Defending Causal Decision Theory

Brian Weatherson

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'm going to argue that causal decision theory is not undermined by asymmetric Death in Damascus cases.

Egan's arguments all turn on intuitive judgments about such cases. Those intuitions, insofar as they are reliable, seem to support a quite general principle, that I'll call Egan's Safety Principle or (ESP).

When we are discussing principles in decision theory, there are two things we have to check. One is whether the principle gives plausible results when it is the only principle we need to use to make a decision. And (ESP) does quite well by that standard. That is, in effect, what Egan shows. The second is whether the principle leads to plausible results in more complicated cases when conjoined with other, plausible, principles of practical inference. And I'll argue that (ESP) does very badly on this test. Indeed, combined with some fairly innocuous principles, (ESP) ends up giving us contradictory advice about a case. If we take those other principles to be laws of the *logic* of decision, then (ESP) is inconsistent. Even if we don't draw such a strong conclusion, we'll see that the outputs of (ESP) are confusing at best, and in some cases incoherent. This suggests to me that both (ESP) and the intuitions that support it are unreliable, and hence shouldn't ground an overthrow of causal decision theory.

1 Death in Damascus

Egan's examples are similar in some respects to the Death in Damascus case 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, 1978, 157-158)

> 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.

† Unpublished.

> 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.

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. So I will mostly ignore cases where mixed strategies offer a way out.[1]

The second complicating factor is that 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 not 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, where O_1 and O_2 are choices, S_1 and S_2 are states, x_{ij} is the payoff for making choice O_i in state S_j, and for each i choosing O_i is evidence that the agent is in state S_i.

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

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.

[1] In section 4 I briefly note that even if we allow mixed strategies, we don't end up with a considerably more intuitive outcome.

2 Asymmetric Death in Damascus

An asymmetric Death in Damascus case is simply a Death in Damascus case, as specified above, with $x_{11} \neq x_{22}$.[2] We'll notate our cases so that $x_{11} > x_{22}$. Egan's examples are a subset of 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}.

If those three conditions are met, we'll call the case an 'Egan case', and call O_1 the 'Safe' option and O_2 the 'Risky' option. And we'll say call S_1 the 'PredSafe' state and S_2 the 'PredRisk' state. We'll illustrate these terms with Egan's example *Newcomb's Firebomb*. (Egan, 2007, 109-110)

> 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.

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 (PredSafe)	Two-Boxing Predicted (PredRisk)
Take one box (Safe)	1,000,000	1,000,000
Take two boxes (Risky)	1,001,000	1,000

As we can see, x_{11} (i.e., 1,000,000) is much much greater than x_{22} (i.e., 1,000), and only a little less than x_{21} (i.e., 1,001,000), while x_{12} (i.e., 1,000,000) is also much much greater than x_{22}. So this fits the pattern described above. So the principle that

[2] Such cases appear to be first discussed by Richter (1984). Among other things, he noted some of the ways I listed two paragraphs ago in which the original Death in Damascus case could be an asymmetric case.

the intuitions driving Egan's argument support seems to be that to be that taking the safe option is uniquely rational in Egan cases. This principle will play a big role in what follows, so let's give it a name.

Egan's Safety Principle (ESP) In an Egan case, taking the Safe option is the unique rational choice.

Using this principle we can give a brisk statement of Egan's objection to causal decision theory.

1. In an Egan case, taking the Safe option is the unique rational choice.
2. Causal decision theory does not say that in an Egan case taking the Safe option is the unique rational choice.
3. So, causal decision theory is either mistaken, if it denies (ESP), or incomplete, if it does not say anything about what the rational thing to do is in Egan cases.[3]

So far we've just looked at cases where the agent has two options. In the next section I'll consider certain three option cases, and argue that if we assume (ESP) we end up with implausible conclusions. I conclude that, as plausible as (ESP) looks when we consider cases like Newcomb's Firebomb, it 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. The demon is very good at making predictions. In particular, the demon is very probably correct in her prediction 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. [4]

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

[3] 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 the Risky option. Indeed, given the variety of implementations of causal decision theory, it isn't obvious that causal decision theory as such makes any prescription about Egan cases. But Egan is clearly right that causal decision theory of any stripe doesn't uniquely recommend the Safe choice, and that's enough to get an objection to causal decision theory going if (ESP) is true.

[4] 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	4000

I'm going to argue that if (ESP) is true, choosing A is uniquely rational 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 Safe and B being Risky. And we're assuming (ESP). 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. But those two options, either the demon predicts A or B, or the demon predicts C, exhaust the possibilities. And A is a better choice than B on the first option and just as good a choice as B on the second option. By an application of the sure thing principle, A is at least as good a choice as B, and better unless the probability that the demon predicts C is 1. Slightly more formally, the argument is

1. Conditional on the demon predicting A or B, choosing A is better than choosing B.
2. Conditional on the demon predicting C, choosing A is exactly as good as choosing B.
3. Those two options (the demon predicting A or B; the demon predicting C) are exclusive and exhaustive.
4. So choosing A is at least as good as choosing B, and better if the probability that the demon predicts C is less than 1.

The motivation for the first premise is (ESP). The second and third premises are true by stipulation in the case. And the validity of the argument is guaranteed by the sure thing principle. Our next step involves an application of the transitivity of *better than*.

1. Choosing A is at least as good as choosing B, and better if the probability the demon predicts C is less than 1.

2. Choosing B is at least as good as choosing C, and better if the probability the demon predicts A is less than 1.
3. So choosing A is better than choosing C.

The first premise is what we derived from the previous argument. The second premise is true by weak dominance. Transitivity alone merely gives us that choosing A is at least as good as choosing C. But the two could only be exactly as good if choosing A was exactly as good as choosing B, and choosing B was exactly as good as choosing C. And the conditions under which those two equalities obtain are incompatible, since it would require that both the demon predicting that A is chosen and the demon predicting that C is chosen have probability 1.

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 choosing the Safe option is meant to be a clearly preferable choice in an Egan case.

Here is the second case we'll be looking at, what I'll call the 'DEF choice'.

	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	4000

The reasoning here will be similar to the ABC choice, 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 Safe and E being Risky. So by the assumption of (ESP), 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, given (ESP), there is very little reason for picking 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.

I conclude from all this that we can't always accept (ESP). Given (ESP), there is almost no relevant difference. But (ESP) implies there is all the difference in the world between them. So (ESP) is incoherent, and hence false.

Here's another way of looking at the problem that these choices raise for (ESP). Assume you're a rational agent making the ABC or DEF choice, as described above,

and (ESP) is a true constraint on rational decision making. Then both B and E are ruled out, as shown above. And the demon knows you are rational, so the demon won't predict B or E. So the choices in question look like these.

	Demon predicts A	Demon predicts C
Agent chooses A	4000	5000
Agent chooses C	5000	$4000

	Demon predicts D	Demon predicts F
Agent chooses D	4000	5000
Agent chooses F	5000	4000

It looks like the same choice! Given (ESP), that is, the ABC choice and the DEF choice are on a par. But also given (ESP), it is irrational to choose C over A, and rationally mandatory to choose F over D. That is, (ESP) both says that the choice between A and C is just the same as the choice between D and F, and says that you should treat these choices differently. That seems incoherent to me. So I conclude (ESP) is false.

But what (ESP) says about Egan cases is very intuitive. That's the point of Egan's paper; it's (ESP), not causal decision theory, that tracks intuitions around here. From this I conclude that intuitions about these problems are not to be trusted. So even if causal decision theory says somewhat counterintuitive things about Egan cases, and Egan quite clearly shows that it does, the right conclusion is that intuition is untrustworthy, and causal decision theory is not undermined.

4 Objections and Replies

I'll conclude with four possible objections to my argument, and brief replies to each of them.

Objection: The difference between the A/C choice and the D/F choice is in what happens if the demon predicts B/E. And F is much better than C conditional on the demon making this 'middle column' choice. That explains why (ESP) recommends choosing A and F.
Reply: If anything, this reasoning should point us in the opposite direction. Since the demon can 'see through' the reasoning of the objector, it is less likely that the demon will predict A is chosen than that the demon will predict D is chosen. And given that the demon predicts A will be chosen, the last thing you want to choose is A. So there's no justification here for (ESP)'s flipping between A and F.

Objection: The argument so far has only shown that there's a small gain to choosing A over C, and a small gain to choosing F over D, assuming (ESP). And perhaps the difference in the middle column could explain this.

Reply: We should reject the premise of the objection. If Egan's objection to causal decision theory is to work, we have to *know* (ESP) is correct. Given standard safety principles for knowledge, that implies that the Safe option should be much better than the Risky option in an Egan case. That's hardly an uncharitable inference to draw from Egan's paper; it seems clear that in the Egan cases he discusses, the Safe option is taken to be easily superior. That implies that A should be much better than B, and hence than C. And F should be much better than E, and hence than D. But that's absurd, given that they only differ if the demon makes a prediction that (ESP) says shouldn't be made.

Objection: Given evidential decision theory, it's a wash whether we choose A or C, and a wash whether we choose D or F. So there's nothing wrong with picking, somewhat arbitrarily, A and F.
Reply: For one thing, as the previous reply shows, Egan's argument against causal decision theory requires that the choice between A and C not be a wash, but in fact be clearly in A's favour. For another, this objection turns on trivial features of the case. Imagine the following slight alternative to the ABC choice.

	Demon predicts G	Demon predicts H	Demon predicts I
Agent chooses G	4000	5000	5000
Agent chooses H	5000	1000	5000
Agent chooses I	5000	800	4500

The only difference is in the bottom right corner of the table. Since the argument for A (now an argument for G) only uses the fact that the middle row dominates the bottom row, and the middle row does indeed still dominate the bottom row, that argument still goes through. So (ESP) says that in this choice, you should choose G. But the evidential decision theorist says that, if the demon is good enough, you should choose I. Since (ESP) is inconsistent with evidential decision theory, it can't use evidential decision theory in its defence.[5]

Objection: Imagine these choices, the ABC choice and the DEF choice, are games, and the demon's payout is 1 if the prediction is correct, 0 otherwise. Then the Nash equilibrium in ABC includes A but not C, and the Nash equilibrium in DEF includes F but not D. This justifies treating the cases differently.
Reply: As with the previous reply, the primary point will be that (ESP) can't use theories that it is inconsistent with to defend its strange consequences. But the idea behind the objection is interesting enough to think through. The ABC and DEF choices are a little complex, so let's take the same idea but apply it to Newcomb's

[5]The GHI choice is problematic for (ESP) for another reason. To the extent I have the intuitions driving (ESP), I also intuit that G is better than H, that H is better than I, and that I is better than G. So in this case, I think the intuitions behind (ESP) imply intransitivity of better than. That's another sign that the intuitions are unreliable, and hence not a source of evidence.

Firebomb. And we'll assume, contrary to what was stipulated to date, that the demon won't get flustered and put no money in any box if a mixed strategy is detected. Then we'll have the following game on our hands, where the first number in each cell represents the agent's payout, and the second cell represents the demon's payout.

	One-Boxing Predicted (PredSafe)	Two-Boxing Predicted (PredRisk)
Take one box (Safe)	(1000000, 1)	(1000000, 0)
Take two boxes (Risky)	(1001000, 0)	(1000, 1)

There is a Nash equilibrium to this game, but it isn't one that helps (ESP). The equilibrium is that the demon plays PredSafe with probability 0.999, and plays PredRisk with probability 0.001. And the agent plays Safe with probability 0.5, and Risky with probability 0.5. That is, the agent simply tosses a fair coin to choose. Since (ESP) is motivated by the thought that there's only one rational choice here, the (ESP) theorist must think that playing the mixed strategy that is part of the unique Nash equilibrium is deeply misguided. If the (ESP) theorist says that, I won't object, but I would object if they then turn around and use equilibrium considerations in defending (ESP), as this objection purports to do.

References

Arntzenius, Frank, (2008). "No Regrets; or, Edith Piaf Revamps Decision Theory." *Erkenntnis* 68: 277-297, doi:10.1007/s10670-007-9084-8. (4)

Egan, Andy, (2007). "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116: 93-114, doi:10.1215/00318108-2006-023. (1, 3)

Gibbard, Allan and Harper, William, (1978). "Counterfactuals and Two Kinds of Expected Utility." In C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory*, 125-162. Dordrecht: Reidel. (1)

Lewis, David, (1981). "Causal Decision Theory." *Australasian Journal of Philosophy* 59: 5-30, doi:10.1080/00048408112340011. Reprinted in *Philosophical Papers*, Volume II, pp. 305-337. (4)

Richter, Reed, (1984). "Rationality Revisited." *Australasian Journal of Philosophy* 62: 393-404, doi:10.1080/00048408412341601. (3)

Weirich, Paul, (2008). "Causal Decision Theory." In Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, volume Winter 2008. Metaphysics Research Lab, Stanford University, Available from: http://plato.stanford.edu/archives/win2008/entries/decision-causal/. (2)