deciding to stay in Damascus would give him strong grounds for thinking that he ought to go to Aleppo.

*Thanks to Ishani Maitra and Wolfgang Schwartz.

In cases like this, the agent is in a real dilemma. Assuming that he goes to Aleppo, probably he would have been better off had he gone to Damascus. And if he stays in Damascus, then probably he would have been better off if he had left. As soon as he does something, it will be the thing that is irrational to do, given his evidence.

The case as presented has two complicating features. First, given that there is only one Death, the man can avoid Death's predictive powers by using some kind of randomising device to choose where he goes. In game theoretic terminology, the man could play a mixed strategy. (This is recommended in Weirich 2008.) If Death could be in multiple places, and would be if he predicted the man would do this, this option would be closed off. Second, it isn't clear how much disutility the man puts into buying a camel, riding to Aleppo etc. It seems from the case that the utility or disutility of this is supposed to be minimal, but it would be good to be more specific, and to think about cases where that disutility is minimal. For instance, we could imagine a case where buying the camel would bankrupt the man's heirs. Formally, we'll consider cases that have the following structure.

	S_1	S_2
O_1	x_{11}	x_{12}
O_2	x_{21}	x_{22}

We assume that taking option O_1 is very good evidence that the agent is in state S_1, and taking option O_2 is very good evidence the agent is in state S_2. We also assume that $x_{11} < x_{21}$ and $x_{22} < x_{12}$, so whatever the agent does, they have evidence that they would have been better choosing otherwise. We'll also assume, though the grounds for this assumption will need to be specified, that mixed strategies are unavailable, or unadvisable, for the agent. Any such case is a Death in Damascus case.

2 Asymmetric Death in Damascus

An asymmetric Death in Damascus case is simply a Death in Damascus case, as specified above, with $x_{11} \neq x_{22}$.[1] We'll notate our cases so that $x_{11} > x_{22}$. Egan's examples are all asymmetric Death in Damascus cases with three distinguishing characteristics.

- x_{11} is much much greater than x_{22}.
- x_{12} is much much greater than x_{22}.
- x_{21} is just a little greater than x_{11}.

Here, for instance, is his example *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 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 be detonated (burning up the $1,000,000) if you press the two-box button. 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.

[1] Such cases appear to be first discussed by Richter (1984).

Egan doesn't make explicit what happens if the demon predicts you'll play a mixed strategy, but let's assume, as in the original Newcomb case, that the predictor will treat this like two-boxing, and include the bomb. And let's further assume, as seems reasonable, that given this mixed strategies are a very bad idea in the circumstances. Now let's look at the payoff table for Newcomb's Firebomb. I'll assume, as seems harmless enough in these cases, that payoffs in dollars translate easily and linearly to payoffs in utilities.

	One-Boxing Predicted	Two-Boxing Predicted
Take one box	1,000,000	1,000,000
Take two boxes	1,001,000	1,000

As we can see, $x_{11} = 1,000,000$ is much much greater than $x_{22} = 1,000$, and only a little less than $x_{21} = 1,001,000$, while $x_{12} = 1,000,000$ is also much much greater than x_{22}. So this fits the pattern described above. I'll call any such case an Egan case, and take Egan's core premise to be that O_1 is uniquely rational in Egan cases. Without this premise, he doesn't have an objection to causal decision theory. Indeed, his objection to causal decision theory can be summed up as follows.

1. In Egan cases, taking O_1 is uniquely rational.
2. Causal decision theory does not say that taking O_1 in Egan cases is uniquely rational.
3. So, causal decision theory is either mistaken (if it says that taking O_1 is not uniquely rational in Egan cases) or incomplete (if it does not say anything about what the rational thing to do in Egan cases).[2]

In the next section I'll consider certain three option cases, and argue that if we assume that taking O_1 in any Egan case is clearly rational, then we end up with an implausible set of commitments. My conclusion is that, intuitive as it may be that O_1 is the uniquely rational choice in an Egan case, this intuition cannot ultimately be accepted.

3 Egan Cases with Alternatives

In each of the following cases, the agent has a choice between three boxes, of which they can choose exactly one. In each case there is a demon that predicts what the agent will choose, and who is very good at making predictions. In particular, the demon is probably correct conditional on any choice the agent makes. How much money the demon puts into each box is dependent on her predictions of the agent's choice. I won't specially notate this in any way, but in each case, if the demon predicts that the agent is using a mixed strategy, then the demon will put no money in any box. And I'll assume this is sufficient for the agent to not to play a mixed strategy. [3]

We'll be interested in two cases - here is the first of them. The row's represent the player's choices, the columns represent the demon's predictions. The cells represent how much money the agent gets given the prediction, as specified in the row, and the prediction, as specified in the column.

[2]Egan notes that given a particular implementation of causal decision theory, that in Lewis (1981), and some particular assumptions about the agent's credences, the agent will choose O_2, which he regards as irrational. But Lewis's implementation is not the only implementation, and the credences Egan ascribes are neither obviously correct nor obviously part of causal decision theory. So it isn't obvious, I think, that causal decision theory as such recommends choosing O_2. But I think it is clear that it doesn't uniquely recommend O_1, as Egan thinks a good theory should.

[3]Arntzenius (2008) argues that the agent should use a mixed strategy in Egan cases as originally described. This is less plausible given my stipulations about the demon.

	Demon predicts A	Demon predicts B	Demon predicts C
Agent chooses A	\$4000	\$5000	\$5000
Agent chooses B	\$5000	\$1000	\$5000
Agent chooses C	\$5000	\$800	\$4800

I'm going to argue that if O_1 is the uniquely rational choice in Egan cases, the choosing A is the uniquely rational choice here.

Note first that choosing B weakly dominates choosing C. If the demon predicts that the agent will choose A, then B and C are just as good, and otherwise B is better than C. So B is at least as good as C, and better if the probability that A is predicted is less than 1.

Now we'll just look at the comparison between A and B. Assume temporarily that the demon will either predict that A will be chosen or predict that B will be chosen. Conditional on that assumption, the choice between A and B looks like this.

	Demon predicts A	Demon predicts B
Agent chooses A	\$4000	\$5000
Agent chooses B	\$5000	\$1000

But this is an Egan case by our definition, with A being O_1 and B being O_2. By hypothesis, O_1 is uniquely rational in Egan cases. So conditional on the demon predicting A or B, A is a better choice than B. It is clear that conditional on the demon predicting C, that A and B are equally good choices. Since those two options, either the demon predicts A or B, or the demon predicts C, exhaust the possibilities, and A is a better choice on one and just as good a choice on the second, then A is at least as good a choice as B, and better unless the probability of the second option, i.e. C being predicted, is 1. (This is just an application of the sure thing principle.)

Since A is at least as good as B, and better unless the probability of C being predicted is 1, and B is at least as good as C, and better unless the probability of A being predicted is 1, it follows that A is better than C by the transitivity of preference. Indeed, it seems that A must be considerably better, since O_1 is meant to be a clearly preferable choice in an Egan case.

Here is the second case we'll be looking at.

	Demon predicts D	Demon predicts E	Demon predicts F
Agent chooses D	\$4000	\$800	\$5000
Agent chooses E	\$5000	\$1000	\$5000
Agent chooses F	\$5000	\$5000	\$4800

The reasoning here will be similar to the previous case, so I won't go through it in anything like the same detail. Since E weakly dominates D, E must be better than D. Conditional on the demon predicting E or F, the choice between E and F is an Egan case, with F being O_1. So by hypothesis, F is better than E conditional on E or F being predicted. If the demon predicts D, then E and F are equally good. So by the sure thing principle, F is simply better than E, unless the probability that the demon will predict D is 1, in which case they are equally good. By transitivity, F is better than D.

But this all seems exceedingly odd. The difference between the A/C comparison and the D/F comparison is simply that, if the demon predicts that neither of them will be chosen, then A is better than C, and D is better than F. But since there is very little reason for the 'middle' option, i.e. B or E, to be chosen, and the demon knows this, and the agent knows the demon knows this, the probability of the middle option being predicted is vanishingly small. So it can't explain much by way of why one option would be better than another.

Yet given merely the assumption that O_1 is uniquely rational in Egan cases, we concluded that A is better than C, and D is worse than F. If O_1 is clearly the rational choice in Egan cases, then A should be clearly better than C, and D clearly worse than F. But given the similarities between A and D, and between C and F, that seems implausible. On the other hand, if O_1 is not clearly better in Egan cases, then it's hard to see how we could know it to be better, and hence hard to see how we could use the fact that it is better as a premise in an argument against causal decision theory.

4 Conclusions

I've argued that if we start with the assumption that O_1 is always the right thing to do in Egan cases, then in three way choices like the choice between A, B and C, or between D, E and F, the comparison between the outer cases turns entirely on how lucrative those choices are if the demon predicts that the 'middle' choice will be made. And that's despite the fact that if it is always irrational to take O_2 in an Egan case, it is very improbable that when a rational player faces these choices, the demon will predict that they will take the middle choice. That claim about probability is true, by the way, whether we consider the probability either before or after the agent has made their choice.[4] Since this is implausible, I think we should reject the idea that option O_1 is uniquely rational in Egan cases. And with that, we should reject the argument from Egan cases to the conclusion that we should either revise or extend causal decision theory.

I have used a few other assumptions in making this argument. I've assumed the sure thing principle, and the principle that dominating options are always better than dominated options, and the transitivity of rational preference. But I assume these are safe principles to use in context. If these assumptions were false, we wouldn't need anything as fancy as asymmetric Death in Damascus cases to raise troubles for causal decision theory. If one doubts those principles then, perhaps the best conclusion to be drawn from this paper is that asymmetric Death in Damascus cases raise no *new* problem for causal decision theory.

[4]The prior probability is what causal decision theorists take to be salient; evidential decision theorists are more interested in the probability conditional on the agent's choice.