Push the Button

Arif Ahmed*†

The article discusses a well-known type of counterexample to Causal Decision Theory (CDT), in which CDT recommends an option that probably causes the best outcome while itself being evidence that it causes the worst. Intuition disagrees. Many philosophers accept that this justifies either modifying CDT or dropping it altogether. I argue to the contrary that (*a*) if intuition is right about this case, then transitivity of preference must be violated in another, but (*b*) this violation is untenable. I conclude that CDT stands.

Introduction. In a family of cases due to Gibbard, Egan, and others, causal decision theory (CDT) is alleged to give advice that is plainly counterintuitive (Gibbard 1992, 218; Egan 2007). Decision theorists have generally responded in fairly elaborate ways. Some have argued that in these 'Egan cases' (as I call them), CDT or something in its spirit does not really recommend the counterintuitive option that it seems to recommend (Arntzenius 2008, 292ff.; Edgington 2011). Others have dropped CDT and have devised new decision theories that endorse intuition in those cases (Gustafsson 2011; Wedgwood 2011).

This article takes a simpler approach. I argue that CDT as it stands gets Egan cases right. Intuition gets them wrong, and for a very simple reason.

Pair-Wise Comparisons. Andy Egan and others have presented a number of cases of which the following example is entirely representative and especially compelling.

Received August 2011; revised December 2011.

*To contact the author, please write to: Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA; e-mail: ama24@cam.ac.uk.

†The author wishes to thank Rachel Briggs and Huw Price for helpful discussion and two *Philosophy of Science* referees for helpful comments. The author began this article while on a Leverhulme Trust Research Fellowship, and he gratefully acknowledges the trust's support.

Philosophy of Science, 79 (July 2012) pp. 386–395. 0031-8248/2012/7903-0004$10.00
Copyright 2012 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/666065 Published online by Cambridge University Press

TABLE 1. PSYCHO BUTTON A PAYOFF

	P: You Are a Psycho	¬P: You Are Normal
A: Push button A	−90	10
Z: Do nothing	0	0

Psycho Button A. You can push this button (button A), and you can do nothing. Pushing button A will cause all psychopaths to die; doing nothing will cause nothing. You want to live in a world without psychos; you want more not to die yourself.

If you do push button A then it is 99 to 1 that you are a psycho. This is not because pushing the button makes you a psychopath (it does not) but because in order to push it you would almost certainly have to have been one yourself. So your payoffs (in 'utiles') are as in table 1.

Finally, you are 95% confident that you are not in fact a psycho. What do you do?

CDT advises whichever option maximizes expected *utility*, where the utility of an option measures the payoff of its *effects*. Since you are 95% confident that you are not a psycho, you are 95% confident that option A (and only option A) will cause the best outcome; given your payoffs it is therefore option A that maximizes expected utility. The following simple calculation—in which U(O) is the expected utility of an option O—confirms this:

$$\mathrm{U(A)} = 0.05(-90) + 0.95(10) = 5, \tag{1}$$

$$\mathrm{U(Z)} = 0. \tag{2}$$

We have $\mathrm{U(A)} > \mathrm{U(Z)}$, so CDT advises pushing the button.

Egan rejects this advice on the grounds that intuition tells you not to push, since anyone who pushes is almost certainly a psychopath, to whom option A is therefore fatal (2007, 97).[1] What drives the antipushing intuition in *Psycho Button A* is the fact that although pushing does not *make* you a psycho, it is a sign that you are one and so also of its own liability to kill you. Similarly in all other Egan cases, CDT recommends an option that is (*a*) likely to cause the best outcome but (*b*) a strong sign that it will cause the worst. CDT responds only to *a*, but intuition seems also sensitive to *b*, hence their divergence over *Psycho Button A*.

1. That is not what my intuition says, but I grant Egan's claim that it is what everyone else's intuition says (Egan 2007, 97 n. 3).

https://doi.org/10.1086/666065 Published online by Cambridge University Press

TABLE 2. PSYCHO BUTTON B PAYOFF

	P: You Are a Psycho	¬P: You Are Normal
B: Push button B	$-90-\Delta$	$10-\Delta$
Z: Do nothing	0	0

This is what intuition says about *Psycho Button A*:

Option Z is more rational than option A
in a straight choice between them. (3)

I argue that (3) is wrong. If intuition and CDT diverge over *Psycho Button A* then it is intuition that must go.

Consider:

> **Psycho Button B.** You can push this button (button B), and you can do nothing. Pushing button B will cause all psychopaths to die; doing nothing will cause nothing. You want to live in a world without psychos; you want more not to die yourself.
>
> Unlike the pushing of button A, the pushing of button B neither makes you a psycho nor indicates that you are already a psycho. However, it does involve a small financial cost whose disutility to you is Δ, where $0<\Delta<5$. So your payoffs are as in table 2. Again, you are 95% confident that you are not in fact a psycho. What do you do?

The utilities of your options are

$$U(B) = 0.05(-90-\Delta) + 0.95(10-\Delta) = 5-\Delta \quad (4)$$

$$U(Z) = 0. \quad (5)$$

Since $\Delta<5$, it follows that $U(B)>U(Z)$: CDT advises you to push button B.

This should be uncontroversial. Pushing B does not make you a psycho, nor is it a sign that you are already one. Your choice between B and Z is completely irrelevant to whether you are a psycho. And so your very high confidence that you are not a psycho means that any sane decision theory should take the expected benefits of pushing B to outweigh the associated risk:

Option B is more rational than option Z
in a straight choice between them. (6)

https://doi.org/10.1086/666065 Published online by Cambridge University Press

TABLE 3. A OR B PAYOFF

	P: You Are a Psycho	¬P: You Are Normal
A: Push button A	−90	10
B: Push button B	−90−Δ	10−Δ

Now consider:

A or B? This time you get to choose between pushing A and pushing B (but you must push one of them). Pushing either button will kill all psychos. If you are a psycho, then you are very likely to push A and not B; conversely nonpsychopaths tend to push B. More precisely, if you push A, then it is 99 to 1 that you are a psycho; if you push B, then it is practically certain that you are not. However, pushing B incurs the same small cost as in *Psycho Button B*. So your payoffs are as in table 3.

As before, you are 95% confident that you are not a psycho. What do you do?

Calculations (1) and (4) apply to this case and imply that CDT prefers A to B. That is also intuitively plausible: nothing that you can do in *A or B* affects whether you are a psycho. But if you are a psycho, then you are better off (by Δ) pushing A. And if you are not a psycho, then you are still better off pushing A. So you should push A.

And not only intuition but also all parties to the present debate should agree, because *A or B*, is a *Newcomb problem* (Nozick 1970), in which one option strictly dominates another over a partition that is causally independent of the choice between them. CDT, common sense, and Egan himself (2007, 94–96) all agree that in Newcomb problems one should take the dominant option.[2]

So the terms of the debate support a conclusion that many will anyway find plausible:

Option A is more rational than option B in a straight choice between them. (7)

Conditions (3), (6), and (7) entail that Z is more rational than A in a straight choice between them, that B is more rational than Z in a straight choice between them, and that A is more rational than B in a straight choice between *them*. And it is hard to see how any party to this dispute

2. In any event, Egan had better concede this. If his argument against CDT needs to assume that CDT gets Newcomb problems wrong, then it begs the question against CDT.

https://doi.org/10.1086/666065 Published online by Cambridge University Press

TABLE 4. ABZ PAYOFF

	P: You Are a Psycho	¬P: You Are Normal
A: Push button A	−90	10
B: Push button B	−90−Δ	10−Δ
Z: Do nothing	0	0

could reject (6) or (7). So if (3) is true then pair-wise rational preference is *intransitive* over A, B, and Z. The rational agent chooses A over B, B over Z, and Z over A.

This is not yet a reductio of (3). Pair-wise rational preference is not obviously transitive—at least it is not obvious to me.[3] In any case I do not assume it here. Still, the facts behind this result do nonetheless cause trouble for (3). To see this, consider:

> **ABZ.** This time you get to choose among pushing button A, pushing button B, and doing nothing. Being a psycho makes you very likely to push A and equally (and very) unlikely to push B or to do nothing; not being one has the opposite effect. As before, pushing A or B causes all psychos to die; pushing B also incurs the small cost Δ. So your payoffs are as in table 4.
>
> As before you are 95% confident that you are not a psycho. What do you do?

I claim that the reasons for preferring Z to A, A to B, and B to Z in the pair-wise cases apply equally well to *ABZ*. To see this let's briefly recap them (but see app. A).

Consider first the choice between A and Z. Intuition favored Z in *Psycho Button A* because A although unlikely to cause a bad outcome is a sign that it will. But this is also a feature of *ABZ*: pushing A is unlikely to kill you but is a strong sign that it will. So if the intuition behind (3) was sound, then it applies here too.

We cannot infer that Z is rationally preferable to A in *ABZ*: Z may still be clearly inferior to B, in which case both Z and A are utterly irrational—neither is rationally preferable to the other. Still, we can be certain that if the intuition behind (3) is sound, then A cannot be rationally optimal in *ABZ*. For either Z is optimal, in which case Z beats A, or Z is not optimal, in which case neither is A. So we can infer that

$$\text{A is not rationally optimal in } \textit{ABZ}. \tag{8}$$

3. For some arguments against transitivity see Anand (1993, chap. 4); for some arguments in favor see Binmore (2009, 17–19).

https://doi.org/10.1086/666065 Published online by Cambridge University Press

It is equally clear that the case for preferring A to B is unaffected by the presence of an additional option Z. In *A or B*, the case was that A strictly dominates B over a partition to which your choice is causally irrelevant. That is also true in *ABZ*. So reasoning analogous to that behind (8) gives

$$\text{B is not rationally optimal in } ABZ. \tag{9}$$

Finally, the additional option A does nothing to illegitimate our transferring to *ABZ* the reason for preferring B to Z in *Psycho Button B*. It is still true that nothing in the choice between B and Z has any bearing (evidential or causal) on whether you are a psycho; also, your 95% confidence that you are not a psycho ensures that the risks plus the fixed cost of B do not outweigh its expected benefits. So

$$\text{Z is not rationally optimal in } ABZ. \tag{10}$$

Putting together this and the three preceding paragraphs, if (3) is true then (8)–(10) are true.

But (8)–(10) jointly imply that there is no rational option in *ABZ*. This is not the innocuous conclusion that there is no unique rational option in *ABZ* (as is the case, e.g., in the decision problem facing Buridan's Ass). It is the catastrophic conclusion that whatever you do in *ABZ* is irrational: whichever option you take, some other option is (and always was) rationally preferable to it.[4]

It may be fair to say this about an agent whose beliefs or preferences are insane or incoherent and also about one whose options are infinite. But we are not facing either case here. Your beliefs and preferences in *ABZ* are, despite their science-fictional character, plainly sane and coherent.[5] And it is equally clear that *ABZ* does not offer an infinitude of options but only three.

We are left with *two* options (but see app. B). If we grant (8)–(10), then we can accept (3) and hence also Egan's intuition that CDT gives bad advice in *Psycho Button A*. But then Egan has understated the gravity of the situation. That in some cases CDT advises an irrational option is now the least of our concerns. The real trouble is that in other cases there is no

4. You might think that CDT itself faces a similar difficulty in *Psycho Button A* and the very similar case of *Death in Damascus* (Gibbard and Harper 1978/1988, 373): whatever you decide to do, once you know that you are going to do it, CDT considers some other option to be rationally preferable. Does this not mean that CDT reckons every option irrational in these cases? The difference is that in these cases there is always a CDT-optimal option for any given state of belief. Which option is CDT-optimal might change as your beliefs evolve, but that is still compatible with the basic requirement that your available options are rationally comparable in the light of any given pattern of beliefs and desires. Taken together, (8)–(10) are not even compatible with this.

5. Or if they are not sane and coherent, then why can we not (*a*) say the same about *Psycho Button A* and (*b*) refuse for that reason to learn anything from it?

https://doi.org/10.1086/666065 Published online by Cambridge University Press

rational option. We must abandon the very idea that an agent can so much as rationally choose from among finitely many options in as straightforward a scenario as *ABZ*.

The only—and so in my view the compelling—alternative is that at least one option is rationally optimal in *ABZ*. So at least one of (8)–(10) is false.[6] Hence (3) is false: pushing A is rational in *Psycho Button A*, and CDT gives the right advice to anyone who is facing it. But then the complaint against Egan's case is not that it is understated but that it fails altogether.

Appendix A

You might worry that *ABZ* fails to preserve the probabilities that I stipulated in the three two-option situations. For instance, that you are only 5% likely to be a psycho if you take option B or if you take option Z in *Psycho Button B* does not entail that the same probability holds when option A is also available. That is true but irrelevant because we can stipulate a probabilistic structure for *ABZ* that retains whatever features of the two-option situations motivated the preferences (3), (6), and (7) in the first place.

To see this, consider a model in which your subjective probabilities reflect a distribution of results for 4,000 trials of each of the four choice situations as in table A1. So for instance, in *Psycho Button A*, your confidence that you are a psycho given that you *Push Button A* is $198/(198 + 2) = 198/200 = 0.99$. And in *ABZ*, your initial confidence that you are a psycho is $(198 + 1 + 1)/(198 + 1 + 1 + 2 + 1,899 + 1,899) = 200/4,000 = 0.05$.

The reader may check that the distribution in this table validates all of my other assumptions concerning your relevant subjective probabilities in the four cases. It is therefore coherent to stipulate probabilities on all four cases in a way that suffices for present purposes. In particular, if your probabilities for *ABZ* reflect the distribution in the table, then the arguments for (8)–(10) are as applicable to it as the arguments for (3), (6), (7) are to *Psycho Button A*, *Psycho Button B*, and *A or B*, respectively.

You might still worry that no such sequence of trials can have taken place—if it had, all the psychos would be dead by now. To get around this, we might insist on interpreting table A1 as a model of subjective probabilities but not of actual trials—after all, the coherence of these probabilities does not require that they reflect the real relative frequencies of anything. Alternatively, we might maintain that the table reflects actual trials but imagine that in all such trials the device connected to buttons A and B is

6. The defender of CDT should in fact reject (8), but nothing in my case depends on this.

https://doi.org/10.1086/666065 Published online by Cambridge University Press

TABLE A1. Hypothetical Distribution

	P. Button A		P. Button B		A or B		ABZ		
	A	Z	B	Z	A	B	A	B	Z
P	198	2	100	100	198	2	198	1	1
¬P	2	3,798	1,900	1,900	2	3,798	2	1,899	1,899

set to 'stun' rather than 'kill'—and so can be used repeatedly—and that the relevant utilities are the same on either setting. This is a little farfetched. But if you were willing to swallow Egan's device in the first place then objecting at this point on that score would perhaps be straining at a gnat.

Appendix B

I am ignoring the possibility of a 'deliberational' CDT that countenances 'mixed' acts, that is, probability distributions over the set of pure options (in this case {A, B, Z}). Roughly, the idea is that for each probability distribution P, we calculate the expected (causal) utility relative to a credence function that has been updated to take account of the fact that you are playing each option O with probability P(O) (which in this case will affect the probability that you are a psycho). The deliberational theory then searches for an equilibrium, that is, a distribution P* such that from the credential perspective of somebody about to play the strategy P* (say, with the use of a randomizing device), P* still appears to maximize expected utility compared to any other distribution P over the option set. Arntzenius (2008, 292–95) gives more details.

It is well known that equilibria are guaranteed to exist given only light assumptions about the agent's utility function. In particular there is an equilibrium for *ABZ*: for instance, an agent who plays A, B, and Z with respective probabilities of about 0.1, 0, and 0.9 will have confidence of about 0.1 that he is a psycho, and from this perspective there is no incentive to exchange this probabilistic strategy for any other. The equilibrium is *self-ratifying* in the sense that as one's credence over one's acts converges upon it, one remains confident that it is optimal. By contrast, no pure strategy in *ABZ* is thus self-ratifying; hence this case like others generates 'instability' problems for unconditional expected utility theories like CDT.

Deliberational decision theory deserves more extended discussion than is appropriate here, but two brief comments are perhaps in order. First, I share Arntzenius's concern over how to interpret a deliberational theory that favors some mixed strategy. It cannot be advising you what to do in a problem like *ABZ* because here there is no randomizing device: the only available options are the pure options A, B, and Z themselves. The alternative is to interpret it as advice about what to believe about what you are

https://doi.org/10.1086/666065 Published online by Cambridge University Press

going to do. This is fine as far as it goes, but then such a theory is not a rival to 'straight' CDT insofar as the latter tells you what to do, and so my argument, that Egan has not identified any flaw in 'straight' CDT, remains unaffected.

Second, Arntzenius's own motivation for the deliberational theory is not that 'straight' CDT recommends an option in *Psycho Button A* that is suboptimal among the pure options. It is rather that both options violate his 'Edith Piaf Principle': do not do something that you know you will regret (Arntzenius 2008, 291). In *Psycho Button A*, pushing the button violates the principle (because then you will think that you are a psycho and so will wish that you had not) and so does not pushing it (because then you will think you are not a psycho and so will wish that you had). It is violation of the Piaf principle that causes instability to arise among each of the pure options in this case as well as in *ABZ*.

But the Piaf Principle is contestable: an alternative that seems to me at least as plausible, but which poses no threat to CDT, is that foreseeable regrettability does not automatically foreclose an option but should merely be factored into its evaluation as a cost, appropriately discounted for its futurity. To take Nagel's example (1970, 74 n. 1), a young man who now values spontaneity and creativity above status and security might know that in 20 years his preferences will be reversed. There is nothing irrational about his now choosing to be a painter rather than (say) a doctor. If somebody objects that he will almost certainly regret his present choice in 20 years, he might retort that that is because in 20 years he will adhere to values that he now holds in contempt, and which values should now concern him more than his present ones?

Clearly there is more to be said on either side. My point here is not that these complications for the deliberational theory are insuperable but only that their existence gives some point to our seeking as simple a response to Egan cases as that defended here.

REFERENCES

Anand, Paul. 1993. *Foundations of Rational Choice under Risk*. Oxford: Clarendon.

Arntzenius, Frank. 2008. "No Regrets; or, Edith Piaf Revamps Decision Theory." *Erkenntnis* 68 (2): 277–97.

Binmore, Kenneth. 2009. *Rational Decisions*. Princeton, NJ: Princeton University Press.

Edgington, Dorothy. 2011. "Conditionals, Causation and Decision." *Analytic Philosophy* 52 (2): 75–87.

Egan, Andrew. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116 (1): 93–114.

Gibbard, Alan. 1992. "Weakly Self-Ratifying Strategies: Comments on McClennen." *Philosophical Studies* 65 (1–2): 217–25.

Gibbard, Alan, and William Harper. 1978/1988. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, ed. Clifford A. Hooker, James J. Leach, and Edward F. McClennen, 125–62. Dordrecht: Reidel. Repr. in Peter Gärdenfors and Nils-Eric Sahlin, eds. *Decision, Probability and Utility* (Cambridge: Cambridge University Press).

https://doi.org/10.1086/666065 Published online by Cambridge University Press

Gustafsson, Johan. 2011. "A Note in Defence of Ratificationism." *Erkenntnis* 75 (1): 147–50.
Nagel, Thomas. 1970. *The Possibility of Altruism*. Oxford: Oxford University Press.
Nozick, Robert. 1970. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher, 114–46. Dordrecht: Reidel.
Wedgwood, Ralph. 2011. "Gandalf's Solution to the Newcomb Problem." *Synthese*. doi:10.1007/s11229-011-9900-1.

https://doi.org/10.1086/666065 Published online by Cambridge University Press