***Smokers and Psychos***

Andy Egan has described a family of ingenious cases that he calls counterexamples to Causal Decision Theory (CDT). They are cases where CDT recommends an irrational act (or so he says). Egan’s reasoning that this is so is the same for all the cases. And in fact he supplies a recipe for constructing the cases. It is as follows:

Start with a counterexample to EDT [Evidential Decision Theory] 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 that 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, whereas CDT’s recommendations remain the same.[[1]](#footnote--1)

In this paper I’ll follow the recipe to construct an Egan-style case. Then I’ll argue that it *isn’t* a counterexample to CDT.

Here is how I’ll argue. First (in section 1) I’ll describe a ‘basic’ case, that is, a supposed counterexample to EDT. Then (in section 2) I’ll describe an ‘Eganized’ case, that is, a supposed counterexample to CDT got by applying Egan’s recipe to the basic case. Then (in section 3) I’ll argue that CDT only switches between the basic case and the Eganized case if the agent’s utilities meet a certain condition. Then (in section 4) I’ll describe a third, ‘straight’ case, that is like the Eganized case except that (a) the payoffs are slightly different; and (b) there is none of the funny business that we find in the Eganized and basic cases—that is, no relevant states of nature are causally independent of, but evidentially dependent on, the agent’s act. I’ll argue that what CDT recommends in the straight case is clearly the rational option in the straight case. Then (in section 5) I’ll construct a fourth, ‘mixed’ case that combines options from the Eganized case and the straight case. I’ll argue that if CDT goes wrong in the Eganized case then it must also go wrong in the mixed case. But if it goes wrong in the mixed case then it must already be going wrong in the basic case, which was supposed to be a counterexample to EDT, not CDT.

The upshot is that if it’s rational to follow CDT in the basic case then it’s rational to follow CDT in the Eganized case. So it’s rational to follow CDT in the Eganized case *unless* the basic case is *also* a counterexample to CDT, and not a counterexample to EDT. So the Eganized case is only a counterexample to CDT if it adds nothing to the case against CDT that the basic case *already* supplied. In any event Egan’s cases should not convince anyone to *drop* CDT.

1. *The basic case*

I’ll start with Egan’s own counterexample to *Evidential* Decision Theory.

*Smoking Lesion A*

Susan is debating whether to smoke cigarettes of Brand A or not to smoke at all. She believes that the smoking of these cigarettes is strongly correlated with lung cancer, but only because there’s a common cause—a lesion L that tends to cause *both* the smoking of these cigarettes *and* the 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 abstention without cancer, she prefers smoking with cancer to abstention with cancer, and she prefers abstention without cancer to smoking with cancer.

Should Susan smoke cigarettes of Brand A or should she abstain from smoking? It seems clear that she should smoke.[[2]](#footnote-0)

The following table summarizes Susan’s possible acts, the relevant states of nature, the possible outcomes and my abbreviations for each of these:

|  |  |  |
| --- | --- | --- |
|  | L: Lesion present | ¬L: Lesion absent |
| A: Smoke A  | O1: Smokes, cancer | O2: Smokes, no cancer |
| Z: Abstain  | O3: Abstains, cancer | O4: Abstains, no cancer |

 **Table 1: Smoking Lesion A**

The argument that A is rationally optimal in {A, Z}[[3]](#footnote-1) is intuitively plausible. It goes like this. Either the lesion is present or the lesion is absent. In either case there is nothing that Susan can do about it now. So if it is present then she is better off smoking; and if it is absent then she is better off smoking. So she is better off smoking.

 Furthermore, Causal Decision Theory recommends smoking over non-smoking in *Smoking Lesion A*.[[4]](#footnote-2) To see this, let V be Susan’s utility function: it takes each possible outcome to a real number that measures Susan’s utility for that outcome. I’ll write ‘Vi’ for V (Oi), Susan’s utility for outcome Oi. Let Cr be Susan’s credence function: it takes each proposition to a real number in [0, 1] that measures Susan’s confidence in its truth. I’ll write ‘X CDT Y’ to mean that Causal Decision Theory recommends option X over option Y.

Then the information about Susan’s preferences implies that:

1. V1 > V3
2. V2 > V4
3. V4 > V1

This is just a codification, in terms of Susan’s utility function, of the fact that she prefers smoking with cancer to abstention with cancer, smoking without cancer to abstention without cancer, and abstention without cancer to smoking with cancer.

Now given that the presence of the lesion is *causally* independent of whether or not Susan smokes, we can write down a necessary and sufficient condition for the claim that Causal Decision Theory recommends smoking over abstention:

1. A CDT Z iff V1 Cr (L) + V2 Cr (¬L) > V3 Cr (L) + V4 Cr (¬L)

It is easy to see from (1), (2) and (4) that Causal Decision Theory *does* recommend smoking over abstention in *Smoking Lesion A*.

On the other hand Evidential Decision Theory recommends abstention over smoking in *Smoking Lesion A*. We needn’t go very deeply into this. But briefly: Susan’s smoking of Brand A cigarettes is very good evidence that she *will* get cancer, even if it doesn’t *cause* cancer. Susan’s abstaining is very good evidence that she will *not* get cancer, even if it doesn’t *prevent* it. Susan prefers abstention *without* cancer to smoking *with* cancer, so Evidential Decision Theory advises her (wrongly it seems) to abstain.

 So we can take the following from this case:

1. It’s uniquely rationally optimal for Susan to smoke in *Smoking lesion A*
2. CDT recommends smoking over abstention in *Smoking Lesion A*
3. EDT recommends abstention over smoking in *Smoking Lesion A*

*Smoking Lesion A* is the basis to which we can apply Egan’s recipe to construct a supposed counterexample to CDT. Let me now turn to that.

1. *The Eganized case*

Applying Egan’s recipe to *Smoking Lesion A* gives us something like this:

*Smoking Lesion B*

Susan is debating whether to smoke cigarettes of Brand B or not to smoke at all. She believes that the smoking of cigarettes of this brand is strongly correlated with lung cancer, but only because there is a common cause—a lesion L\*, which tends to cause smoking of brand B *and* to make smoking carcinogenic. The lesion does no harm to anyone who does not smoke. So if the lesion is present in Susan then smoking cigarettes of Brand B will cause lung cancer. But if the lesion is absent then it will not cause lung cancer, and so she will certainly not get lung cancer.

Susan prefers smoking without cancer to abstention without cancer, she prefers smoking with cancer to abstention with cancer, and she prefers abstention without cancer to smoking with cancer.

Should Susan smoke cigarettes of Brand B or should she abstain from smoking altogether?[[5]](#footnote-3)

The following table summarizes Susan’s options and the possible outcomes.

|  |  |  |
| --- | --- | --- |
|  | L\*: Lesion present | ¬L\*: Lesion absent |
| B: smoke B  | O1: Smokes, cancer | O2: Smokes, no cancer |
| Z: abstain | O4: Abstains, no cancer | O4: Abstains, no cancer |

**Table 2: Smoking Lesion B**

 The difference between *Smoking Lesion A* and *Smoking Lesion B* is that in Smoking Lesion B the lesion *doesn’t* cause cancer. Instead the lesion enables *smoking* to cause cancer. The similarity between them is that in both cases the presence of the lesion is strongly correlated with smoking.

Egan claims that *Smoking Lesion B* is a counterexample to Causal Decision Theory. That is, he says:

1. It’s uniquely rationally optimal for Susan *not* to smoke in *Smoking Lesion B*
2. CDT recommends smoking over abstention in *Smoking Lesion B*

Note first the contrast between (5) / (8) on the one hand and (6) / (9) on the other hand. (5) and (8) make opposite claims about their respective cases, whereas (6) and (9) make the same claim about their respective cases. That is why Egan says that intuition *switches* between *Smoking Lesion A* and *Smoking Lesion B*, but CDT makes the *same* call in both cases.

 I’m going to argue that (8) and (9) can’t both be true if (5) is true. First I’ll argue briefly for what should anyway be obvious, that the truth of (9) will depend on Susan’s utilities for the outcomes in *Smoking Lesion B*. Not *every* assignment of utilities that is compatible with her stated preferences over outcomes is going to make (9) true. Or so I’ll argue.

 The question then arises: does *any* assignment of credences and utilities that makes (9) true *also* make (8) true? I’ll argue not. If CDT recommends the smoking of Brand B cigarettes in *Smoking Lesion B* then Susan’s credences and utilities make it *rational* to smoke cigarettes of Brand B in *Smoking Lesion B*. At least: this is true if (5) is true. And the causalist is entitled to maintain that (5) *is* true, at least for all that Egan says. Or so I’ll argue.

 But if no assignment of utilities makes (8) and (9) both true then *Smoking Lesion B* isn’t a counterexample to CDT after all. CDT recommends smoking over abstention in *Smoking Lesion B*, given *some* assignments of credences and utilities. It’s uniquely rationally optimal *not* to smoke in that situation, given *other* assignments of credences and utilities. But there’s no assignment of credences and utilities that makes smoking *both* CDT-optimal *and* rationally suboptimal in *Smoking Lesion B*. So if I can show that (8) and (9) are never both true then I’ve done enough to defend CDT.

 First I need to show that (9) is sensitive to Susan’s utilities and credences.

1. *Sensitivity of utilities in the Eganized case*

In *Smoking Lesion B*, Causal Decision Theory assigns the following values to Susan’s options:

1. VALCDT (B) = V1 Cr (L\*) + V2 Cr (¬L\*)
2. VALCDT (Z) = V4 Cr (L\*) + V4 Cr (¬L\*)

It will do no harm to set V4 at 0. It then follows from (10) and (11) that CDT recommends smoking Brand B over abstention if and only if:

1. V1 Cr (L\*) + V2 Cr (¬L\*) > 0

Susan is rational. So Cr (L\*) + Cr (¬L\*) = 1. And since Susan has the same preferences in *Smoking Lesion A* and *Smoking Lesion B*, we know from (1)-(3) that V2 > V1. So CDT recommends smoking if and only if Susan’s credences and utilities satisfy:

1. V2 / (V2 – V1) > Cr (L\*)

(12) and (13) are equivalent statements of the crucial condition. They tell us this: given a certain credence α that the lesion is present, Causal Decision Theory only recommends smoking Brand B over abstention if, for Susan, smoking without cancer is better than abstention by more than α as much as it is better than smoking *with* cancer.

 Now as I’ve already mentioned, Susan’s preferences over the outcomes in *Smoking Lesion B* follow from those in *Smoking Lesion A* as stated at (1)-(3). Her preferences over the outcomes in *Smoking Lesion B* are:

1. V2 > V4 > V1

But (14) does *not* force (13) to be true. It all depends on Susan’s confidence Cr (L\*) that she has the lesion.

For instance, if Susan’s credence that the lesion is present is 0.2, and she considers smoking with cancer to be much, much worse than either smoking without cancer or abstention, but smoking without cancer to be only slightly better than abstention, Causal Decision Theory will recommend abstention over smoking of Brand B in *Smoking Lesion B*. Thus if we set V1 = -1000, V2 = 1, V4 = 0 and Cr (L\*) = 0.2, then:

1. VALCDT (B) = -1000 × 0.2 + 1 × 0.8 = -199.2
2. VALCDT (Z) = 0

In this case Causal Decision Theory recommends abstention over smoking in *Smoking Lesion B*, and (9) is false.

 More generally, (9) is only true if Susan’s utility-distance between smoking without cancer and abstention is large enough, relative to the distance between abstention and smoking *with* cancer. The ration between the first quantity and the second must exceed her Cr (L\*). Clearly then (9) is sensitive to Susan’s precise credences and utilities in *Smoking Lesion B*.[[6]](#footnote-4)

1. *The straight case*

Let’s suppose that some assignment of credences and utilities *does* make (9) true. On that assignment then, CDT recommends smoking over abstention in *Smoking Lesion B*. Let Cr (L\*), V1, V2, V4 represent that assignment.

I’ll now argue that anyone with those credences and utilities is *rational* to smoke in *Smoking Lesion B*. I’ll do this in two steps. In the first step I’ll construct a third case that partly resembles *Smoking Lesion B* and in which smoking is clearly and uniquely rationally optimal. In the second step I’ll construct a *fourth* case that recombines elements of the preceding two cases. And I’ll argue that anyone who thinks it’s rational to smoke in the third case but not in *Smoking Lesion B* must say something absurd about this fourth case. At least: he must say something that is absurd if smoking is rationally optimal in *Smoking Lesion A*. I’ll conclude that smoking is rationally optimal in *Smoking Lesion B*. At least: it is if it’s rationally optimal in *Smoking Lesion A*.

First then, I’d like you to hold fixed Susan’s Cr (L\*), V1, V2 and V4 whilst considering:

*Smoking Lesion C*

Susan is debating whether to smoke cigarettes of Brand C or to abstain from smoking. She believes that lesion L\* (the same lesion as before) does *not* tend to cause people to smoke Brand C. Moreover, she believes that the smoking of Brand C is *not* positively correlated with possession of the lesion. However, she believes that L\* has the same effect on smokers of Brand C as it does on smokers of Brand B. That is, it makes smoking carcinogenic for both types of smoker. On the other hand those who do *not* possess lesion L\* will come to no harm from smoking cigarettes of Brand C.

Finally, cigarettes of Brand C give Susan just as much pleasure as cigarettes of Brand B. But they have a small *cost*: she has to pay just a little more for cigarettes of Brand C.

Should Susan smoke cigarettes of Brand C or abstain from smoking altogether?

Susan’s options, the relevant states of nature and the possible outcomes are as follows:

|  |  |  |
| --- | --- | --- |
|  | L\*: Lesion present | ¬L\*: Lesion absent |
| Z: Abstain | O4: Abstains, no cancer | O4: Abstains, no cancer |
| C: Smoke C  | O5: Smokes, cancer, small cost | O6: Smokes, no cancer, small cost |

 **Table 3: Smoking Lesion C**

Note well that in *Smoking Lesion C*, there is none of the Newcomb-style funny business that we saw in *Smoking Lesion A* and *Smoking Lesion B*. That is, Susan’s options aren’t correlated with any states of interest that those options do not *cause*. More specifically: just as in *Smoking Lesion A* and *Smoking Lesion B*, the presence or absence of the lesion is *causally* independent of whether she smokes. But in contrast with those cases, whether or not she smokes Brand C has no *evidential* bearing on whether or not she already has the lesion. Smokers of Brand C are no more ore less likely to have L\* than non-smokers.

Let’s suppose that the disutility of the extra cost of Brand C is some ε > 0, where ε satisfies the following condition:

1. 0 < ε < V1 Cr (¬L\*) + V2 Cr (¬L\*)

Certainly some ε satisfying (17) exists if V1, V2 and Cr (L\*) satisfy (9) and hence (12), which we are supposing that they do. Comparison of the bottom row of Table 3 with the top row of Table 2 then assures us that:

1. V5 = V1 - ε
2. V6 = V2 - ε

I claim now that Causal Decision Theory recommends smoking over abstention in *Smoking Lesion C*. To see this consider that

1. VALCDT (C) = V5 Cr (L\*) + V6 Cr (¬L\*)
2. VALCDT (Z) = 0

Substituting (18) and (19) into (20) and simplifying we get:

1. VALCDT (C) = V1 Cr (L\*) + V2 Cr (¬L\*) - ε

By (17) and (22) we have:

1. VALCDT (C) > 0

And from (21) and (23) it follows that:

1. Causal Decision Theory recommends smoking over abstention in *Smoking Lesion C*.

The small extra cost of Brand C doesn’t make abstention preferable to smoking or even equivalent to smoking, at least not from CDT’s perspective.

 But it follows that it’s *uniquely rationally optimal* for Susan to smoke cigarettes of Brand C in *Smoking Lesion C*. Remember that there isn’t any Newcomb-like funny business in *Smoking Lesion C*. It is a *straight* case of decision-making under risk. That is, it is a case of just that central and everyday sort that any decision theory *must* address and on which all decision theories agree. Evidential Decision Theory, ratificationist versions of CDT and EDT, and indeed every sane decision theory that I’ve ever seen, agrees with CDT about *Smoking Lesion C*.

 In short, either CDT gets it right in *Smoking Lesion C*, or decision theory is bunk. But decision theory isn’t bunk. So CDT gets it right in *Smoking Lesion C*. From this claim and (24) follows:

1. Smoking is uniquely rationally optimal in *Smoking Lesion C*

That is the claim that I want to take away from this section.

1. *The mixed case*

Now I want you to continue holding fixed Susan’s credences and utilities whilst considering a *fourth* scenario:

 *B or C?*

Susan *must* smoke; but she can smoke just cigarettes of Brand B or she can smoke just cigarettes of Brand C (and pay a small fee whose disutility is ε).

Smoking cigarettes of Brand B is just as pleasurable to Susan as smoking cigarettes of Brand C. Cigarettes of both types are carcinogenic if and only if she possesses lesion L\*. She knows that lesion L\* tends to cause people who possess it to smoke cigarettes of Brand B. It does *not* tend to cause people who possess it to smoke cigarettes of Brand C.

Should Susan smoke cigarettes of Brand B or cigarettes of brand C?

Susan’s options, the relevant states of nature and the possible outcomes are as follows:

|  |  |  |
| --- | --- | --- |
|  | L\*: Lesion present | ¬L\*: Lesion absent |
| B: Smoke B  | O1: Smokes, cancer | O2: Smokes, no cancer |
| C: Smoke C | O5: Smokes, cancer, extra cost | O6: Smokes, no cancer, extra cost |

 **Table 4: B or C?**

I’ll argue that if (8) is true—that is, if smoking Brand B is rationally *sub*optimal in *Smoking Lesion B*—then smoking Brand C is rationally *optimal* in *B or C*. I’ll infer that (8) is false, because smoking Brand C is *not* rationally optimal in *B or C*.

 My first task is to argue that if (8) is true then smoking Brand C is rationally optimal in *B or C*. To that end I’ll assume the *Independence Axiom*. That axiom has many formulations of which the following is straightforward, plausible, and suited to my purposes:

1. **Independence:** Let S be a set of options and let T ⊆ S be some possibly restricted subset of options. Suppose that P ε T is rationally optimal in S. Then P is rationally optimal in T.[[7]](#footnote-5)

I’ll also assume the following *Existence Axiom*:

1. **Existence:** Let K be a finite set of options. Then at least one element of K is rationally optimal in K.

Now suppose (8) is true and consider the set of options K = {B, C, Z}. Which of K’s elements are rationally optimal in K?

Suppose first that B is rationally optimal in K. Then by (26), B is rationally optimal in {B, Z} ⊆ K: {B, Z} is the set of options in *Smoking Lesion B*. But on the supposition that (8) is true, B is *not* rationally optimal in {B, Z}, because (8) says that Z is *uniquely* rationally optimal in {B, Z}. So:

1. B is not rationally optimal in K = {B, C, Z}

Suppose next that Z is rationally optimal in K. Then by (26), Z is rationally optimal in {C, Z} ⊆ K: {C, Z} is the set of options in *Smoking Lesion C*. But we know that Z is *not* rationally optimal in {C, Z} because we know by (25) that C is *uniquely* rationally optimal in {C, Z}. So:

1. Z is not rationally optimal in K = {B, C, Z}

Now it follows from (27), (28) and (29) that:

1. C is uniquely rationally optimal in K = {B, C, Z}

Moreover, it follows from (26) and (30) that:

1. C is rationally optimal in {B, C}

And (31) says that C is rationally optimal in *B or C*. So my first claim is true:

1. If (8) is true then C is rationally optimal in *B or C*.

 My second task is to argue that C is *not* rationally optimal in *B or C*. Note first that *B or C* has the very same causal and evidential structure as *Smoking Lesion A*. In both cases, one of the options strictly dominates the other with respect to an unwanted condition that is causally independent of either option but correlated with the dominant option. So the intuitive argument for smoking’s being uniquely rationally optimal in *Smoking Lesion A* (or indeed any other argument for smoking in *Smoking Lesion A*) is also an argument for B’s being uniquely rationally optimal in *B or C*.

 Transposed to *B or C*, the argument goes like this. Either the lesion L\* is present or it is absent. In either case there is nothing that Susan can do about it now. So if it is present then she is better off smoking Brand B, because she makes a small saving. And if it is absent then she is better off smoking Brand B, again because she makes a small saving. So she is better off smoking Brand B. So:

1. B is uniquely rationally optimal in *B or C*.

By (32) and (33), (8) is false: abstention is *not* uniquely rationally optimal in *Smoking Lesion B*. So Egan is wrong: *Smoking Lesion B* is not a counterexample to Causal Decision Theory.[[8]](#footnote-6)

1. *Possible responses*

The falsity of (8) follows validly from these premises: (25), (26), (27) and (33). Can Egan reject any of them?

 There are two possible bases for rejecting (25). One would be that Causal Decision Theory, and Evidential Decision Theory, and every other decision theory that anyone has seriously proposed, makes the wrong recommendation in an utterly straightforward case of decision-making under risk, namely *Smoking Lesion C*. That is, it would be to say that decision theory is bunk. There is no reason to think that decision theory is bunk. Or if there is, then Egan should say what it s.

 The other basis for rejecting (25) would be that no value of ε in *Smoking Lesion C* makes (25) true. But the existence of some such value depends only upon the existence of a value of ε that makes (17) true. And there *is* a value of ε that makes (17) true if (12) is true, hence if (9) is true i.e. if Causal Decision Theory recommends smoking over abstention in *Smoking Lesion B*.

So to deny (25) on this basis you would have to deny (9) i.e. deny that CDT recommends smoking over abstention in *Smoking Lesion B*. But then *Smoking Lesion B* is *still* not a counterexample to CDT: if smoking is suboptimal in *Smoking Lesion B* then CDT is right not to recommend it. In short: *this* reason for thinking that (8) is true is a reason for thinking that (9) is false.

 People sometimes suggest the following sort of counterexample to (26): Stanley is very concerned to appear respectable. He also likes drinking tea. He is attending a party. Given a choice between (S1) leaving the party and (S2) staying at the party and drinking tea, his uniquely rationally optimal option is S2. But given that the choice is between (S1) leaving the party, (S2) staying at the party and drinking tea, and (S3) staying at the party and *snorting cocaine*, he wouldn’t be seen dead at the party: his uniquely rationally optimal option is S1. This violates (26) because S1 is rationally optimal in {S1, S2, S3} but not in {S1, S2}.

 But surely the explanation for a case like this is that Stanley’s valuation of the outcomes (which in this case are just his options) depends on what options are available. The pleasure (or other sort of utility) that he gets from attending a respectable party at which cocaine is *not* available exceeds the pleasure that he gets from leaving it; but the pleasure that he gets from leaving a party at which cocaine *is* available exceeds the pleasure that he gets from attending *it*.

 Nothing like that is true in Susan’s case. The pleasure that she gets from smoking Brand B or Brand C, and the disutility ε that she attaches to the cost of Brand C, are all independent of whether or not abstention is an option. So even if Stanley’s case was grounds for rejecting (26) in its full generality, (8) is still false. The falsity of (8) still follows from (25), (27), (33) and:

1. Suppose that P ε {B, C} is rationally optimal in {B, C, Z}. Then P is rationally optimal in {B, C}.

And as we’ve just seen, Stanley’s case is not any sort of grounds for rejecting (34).

 It is hard to see why anyone might reject (27). Note: (27) *doesn’t* demand that every finite option set has *just one* rationally optimal element. (That is plainly false.) As far as (27) is concerned, *every* element of *every* finite option set might be rationally optimal in that set. All it’s saying is that for any finite option set, there is *at least* one option in that set such that no option is definitely better than it. And I can’t defend (27) against any objections because I can’t think of any objections to it.

 That leaves (33). Rejecting (33) amounts to saying that smoking Brand C is rationally optimal in *B or C*. Since *B or C* is analogous in every relevant respect to *Smoking Lesion A*, the cost of rejecting (8) on these grounds is that you have to reject (5) as well. That is, you have to say that non-smoking is rationally optimal in *Smoking Lesion A*. (And in general that in any other Newcomb-type case, whatever plays the role of ‘one-boxing’ there is rationally optimal.) This is clearly counterintuitive.

 Worse than that: if there *are* grounds for rejecting (5) then Egan’s argument is redundant. That is because CDT clearly recommends smoking over abstention in *Smoking Lesion A*; so if (5) is false then CDT is clearly undervaluing a rationally optimal option.[[9]](#footnote-7) So Egan’s recipe, supposedly intended to construct a counterexample to CDT on the basis of a counterexample to EDT, is unnecessary. The basic case itself is already a counterexample, not to EDT but to CDT.

 To put it contrapositively, anyone who was a causalist when she started reading Egan’s paper would have assumed from the outset that (5) is true. Egan’s paper gives no reason for rejecting (5). So anyone who was a causalist about decision theory when she started reading the paper should still be a causalist about decision theory by the time that she finishes it.

 Assuming that the options of denying (25), (26) and (27) are indeed foreclosed, this means the following. *Smoking Lesion B*, and other cases like it, will only look like counterexamples to CDT to someone who *already* rejects CDT on the distinct and simpler grounds that CDT gets it wrong in well-known cases like *Smoking Lesion A*.[[10]](#footnote-8) On the other hand a *defender* of Causal Decision Theory—who presumably considers CDT perfectly correct about *Smoking Lesion A*—should find in *Smoking Lesion B* no reason to change her mind.

1. *The psychopath button*

I’ll conclude by briefly illustrating my argument on another Egan-type case. The aim is to convince the reader that nothing in my treatment of *Smoking Lesion B* depends on features that are peculiar to that case. *Any* attempt to apply Egan’s recipe must face the same objection.

The case that I’ll consider is probably the most intuitively compelling of all. It is:

*Psycho Button*

 Paul is debating whether to push the ‘kill all psychopaths’ button. It would, he thinks, be much better to live in a world with no psychopaths. And he knows that whether or not he presses the button makes no difference to whether or not he is a psychopath. Unfortunately, Paul is quite confident that anyone who pushes the ‘kill all psychopaths’ button is *already* a psychopath. Paul very strongly prefers living in a world with psychopaths to dying.

 Should Paul push the button?

Paul’s two options, the two relevant states of nature and the three possible outcomes are as represented in this matrix:

|  |  |  |
| --- | --- | --- |
|  | ψ: Paul is a psycho | ¬ψ: Paul is not a psycho |
| D: push button | O7: All psychos die; Paul dies | O8: All psychos die; Paul lives |
| Y: Do nothing | O9: Things go on much as before | O9: Things go on much as before |

 **Table 5: Psycho Button**

Let V be Paul’s desirability function: it takes each proposition to a number measuring the value to Paul of that proposition’s truth. Let Cr be Paul’s credence function: it takes each proposition to a number between 0 and 1 measuring Paul’s confidence in its truth.

Egan takes it to be obvious that:

1. CDT recommends pushing the psycho button over doing nothing in *Psycho Button*.
2. It’s uniquely rationally optimal *not* to push the psycho button in *Psycho Button*.

But as before, whether (35) is true depends crucially on Paul’s credences and utilities. For instance, if Paul’s credences and utilities reveal that he (a) considers himself almost certainly psychopathic and (b) would delay his own death at almost any cost, then (35) is false. So Paul’s credences and utilities have got to be just right for the story to work.

Let us write V7, V8 and V9 for Paul’s cardinal utilities for the outcomes O7, O8 and O9. I’ll assume that V9 = 0. Then Paul’s credences and utilities must satisfy:

1. V7 Cr (ψ) + V8 Cr (¬ψ) > 0; or equivalently
2. V8 / (V8 – V7) > Cr (ψ)

Suppose that they do.

Now I want you to hold fixed those credences and utilities whilst considering:

*Lever of Death*

Paul is debating whether to pull the ‘kill all psychopaths’ lever. It would, he thinks, be much better to live in a world with no psychopaths. And he knows that whether or not he pulls the lever is *not* positively correlated with *his* being a psychopath. In particular the pulling of the lever is *neither* a cause, *nor* an *effect* of his becoming or already being a psychopath.

 However, Paul also knows that the lever is coin-operated and he will have to pay to pull it. The disutility of that payment is some small positive quantity ε’ satisfying 0 < ε’ < V7 Cr (ψ) + V8 Cr (¬ψ).

 Should Paul pull the lever?

Paul’s two options, the two relevant states of nature, and the three outcomes are as represented in the following matrix:

|  |  |  |
| --- | --- | --- |
|  | ψ: Paul is a psycho | ¬ψ: Paul is not a psycho |
| Y: Do nothing | O9: Things go on much as before | O9: Things go on much as before |
| E: Pull lever | O10: All psychos die; Paul dies; cost | O11: All psychos die; Paul lives; cost |

 **Table 6: Lever of Death**

For the same reasons as in section 4 I maintain that:

1. CDT recommends pulling the lever over doing nothing in *Lever of Death*.
2. Pulling the lever is uniquely rationally optimal in *Lever of Death*.

In defence of (39), note that Paul’s credences and cardinal utilities satisfy (37) and (38), and that the causal dependence of outcomes upon acts is exactly the same in *Lever of Death* as it is in *Psycho Button*. In defence of (40), note that there is no Newcomb-like funny business in *Lever of Death*: it is just the sort of straightforward case on which EDT and every other decision theory agrees with CDT. So unless decision theory is wrong in some deeper way than anyone has realized, (40) is true if (39) is true.

 Now whilst still holding fixed Paul’s credences and utilities, consider:

 *Button or Lever?*

 This time Paul must *either* push the psycho button *or* pull the lever of death.

Pushing the button and pulling the lever both have the same effect of causing all psychos to die. Neither pushing the button nor pulling the lever has any tendency to *make* Paul a psycho. Paul knows however that only somebody who is already a psycho would push the button; pulling the lever, by contrast, is something that a non-psycho would do as readily as a psycho.

As before, pulling the lever costs money. The disutility of the cost is again ε’.

Should Paul push the button or pull the lever?

Paul’s options, the two relevant states of nature, and the four possible outcomes are as represented in the following matrix[[11]](#footnote-9):

|  |  |  |
| --- | --- | --- |
|  | ψ: Paul is a psycho | ¬ψ: Paul is not a psycho |
| D: Push button | O7: All psychos die; Paul dies | O8: All psychos die; Paul lives |
| E: Pull lever | O10: All psychos die, Paul dies; cost | O11: All psychos die, Paul lives; cost |

 **Table 7: Button or Lever?**

 The argument now proceeds as before. (36) says that Y is uniquely rationally optimal in *Psycho Button*. So if (36) is true then Y is rationally optimal in {D, Y}. Since (40) is true, E is rationally optimal in {E, Y}. It then follows from (26) and (27)—the Independence and Existence assumptions—that: if (36) is true then E is rationally optimal in {D, E}. That is: if not pushing the button is uniquely rationally optimal in *Psycho Button*, then pulling the lever is rationally optimal in *Button or Lever*.

 But it’s crazy to think that pulling the lever is rationally optimal in *Button or Lever*. At any rate, that is what Egan’s target audience should think. It’s clear from Table 7 that D strictly dominates E with respect to a partition {ψ, ¬ψ} that is causally independent of whether Paul does D or E. So it should be clear to any causalist—that is, to anyone whom *Smoking Lesion A* has not *already* convinced to abandon CDT—that:

1. Pushing the button is uniquely rationally optimal in *Button or Lever*

In English: either Paul is already a psychopath or he is already a non-psychopath: nothing that he now does will *make* any difference to that. If he is a psychopath then he is better off pushing the button than pulling the lever. In either case he’ll die, but at least if he pushes the button he’ll avoid the extra cost associated with pulling the lever.[[12]](#footnote-10) And if he is already a *non*-psychopath then again he is better off pushing the button than pulling the lever. In either case he’ll get to live in a world without psychopaths. But if he pushes the button then he also gets to avoid the extra cost associated with pulling the lever. So pushing the button is uniquely rationally optimal in *Button or Lever*.

It follows that (36) is false. Pushing the button is rationally optimal in *Psycho Button*. As before, the only ways to avoid this conclusion are: reject (26), reject (27), reject (40) or reject (41). Rejecting (26) or (27) on Paul’s behalf is no more plausible than was rejecting either on Susan’s. Rejecting (40) commits you to saying one of two things. *Either* you must say that Decision Theory is bunk, which is hardly an option. *Or* you must say that no value of ε’ satisfies the conditions in Lever of Death, which in turn rules out (35). But if (35) is false and (36) is true then *Psycho Button* is still not a counterexample to CDT. Finally, nobody who thinks that *Smoking Lesion A* is not already a counterexample to CDT—that is, nobody whom Egan stood any chance of convincing of anything—is in any position to reject (41).

I conclude that *Psycho Button* raises no new difficulty for CDT. By similar reasoning we can reject all of the other supposed counterexample that Egan’s recipe throws up.

In brief, my antidote to Egan’s recipe is as follows. Egan’s recipe starts with a counterexample to EDT—a *basic case*—in which some strictly dominant act A is caused by a causally independent state of nature N that independently causes some undesirable state O. CDT recommends A in the basic case, and this seems right. The supposed counterexample to CDT—an *Eganized case*—is a case in which N gives way to N\*, a condition that enables an act B to cause O but which is itself causally independent of B. CDT recommends B in the Eganized case, but this seems wrong.

My reply is in two steps. First, construct a *straight case* in which the effects of an action C depend on the same causally independent state N\* that figured in Case 2. In addition: (a) N\* does not cause C; (b) C costs a little more than B. Then argue that if CDT recommends B in the Eganized case then C is rationally optimal in the straight case. Then, construct a *mixed case* in which the agent chooses between B and C, and argue that if CDT gets it wrong in the Eganized case then C is rationally optimal in the straight case. But that’s absurd: C is clearly worse than B—or so you must think unless you already rejected CDT for the simpler reason that it gets even the basic case wrong. So considered as an argument against CDT, the Eganized case is either ineffective or redundant.

1. *Conclusion*

I believe, and sections 1-6 aimed to convince you, that *Smoking Lesion B* is not a counterexample to Causal Decision Theory. I also believe, and section 7 aimed to convince you, that Egan’s recipe is not capable of producing such counterexamples.[[13]](#footnote-11)

 But I also believe that Causal Decision Theory is *false*—although nothing in this paper aimed to convince you of *that*. I think that properly understood, Newcomb-type cases (like *Smoking Lesion A*) *are* counterexamples to Causal Decision Theory.

So the message of this paper is not that Causal Decision Theory is *true*. The message of this paper is that if Newcomb cases have not moved you to reject CDT, then Egan cases should not move you to reject CDT.[[14]](#footnote-12)

***References***

Egan, A. 2007. Some counterexamples to causal decision theory. *Philosophical Review* 116: 93-114.

Kreps, D. 1988. *Notes on the Theory of Choice*. Boulder: Westview.

Lewis, D. 1981. Causal decision theory. *Australasian Journal of Philosophy* 59: 5-30. Reprinted in Gärdenfors, P. and N.-E. Sahlin (eds), Decision, Probability and Utility: Selected Readings. Cambridge: CUP 1988.

1. Egan 2007: 103-4. [↑](#footnote-ref--1)
2. Egan 2007: 94. I have modified the example in a few trivial ways. [↑](#footnote-ref-0)
3. In case it’s not obvious, a brief note on terminology: when I say that an option is *rationally optimal* in a set of options, I’ll mean that it’s *among* the best options in that set, *not* that it’s uniquely *the* best option in that set. When I say that an option is *uniquely rationally optimal* in a set of options, I’ll mean that it’s rationally optimal in that set, and that no *other* option in that set is rationally optimal in that set. [↑](#footnote-ref-1)
4. Again in case it’s not obvious: when I say that some decision theory *recommends an option A* in a decision problem D, I’ll mean that A scores at least as highly, on that theory’s measure of expected utility, as any other option in D. When I say that a decision theory *recommends an option A over an option B* in D, I’ll mean that A gets a strictly higher score than B, on that theory’s measure of expected utility. [↑](#footnote-ref-2)
5. This example accords roughly with Egan’s own sketch at p. 103. Again I have modified it slightly. [↑](#footnote-ref-3)
6. The reader of Egan’s original paper will have noticed his advice to ‘put *yourself* in Susan’s situation. Would *you* smoke? Would *you* take *yourself* to be irrational for not doing so?’ (Egan 2007: 94). Does applying this advice to *Smoking Lesion B* make any difference to the argument in this section?

 It makes no difference at all. Either Egan is advising you to consider what *Susan* ought to do if *she* had *your* utilities, or he is asking you to consider what *you* would do if *you* had *her* utilities. In either case the argument of this section is that (9) depends crucially on the credences and utilities of the agent involved, and not just on the fact that his or her preferences conform to (14). This is just as true if *you* are the agent as it is if *Susan* is the agent. [↑](#footnote-ref-4)
7. For this statement of the principle see Kreps 1988: 13. Kreps calls the principle ‘Sen’s α’. [↑](#footnote-ref-5)
8. Egan’s objection is that CDT recommends an irrational option in Eganized cases (see Egan 2007: 97-8). That is why it stands or falls with the truth of (8) and its analogues in other Eganized cases. But you might raise a different objection to CDT, which is that in *Smoking Lesion B* it *undervalues* an option that is rationally optimal (namely Z), even if that option is not the *uniquely* rationally optimal option.

 My argument as it stands does not address this other objection. But it is easy to see how you might address it. All we need is the principle—which is as plausible as (26)—that if an option is *uniquely* rationally optimal in any set S then it is *uniquely* rationally optimal in every subset of S that contains that option. Given that principle we can argue that if Z is rationally optimal in {B, Z} then B is not uniquely rationally optimal in {B, C, Z}. It then follows by the same reasoning as before that C is rationally optimal in {B, C, Z} and hence by (26) in {B, C}. But C is *not* rationally optimal in {B, C}. So Z, that is, abstention, does not only fail to be *uniquely* rationally optimal in Smoking Lesion B; it fails to be rationally optimal *at all*. That stronger result refutes this alternative objection to CDT. Still, I only need the weaker result, that it fails of *unique* optimality, for my case against Egan. [↑](#footnote-ref-6)
9. You might think that this is not a serious problem for CDT. After all, that abstention is rationally optimal in *Smoking Lesion A* does not imply that abstention is *uniquely* rationally optimal in *Smoking Lesion A*. So it does not imply that smoking is suboptimal in *Smoking Lesion A*. So the fact that CDT recommends smoking in *Smoking Lesion A* does not imply that it actually advises Susan to do something *irrational*.

 But trouble is just around the corner. Grant that smoking and abstention are *both* rationally optimal in *Smoking Lesion A* and consider *Smoking Lesion A\**. Here Susan must choose between abstention and paying a small amount to smoke cigarettes of Brand A. The disutility of the small payment is some δ satisfying this condition: 0 < δ < [VALCDT (Z) + VALCDT (A)]/2. Now since abstention is rationally optimal in *Smoking Lesion A*, smoking is rationally *sub*optimal in Smoking Lesion A\*, because 0 < δ. And since CDT recommends smoking in *Smoking Lesion A*, CDT recommends smoking in *Smoking Lesion A\**, because δ < [VALCDT (Z) + VALCDT (A)]/2. So CDT recommends a suboptimal and hence irrational act in *Smoking Lesion A\**.

 In short: if (33) is false then either *Smoking Lesion A* refutes CDT or a case with the same Newcomb-like structure but slightly different payoffs refutes CDT. So in any case, if (33) is false then *some* Newcomb-like case refutes CDT. [↑](#footnote-ref-7)
10. When I say that these grounds are ‘simpler’ I mean this in the precise sense that what CDT recommends in *Smoking Lesion A* depends only on *ordinal* facts about Susan’s utilities. We are given that Susan prefers smoking with cancer to abstention with cancer, that she prefers smoking without cancer to abstention without cancer, and that she prefers abstention without cancer to smoking with cancer. CDT recommends smoking to anyone whose utilities satisfy the ordinal constraints that these preferences imply i.e. V2 > V4 > V1 > V3.

 On the other hand and as we have seen, Susan’s preferences over the outcomes in *Smoking Lesion B* (as specified in (14)) are *not* enough to guarantee that CDT recommends smoking in that scenario. For any given Cr (L\*) there are utility schedules that satisfy (14) but do *not* satisfy (12) / (13). But CDT only advises smoking on schedules that *do* satisfy (12) / (13).

Now, one premise of the argument from *Smoking Lesion A* to ¬CDT, *and* of the argument from *Smoking Lesion B* to ¬CDT must be that CDT recommends smoking over abstention. The point of the foregoing is that in the first case this premise depends only on ordinal facts, whereas in the second case it depends also on cardinal facts, about Susan’s utilities. It is in this sense that anyone who *does* take Egan-type cases to be counterexamples to CDT must think that *simpler* counterexamples to CDT already exist. [↑](#footnote-ref-8)
11. You might object that the difference between buttons and levers is far too insignificant to make this case, and *Lever of Death*, coherent to anyone who has the credences assumed in *Psycho Button*? I’m not sure that that’s true: after all, the belief that only a psycho would push the button but not only a psycho would pull the lever is no more outlandish than the beliefs involved in other familiar cases (the standard Newcomb Problem being a clear example).

But if you do think it’s true, then just modify *Psycho Button*, *Lever of Death* and *Button or Lever*, in the following way. Let *Augmented Paul* be just like Paul, only he believes that he belongs to population consisting exclusively of psychopaths and lever fetishists in which nobody is *both* a psychopath *and* a lever fetishist. Then pushing the button confirms that he is a psychopath but pulling the lever does *not* confirm that he is a psychopath (because lever fetishists are, he thinks, just as likely to pull it as psychopaths). That belief is harmless in the context of *Psycho Button*: everything that Egan says about that case, and everything that I say about that case, is true of Paul if and only if it is true of Augmented Paul. And Augmented Paul’s extra belief is certainly no more peculiar than the belief that a button with these powers exists in the first place. But in *Lever of Death* and *Button or Lever*, it would be entirely rational for Augmented Paul to have the beliefs that I am imagining for Paul. [↑](#footnote-ref-9)
12. But what harm does the extra cost do if he’s dead? Assume that Paul cares about the size of his estate as much as he cares about his own bank balance. [↑](#footnote-ref-10)
13. Egan *also* claims (pp. 100-101) that oracle cases and the like—which are *not* products of his recipe—are counterexamples to Causal Decision Theory. Example: you are a mediaeval soldier and tomorrow is the day of the battle. You could buy armour tomorrow. You are highly confident that if you *were* to buy armour tomorrow then you would be *alive* by the end of tomorrow, and that if you were *not* to buy armour tomorrow then you would be *dead* by the end of tomorrow. The (infallible or near-infallible) oracle says that you will be *dead* by the end of tomorrow. CDT recommends buying the armour anyway, but that looks like a waste of money: what is the point in trying to avert a certain or near-certain fate?

 The case deserves more attention than I can give it in a paper that is focused on Egan’s general recipe. But briefly: if CDT goes wrong here because it tells the agent to avoid a near-certain fate, then I can’t see the relevant difference between oracle cases and Newcomb-type cases like *Smoking Lesion A*.

Let’s say that a person is *A-mortal* at a time t if and only if: *either* she is not a smoker of Brand A cigarettes at t *or* she has cancer at t. Then if Susan is confident that she has no lesion in *Smoking Lesion A*, then she is highly confident that if she *were* to smoke Brand A cigarettes then she would *not* be A-mortal tomorrow, say. And she is confident that if she were *not* to smoke them then she *would* be A-mortal tomorrow. She also has (in effect) an oracular assurance that she *will* be A-mortal tomorrow. So the case is in one way parallel to the armour case; and here CDT recommends smoking. So if CDT gets it wrong in the armour case because it asks you to avert what the oracle predicted then it must be getting it wrong in Smoking Lesion A for the same reason.

My tentative conclusion oracle cases are just as inadequate as *Smoking Lesion B* and its like. They don’t supply any reason to deny CDT to anyone who did not already deny CDT on the basis of Newcomb-type cases. But as I said, the case deserves further discussion. [↑](#footnote-ref-11)
14. I wish I had a diagnosis of why some people’s *intuitions* reject CDT in the Eganized cases. Maybe the mistake—if it is one—is that they weight the causal hypotheses by their probabilities *conditional on the agent’s actions*. But applied quite generally this approach leads straight back to EDT (see Lewis 1981: 385) and so sits ill with the fact that those people’s intuitions typically *endorses* CDT in the *basic* cases. There is a great deal more to say about this—though not in this paper. But whatever the explanation of this fact, the fact that a common combination of intuitions *endorses* CDT in the basic case and *rejects* it in the Eganized case implies, if I am right, that that combination of intuitions inconsistent with non-negotiable principles of rational choice. [↑](#footnote-ref-12)