David James Barnett
VERY EARLY DRAFT: 08.16.14

A Note on Egan's Counterexamples to Causal Decision Theory

Abstract. Andy Egan presents counterexamples to causal decision theory, but offers no replacement that itself avoids counterexamples. In this paper, I propose a replacement for causal decision theory on behalf of those who accept Egan's counterexamples, show that the proposal handles the cases that led Egan to reject the potential replacements he considers, and finally consider a striking and perhaps objectionable consequence of the proposal.

1. Background

Andy Egan (2007) has proposed what he takes to be counterexamples to causal decision theory (CDT). Consider an example that Egan attributes to David Braddon-Mitchell:

> **Psychopath Button:** 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. But while Paul's evidence supports that he his not a psychopath, his evidence also overwhelmingly supports that only a psychopath would press such a button. Paul very strongly prefers living in a world *with* psychopaths to dying. Paul's unconditional confidence that he is not a psychopath is high enough that he would see it as a good thing for someone else to push the button. But on the assumption that he will press the button, he thinks he is likely to be a psychopath, and therefore is likely to be killed.

Egan claims, plausibly, that the rational thing for Paul to do is to refrain from pressing the button. For it seems that even if Paul thinks that he is probably not a psychopath, and thus that pressing the button would likely further his aims, matters are different if Paul supposes that he will press the button. For on the assumption that he will press the button, he is likely to be a psychopath, and thus is likely to be killed by his pressing the button. It thus appears that even though Paul thinks that pressing the button is likely to further his aims, he nevertheless should refrain from pressing, for it is likely to thwart his aims on the assumption that he does it.

Egan also claims, again quite plausibly, that CDT says instead that Paul should press the button. I take CDT to have two central commitments: first, that the rationally optimal option is the one with the highest expected utility, and second, that expected utility is given by a particular formula. Where U(A) is the expected utility of A-ing, v(O) is the value one assigns to the outcome O, and the Ks are **dependence hypotheses**—i.e., maximal hypotheses about how outcomes depend causally on one's actions that form a partition—CDT says that

$$U(A) = \sum_K Cr(K)v(KA).$$

Egan's claim that CDT recommends pressing is plausible because on CDT—unlike on evidential decision theory, as Egan explains—what matters for the expected utility of an action are one's unconditional credences about its potential effects. Since Paul's unconditional credence that he is not a psychopath is high, pressing the button maximizes expected utility under CDT. It is only conditional on the assumption that Paul presses that pressing does not maximize utility. And CDT says that it is Paul's unconditional credences that matter.

Both of Egan's claims have been disputed.[1] But if we accept them, CDT will need to be rejected or reformulated. What should we accept in its place? Egan considers and rejects a pair of ratificationist proposals, but he ultimately offers no replacement for CDT which avoids counterexamples. Here I will offer a replacement of CDT on behalf of those who accept Egan's two claims, show that it handles the cases that lead Egan to reject his ratificationist proposals, and finally consider a striking and perhaps objectionable consequence.

2. An Alternative to CDT

Why does refraining seem rationally preferable to pressing in Psychopath Button? I say it is because pressing looks only moderately better than refraining on the assumption that Paul refrains, while it looks far worse than refraining on the assumption that Paul presses. What this intuition motivates is a decision procedure that considers, on the assumption that Paul refrains, the difference in expected utility between refraining and pressing, and compares it to the difference in expected utility of pressing and of refraining on the assumption that he presses. Because the expected utility of pressing is only moderately greater than that of refraining on the assumption that he refrains, but is much lower on the assumption that he refrains, such a decision procedure will result in Paul's refraining.

More formally, we can let our theory understand expected utility in the same way that CDT does, but as recommending a more complicated relationship between the expected utility of an action and the rational preferability of that action. Where U(B|A) is the expected utility of B-ing on the assumption that one As, my proposal is that B-ing is rationally preferable to A-ing if and only if U(B|A) - U(A|A) > U(A|B) - U(B|B).

Given causal decision theory's understanding of expected utility, U(B|A) is given by the following formula:

$$U(B \mid A) = \sum_K Cr(K \mid A)v(KB).$$

And thus, my proposal says that B-ing is rationally preferable to A-ing if and only if

$$\sum_K Cr(K \mid A)v(KB) - \sum \ Cr(K \mid A)v(KA) > \sum_K Cr(K \mid B)v(KA) - \sum_K Cr(K \mid B)v(KB).$$

This proposal correctly says that refraining is preferable to pressing in Psychopath Button. Because pressing looks only somewhat better than refraining on the assumption that Paul refrains, U(press|refrain) - U(refrain|refrain) is positive but low. Because refraining looks

[1] See (Cantwell, 2010) and (Ahmed, Ms.) for critical discussions of Egan's claim that Paul should refrain from pressing. And see (Joyce, 2010) for critical discussion of Egan's claim that CDT says otherwise.

much better than pressing on the assumption that Paul presses, U(refrain|press) - U(press|press) is high. So refraining is preferable to pressing.

Since our proposal understands expected utility in the same way CDT does, the actions it recommends as rational differ from those recommended by CDT only in special cases where one's performing or refraining from an action itself amounts to evidence about the action's likely effects. In all but these special cases, U(B|A), which is the expected utility of B-ing conditional on the assumption that one As, will equal the unconditional expected utility of B-ing—and similarly for U(A|A), U(A|B), and U(B|B). And so our proposal will agree with CDT that B-ing is preferable to A-ing if and only if U(B) - U(A) > U(A) - U(B), which is true if and only if U(B) > U(A). The proposal will thus yield the same verdicts as CDT in familiar Newcomb and smoking lesion cases. In the special cases like Psychopath Button, however, my proposal differs from CDT in recommending the intuitively rational action.

The proposal moreover recommends the intuitively rational action in the cases that Egan presents as counterexamples to the ratificationist proposals that he considers. The first of these counterexamples is this:

> **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, or take both boxes. 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 take both boxes, 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 choose to take only box A, no bomb has been planted.

Egan claims, plausibly, that choosing only box A is preferable to choosing both boxes. And my proposal agrees. On the assumption that one chooses only box A, choosing both boxes is only $1,000 better than choosing just A. But on the assumption that one chooses both boxes, choosing only A is $1,000,000 better than choosing both. And so, U(one-box|two-box) - U(two-box|two-box) > U(two-box|one-box) - U(one-box|one-box).

The proposal also recommends the intuitively correct verdict in an example Egan attributes to Anil Gupta, which Egan presents as a counterexample to a different ratificationist proposal that he once accepted:

> **Three Option Smoking Lesion:** Samantha is deciding whether to smoke. She has three options: smoke cigars, smoke cigarettes, or refrain from smoking altogether. Due to the various ways 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. Similarly, cigarette smokers tend to be worse off than they would be smoking cigars, but better off than they would be refraining from smoking. Finally, nonsmokers tend to be best off refraining from smoking. Samantha thinks that she is most likely to smoke cigars.

Egan's once-favored ratificationist proposal says that refraining from smoking is the only rational option, but Egan says, again plausibly, that it could be preferable to smoke cigars in the right circumstances. Egan appears to suggest that this depends on one's unconditional credence that one will smoke cigars, although he does not explain why. However, it seems better, given Egan's discussion of Psychopath Button, to say instead that it depends on details like how much better off cigar smokers are for smoking cigars rather than refraining, and how much worse off they are for smoking cigars rather than cigarettes. Let's suppose that cigar smokers are much better off smoking cigars than refraining, and only a little worse off smoking cigars than smoking cigarettes. Let's furthermore suppose that cigarette smokers are much worse off smoking cigarettes than cigars, and that nonsmokers are only slightly better off refraining than smoking. If so, our proposal will say, plausibly, that smoking cigars is preferable to the other options. It is preferable to smoking cigarettes because smoking cigarettes is only slightly better than smoking cigars on the assumption that one smokes cigars, while smoking cigars is much better than smoking cigarettes on the assumption that one smokes cigarettes—and thus, U(cigars|cigarettes) - U(cigarettes|cigarettes) > U(cigarettes|cigars) - U(cigars|cigars). And it is preferable to refraining from smoking because refraining is only slightly better than smoking cigars on the assumption that one refrains, while smoking cigars is much better than refraining on the assumption that one smokes cigars—and thus, U(cigars|refrain) - U(refrain|refrain) > U(refrain|cigars) - U(cigars|cigars).

3. A Striking Consequence

The preferability relation suggested by my proposal has a number of attractive features in common with the one suggested by more familiar decision theories such as CDT. For example, the preferability relation is antisymmetric, since if U(B|A) - U(A|A) is greater than U(A|B) - U(B|B), then U(A|B) - U(B|B) cannot be greater than U(B|A) - U(A|A). There is, however, a striking and, some might think, objectionable difference. Unlike the preferability relation suggested by CDT, the one suggested by my proposal is not transitive. So under the proposal, it is possible for A-ing to be preferable to B-ing, and for B-ing to be preferable to C-ing, and yet for C-ing to be preferable to A-ing. This opens the door for **prudential quagmires**—cases in which each of one's options has another option that is rationally preferable to it. Consider and example from Arif Ahmed:

> **Psychopath Lever:** All is as before in the Psychopath Button case, but with one difference. In addition to the option of pressing the button or refraining, Paul is presented with a third option: He may pull a 'kill all psychopaths' lever. The effects of pulling such a lever are the same as those of pressing the button. But Paul knows that, for whatever reason, one's pulling the lever provides no evidence at all that one is a psychopath. Pulling the lever comes with an additional fee, however, which will do damage to Paul's aims regardless of whether he lives or dies.

In Psychopath Lever, my proposal says that each of Paul's three options has another option that is preferable to it. It says that refraining is preferable to pressing the button, for reasons we already have seen. It says that pulling the lever is preferable to refraining, since on the assumption that Paul pulls the lever, it is likely that pulling will result in Paul's living in a world without psychopaths—and so, U(pull|refrain) - U(refrain|refrain) > U(refrain|pull) - U(pull|pull). And it says that pressing the button is preferable to pulling the lever, for no matter whether Paul is a psychopath, the results of pressing and of pulling will be the same aside for the additional fee Paul must pay to pull—and so, U(press|pull) - U(pull|pull) > U(pull|press) - U(press|press).

I will not here try to respond to those, like Ahmed, who think we should reject the possibility of prudential quagmires. Instead, I will note only that for those of us who are already convinced that refraining is the rational thing to do in Psychopath Button, we should not be dissuaded from accepted my proposal on account of its allowing for prudential quagmires. For as Ahmed persuasively argues, *any* view that says that it is irrational to press in Psychopath Button must allow for prudential quagmires in cases like Psychopath Lever. Ahmed considers this result unacceptable, but I think it is the most plausible thing to say about such a case. Those who disagree must say which of the options they think is rationally optimal, and an adapted form of Ahmed's argument shows the difficulty of any answer they might give. If they say it is pressing the button, then it is difficult to say that pressing is irrational in Psychopath Button—for it is hard to see how merely gaining the additional option of pulling the lever could result in refraining no longer being preferable to pressing if it was preferable to begin with. If they say it is pulling the lever, then they will have a hard time avoiding the verdicts about smoking lesion and Newcomb cases that have led others to reject evidential decision theory in favor of causal decision theory. For a case where one is *forced* to choose between pushing the button or pulling the lever is structurally equivalent to such cases—and it is hard to see how gaining the option of refraining could result in pushing no longer being preferable to pulling. If they instead say that refraining is optimally rational, then they seem to flout the weighing of risks and rewards in a manner recommended by every plausible decision theory. For by stipulation, Paul is confident enough that he is not a psychopath that any plausible decision theory would recommend that he pull the lever if given the option between doing so and refraining—and it is hard to see how giving Paul the additional option of pressing the button could result in pulling no longer being preferable to refraining.[2] Because Ahmed thinks we should not be open to the possibility of prudential quagmires, he takes the upshot to be that it is, after all, rational for Paul to press in Psychopath Button. Those of us who are instead convinced that refraining is rationally preferable to pressing Psychopath Button must accept that in Psychopath Lever each of Paul's options has another option that is rationally preferable to it.

[2] In reconstructing Ahmed's argument, I have been careful to avoid his commitment to what I take to be an unnecessarily general Independence principle, which says that whenever A-ing is preferable to B-ing, adding an additional option C can never make B-ing the rationally optimal option (presumably because merely adding C as an option should not make B-ing preferable to A-ing if it was not already). While the particular applications of this principle necessary for Ahmed's argument seem defensible to me, I think there are counterexamples to the general principle. Ahmed himself considers some potential counterexamples to Independence, but plausibly dismisses those he considers as irrelevant. However, he fails to note that anyone attracted to Egan's verdict in Psychopath Button should be open to more germane counterexamples to Independence. For Egan's basic idea is that an agent's selecting a particular one of his options can itself provide evidence as to his options' likely consequences, and that he should take that source of evidence into account when deciding which option to take. Since adding a new option C to those available to an agent has the potential to affect evidential import of his choosing one of his initial options A and B, adding an option should under a view like Egan's have the potential to affect whether A-ing is preferable to B-ing. Consider for example a case where Paul initially has the option of either stepping on a 'kill all psychopaths' pedal or refraining. Stepping on the pedal provides no evidence that one is a psychopath, and it comes with a minor reward. Suppose that Paul is fairly confident that he is no psychopath, but unlike in the other cases, he is not quite confident enough to make stepping on the pedal look like a good bet. Should he be offered the additional option of pressing the 'kill all psychopaths' button, whose pressing does amount to evidence that one is a psychopath, his selecting the pedal instead of the button might be enough additional evidence that he is not a psychopath to tip the scales in favor of stepping on the pedal. And so, having the additional option of pushing the button might, on the account I favor, tilt the scales sufficiently to make stepping on the pedal the rationally optimal action, even though refraining is preferable to stepping when no option to press the button is present.

References

Ahmed, Arif (Ms.) 'Smokers and Psychos: Egan Cases Don't Work'

Cantwell, John (2010) 'On an Alleged Counter-Example to Causal Decision Theory' *Synthese* 173(2): 127-152.

Egan, Andy (2007) 'Some Counterexamples to Causal Decision Theory' *Philosophical Review* 116(1): 93-114.

Joyce, James M. (2012) 'Regret and Instability in Causal Decision Theory' *Synthese* 187(1): 123-145.