ESJP
#6 | 2014

Would You Press a Button that Kills All Psychopaths?

The Two-Order Ratificationism Theorem of Causal Decision Theory

Nora Neuteboom

1. Introduction

> Paul is debating whether to press a button that 'kills all psychopaths'. Surely it would be much better to live in a world without psychopaths, Paul reasons. However, at the same time Paul is quite confident that only a psychopath would press such a button. Paul prefers living in a world with psychopaths to dying. What should Paul do?

In this situation it seems irrational to press the button and rational to refrain from doing so: if Paul would press he would be a psychopath and hence die.

However, in a recent paper Egan (2007) points out that the most sophisticated current version of decision theory, commonly called Causal Decision Theory, actually gives the irrational recommendation to press the psychopath button in this example.

How can this be? Decision theory may be pursued with the aim of finding out how decisions ought to be made in order to maximize one's utility.[1] In the standard view, rational choice is defined as the process of determining what options are available and then choosing the most preferred one. Within the rational actor model, which is the standard for Rational Choice Theory (RCT), the substance of the rational action concept is the maximization of individual utility. The individual can choose between alternative courses of action, taking into account individual preferences and states of nature. The rationality of a decision is evaluated by taking into account in what measure the results of the decision have succeeded in maximizing utility (that is, satisfying individual preferences) under some specific contextual constraints (states of nature).

There is an ongoing debate between two camps of decision theorists, namely, the proponents of Evidential Decision Theory (EDT) and the proponents of Causal Decision Theory (CDT). The difference between the two decision theories consists in how they compute the relative value of actions. Roughly speaking, EDT says to do the thing you would be happiest to learn that you did, and CDT says to do the thing that is most likely to bring about the best results (Egan, 2007: 93). In this paper, I try to contribute to the debate between causal and evidential decision theorists by solving the latest counterexamples that burden CDT by proposing a revisionist account.

In part one I begin by briefly introducing Egan's 2007 paper which shows that there are fatal problems to using both EDT and CDT for choosing rational options. Thereafter I spell out the Ratificationism Theorem (RT). Ratificationism requires the chosen act, A, to have an estimated desirability at least as high as any of the alternative choices on the hypothesis that one's final decision will be to perform A (Jeffrey, 1983: 19). I reconfirm that RT is not sufficient for solving all disputes between CDT and EDT. A problem with ratificationism is that in some cases there are no ratifiable options, but some options still seem rational. To solve this problem, Egan considers a lexical version of Ratificationism. The Lexical Ratificationism Theorem (LRT) recommends at least one option, even in cases where no option is ratifiable. I demonstrate, following Gupta, that also LRT can be refuted with a counterexample. In part two of the paper I proceed to develop an alternative ratificationist account, the Two-Order Ratificationism Theorem (TORT). I show that TORT does give us the correct results to two recently posed (and still unresolved) counterexamples to LRT.

PART I

1.1 Causal and Evidential Decision Theory

In recent philosophical debates the most prominent rival to CDT is EDT. EDT upholds that the best action is the one which, *conditional* on you having chosen it, gives you the best expected outcome. CDT maintains that the expected utility of actions should be *unconditionally* evaluated with respect to their potential consequences. CDT enjoins us to do whatever has the best expected outcome, holding fixed our initial views about the likely causal structure of the world (Egan, 2007: 94-96).

In his 2007 paper *Some Counterexamples to Causal Decision Theory*, Egan shows that there are fatal problems to both EDT and CDT. EDT argues for a policy of performing the action with the greatest evidential value, rather than the action with the best-expected causal upshot (2007: 93-96). CDT requires that the expected utility of an action is unconditionally evaluated with respect to its potential consequences (2007: 96-102). However, both EDT and CDT are mistaken in some cases according to Egan. To solve this, he modifies the CDT account by proposing RT.[2]

1.2 Ratificationism Theorem

(RT)

It is rational to perform an action A*, iff,*

1. A *is ratifiable, and*
2. *There is no other ratifiable option with greater* VAL_{CDT} *than* A.

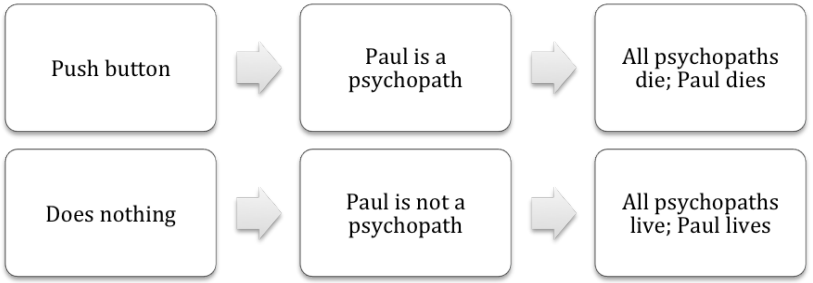
Option A *is ratifiable if, and only if, there is no alternative* B *such that the value of* B ($VAL_{CDT}(B)$) *exceeds the value of* A ($VAL_{CDT}(A)$) *on the supposition that* A *is decided upon. Therefore, it is rational to decide upon an option* A *if, and only if,* A *is the only ratifiable option.*

Egan (2007: 107) promotes RT as a kind of adjusting principle; whenever basic CDT would get it wrong, RT should help to get to the right answer.

But it seems that RT is not useful for all cases. Let us go back to where we started our journey with, the example of the psychopath:

> 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. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul prefers living in a world with psychopaths to dying.

I represent Paul's choices below:



If Paul would press the button, he would be a psychopath and hence die. Pressing is the irrational thing to do since Paul prefers living to dying. In order to have an account of decision theory that gives us the utility-maximizing guidance on rational acts, we want our decision model to advise us to not press the psychopath button.

CDT advises us to unconditionally evaluate the outcome. In this example, the fact that Paul is a psychopath (if he presses the button) is an unfortunate condition of pressing the button. But CDT does not take this condition into account and therefore encourages pressing the button. Also RT cannot help CDT from falling afoul, since neither refraining from pressing the button nor pressing the button is ratifiable. When Paul becomes convinced that he will choose to refrain, he will become quite confident that he is not a psychopath, and pressing will look better than refraining. Thus the option 'not pressing' is not ratifiable because there

is an alternative option, 'pressing', such that the value of that alternative option (*VAL(PRESSING)*) exceeds the value of the initial option (*VAL(NOT-PRESSING)*), on the supposition that the initial option 'not pressing' is decided upon. The same holds for the option 'pressing'. The moment that Paul becomes convinced that 'pressing' is the best option, the value of the option 'not pressing' (*VAL(NOT-PRESSING)*) exceeds the value of 'pressing' (*VAL(PRESSING)*) since as soon as Paul becomes aware of the fact that he prefers to press, he might be a psychopath himself. So when Paul chooses to press, the value of not pressing will exceed the value of pressing since he has a good reason to think that he is actually a psychopath.[3] It seems that no option is ratifiable, and thus RT does not give us any helpful guideline in this example.[4]

1.3 Egan's Lexical Ratificationism Theorem

Since Paul is quite sensitive about dying, we have the obligation to fix this problem. To do so, Egan (2007: 111) considers an adaption of RT, LRT.

(LRT)

It is rational to decide upon an option A *iff,*

1. A *is ratifiable, and there is no other ratifiable option with higher* VAL_{EDT} *than* A,[5] *or*
2. *There are no ratifiable options, and no other (unratifiable) option has higher* VAL_{EDT} *than* A.

Option A *is ratifiable if, and only if, there is no alternative* B *such that value of* B ($VAL_{CDT}(B)$) *exceeds the value of* A ($VAL_{CDT}(A)$) *on the supposition that* A *is decided upon. Therefore, it is rational to decide upon an option* A *if, and only if,* A *is the only ratifiable option.*

Step one advocates to order by ratifiability – that is, if A is ratifiable and B is unratifiable, then A is to be preferred over B. Step two advocates that within each of the two groups, we should order by VAL_{EDT}.[6] Thus, it seems that LRT advances CDT by taking some elements from EDT for rational decision-making.

LRT recommends at least one option even in cases where no option is ratifiable. For example, it yields the rational recommendation not to press the psychopath button. Step one is indecisive since we already saw that neither refraining from pressing the button nor pressing the button is ratifiable. In step two we refer to the EDT principle. EDT says that the rational action is the one such that your expected utility, conditional on you performing it, is greater than the expectations conditional on you performing any other action. So conditional on the fact that *if* you press you would be a psychopath, EDT prescribes not to press. And that seems to be the rational thing to do.

1.4 A refutation of the Lexical Ratificationism Theorem

LRT seemed the way to go until Gupta (Egan, 2007: 112) came up with a decisive counterexample of the Three-Option Smoking Lesion:[7]

> Paul is deciding whether to smoke. Paul has three options: Smoke cigars, smoke cigarettes, or refrain from smoking altogether. Call these options CIGAR, CIGARETTE, and NO SMOKE. Due to the ways that various 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 altogether. Similarly, cigarette smokers tend to be worse off than they would be when smoking cigars, but better off than they would be when refraining from smoking altogether. Finally, non-smokers tend to be best off refraining from smoking.

Assume:

ψ_1 = smoking cigars

ψ_2 = smoking cigarettes

$\neg\psi$ = not smoking

I represent the example formally in Table 1 below.

ψ: Smoker	$\neg\psi$: Non-smoker
$VAL(\psi_1) > VAL(\neg\psi)$	$VAL(\neg\psi) = VAL(\psi_1)$
$VAL(\psi_2) > VAL(\neg\psi)$	$VAL(\neg\psi) = VAL(\psi_2)$
$VAL(\psi_1 \mid \psi_2) > VAL(\psi_2 \mid \psi_1)$	
$VAL(\psi_2 \mid \psi_1) > VAL(\psi_1 \mid \psi_2)$	

Table 1: The values of the Three-Option Smoking Lesion

Following LRT, step one tells Paul that there is one ratifiable option, not smoking ($\neg\psi$), and there are two unratifiable options (ψ_1, ψ_2). Both ψ_1 and ψ_2 are unratifiable since smoking cigarettes is very good evidence that you would be better off smoking cigars and vice versa. Not smoking ($\neg\psi$), however, is ratifiable because it is good evidence that you would be best off not smoking. Thus, based on step one and two in LRT, Paul should not smoke. But this seems irrational: if Paul wants to smoke cigars or cigarettes, one thing Paul knows for sure is that not smoking is not the way to go.[8]

Egan thinks that this example is fatal for LRT:

> 'No ratificationist account will be able to deliver the right results in the sorts of three-option cases that Gupta has pointed out. The real importance of the Gupta cases is not that they refute lexical ratificationism— it's that they refute every form of ratificationism' (Egan, 2007: 113).

In Part II, I will prove that this is not true by providing an alternative account of LRT that does give the rational solution to Gupta's counterexample.

PART II

2.1 An adapted version of the Lexical Ratificationism Theorem

So let us look at what element in the Three-Option Smoking Lesion really causes the structural failure of LRT.

Paul has three options. Option ψ_1 is unratifiable because, conditional on choosing option ψ_1, option ψ_2 looks better than option ψ_1 – and the same applies to option ψ_2. Option $\neg\psi$ is ratifiable because, conditional on choosing option $\neg\psi$, option $\neg\psi$ looks better than either ψ_1 or ψ_2. However, conditional on choosing either one of the options ψ_1 or ψ_2, option $\neg\psi$ looks very bad. What seems clearly irrational, for Paul, who finds himself deciding on either ψ_1 or ψ_2, is to perform action $\neg\psi$ on grounds of its ratifiability. What we need is an account that considers the fact that if Paul finds himself in a situation in which he wants to choose between ψ_1 and ψ_2, the ratifiability of option $\neg\psi$ does not hold as binding anymore. So considering Paul's desire to smoke, option ψ_1 and ψ_2 should also be ratifiable.[9] From this line of thought it seems that we need to consider two sorts of ratifiability. To incorporate two sorts of ratifiability in the account, I develop the Two-Order Ratificationism Theorem version of CDT:

(TORT)

It is rational to decide upon an option A iff,

1. A *is first-order or second-order ratifiable and*
2. *There are no other first-order or second-order ratifiable options with higher* VAL_{EDT} *than* A*, or*
3. *There are no ratifiable options, and no other (unratifiable) option has higher* VAL_{EDT} *than* A.

An option A *is first-order ratifiable if, and only if, there is no alternative* B *such that* $VAL_{CDT}(B)$ *exceeds* $VAL_{CDT}(A)$ *on the supposition that* A *is decided upon.*

An option A *is second-order ratifiable if, and only if, there is no other first-order ratifiable alternative* B *such that* $VAL_{CDT}(B)$ *(on the supposition that* B *is decided upon) exceeds* $VAL_{CDT}(A)$ *(on the supposition that* A *is decided upon).*[10]

To demonstrate the above TORT principle, I will examine the Three-Option Smoking Lesion again. The option not to smoke ($\neg\psi$) is obviously a first-order ratifiable option; there is no alternative (ψ_1 or ψ_2) such that the $VAL_{CDT}(\psi_1)$ or $VAL_{CDT}(\psi_2)$ exceeds the $VAL_{CDT}(\neg\psi)$ on the supposition that Paul is a non-smoker ($\neg\psi$). Then we proceed by arguing that both option ψ_1 and option ψ_2 are second-order ratifiable. Option ψ_1 is a second-order ratifiable option since there is no other first-order ratifiable alternative ($\neg\psi$) such that

the $VAL_{CDT}(\neg\psi)$ exceeds $VAL_{CDT}(\psi_1)$ on the supposition that Paul chooses to smoke. Since Paul already decided to smoke, the value of ψ_1 exceeds the value of $\neg\psi$ in the example. The same applies to option ψ_2. So all options are either first-order or second-order ratifiable. Therefore Paul has to choose the ratifiable option that has the highest VAL_{EDT}. If ψ_1 or ψ_2 is ruled out in favor of $\neg\psi$, it is due to a higher VAL_{EDT} of $\neg\psi$. Hence, unlike LRT, TORT does not rule out ψ_1 or ψ_2 in the Three-Option Smoking Lesion on grounds of $\neg\psi$'s ratifiability, since all options are ratifiable. TORT recommends a rational choice in the Three-Option Smoking Lesion.

One could argue, however, that TORT is not really a pure ratifiability account. Since I introduce a second-order ratifibility, the content of the notion ratifibility changes, thereby changing its relation with rationality. However, one should notice that ratifiability does not necessarily imply rationality. In other words, not all ratifiable options are rational, even though all non-ratifiable options are irrational. What a ratifiability account should do is to rule out the irrational options for the agent. As Egan (2007: 108) points out: 'being unratifiable is sufficient for being ruled out as a rational option'. In the above case, Paul choosing not to smoke ($\neg\psi$) is not ruled out as an irrational option immediately. However, we cannot say it is rational either. What matters is that Paul should prefer smoking (ψ_1 or ψ_2) to non-smoking ($\neg\psi$) and that my TORT mechanism captures this. Thus, according to Egan, ruling out the irrational options is all it takes to be a genuine ratifiability account. Since TORT does exactly that, the objection that TORT is not a genuine ratifiability account is not warranted.[11]

2.2 So can TORT also account for other examples?

A possible objection to TORT could be that it is tailor-made to the specific example of the Three-Option Smoking Lesion. To refute this objection I provide two other examples where TORT is relevant.

Firstly, TORT is also applicable to the first example of the Psychopath Button since it has the same relevant elements as LRT. Recall that the result of step one is indecisive since both pressing and not pressing are not first-order (or second-order) ratifiable. We skip step two since there are no first-order ratifiable options. In step three we decide to apply EDT, and so conditional on the fact that if Paul presses he has a great chance of being a psychopath, Paul does not press the button. And that again seems to be rational.

Secondly, TORT is also applicable in another recent example of 'Picking the Box'. Arntzenius presented Picking the Box in the paper of Gustafsson (2011: 149) in order to show how LRT can be refuted:

> Paul is confronted with three boxes A, B, and C. However, Paul can only choose A or B in this first scenario. A perfect predictor has filled the boxes with money. If the predictor predicted that Paul will take A then he filled the boxes as follows: 2 euros in A, 1 euro in B, and nothing in C. If he predicted Paul will take B then he filled the boxes as follows: 4 euros in A, 3 euros in B, and nothing in C.

See the corresponding Table 2 below.

Choosing A	Choosing B
VAL(A)= 2	*VAL(A)= 4*
VAL(B)= 1	*VAL(B)= 3*
VAL(C)= 0	*VAL(C)= 0*

Table 2: Picking the box

It is rational for Paul to choose option B: he gets 3 euros instead of 2 euros, which is what he would get were he to choose option A. According to LRT, A is the only ratifiable option, and thus LRT would recommend the irrational option, namely A. Option B would be unratifiable because you would be better off choosing option A, on the supposition that B is chosen. However, TORT would get it right. Choosing option A is first-order ratifiable. And option B is second-order ratifiable since the first-order ratifiable alternative (*VAL(A)* = 2 euros) does not exceed the value of option B (*VAL(B)* = 3 euros). Now both option A and B are ratifiable, so accordingly Paul should choose the option with the highest VAL_{EDT}, namely option B.

3. Conclusion

In this essay I have argued for a transformed version of the Lexical Ratificationism Theorem (LRT): the Two-Order Ratificationism Theorem (TORT). We have seen that in order to recommend rational decisions, versions of Ratifiability are often added to Causal Decision Theory (CDT). According to proponents of CDT, imposing a ratifiability requirement will help us to save CDT. The ratifiability requirement teaches us that it becomes rational to perform an action A if and only if A is ratifiable. Two of such accounts are present in the current literature, the (original) Ratification Theorem (RT) and LRT. However, Egan and Gupta (Egan, 2007) came up with decisive counterexamples which showed that ratifiability accounts of CDT sometimes recommend the irrational action. Yet, standard Evidential Decision Theory (EDT) will also not endorse the rational action. This is bad news to decision theorists. Therefore, I introduce a first and second-order form of ratifiability to CDT. TORT can accommodate the intuitively compelling counterexamples to CDT that have recently been articulated by Egan (2007). By incorporating a second order, TORT is able to assess whether the first-order ratifiablity account is the rational action. With the introduction of TORT, I have proved that there is a ratificationist account that is able to deliver the right rational choice in the sort of three-option cases that Gupta has pointed out.

The generalizability of my proposed TORT is yet to be tested. Above all, we should try to find counterexamples that prove my TORT to recommend the irrational decision. Or, we should try to find counterexamples that show TORT does not recommend the rational decision with the highest expected value for the agent.

Acknowledgements

I would like to thank dr. H.C.K. Heilmann for his insightful comments on earlier versions of this paper. He has given me in-depth feedback and remarkable guidance on my philosophical writing along the two-year masters program. Furthermore, thanks to the journal's anonymous reviewers, I was able to substantially improve and refine my work.

Nora Neuteboom (1989) holds her BSc degree in Business Economics from the University of Amsterdam in 2010. She obtained her MSc in Economics, Markets & Policy at the Erasmus University in 2012. She wrote her thesis at the econometrics department on the Influence of Economic Activity on the Onset of Civil War. In September 2012 she started the Research Master programme at the Erasmus Institute for Philosophy and Economics (EIPE). She is currently finishing her thesis on Metaphysics of Free Will and thereafter will do an internship at the Dutch Permanent mission at the United Nations in New York.

'Would You Press a Button that Kills All Psychopaths?' was written for the master's course 'Decision and Game Theory' with dr. H.C.K. Heilmann.

Notes

1. We do not need to worry about which of the many possible interpretations of utility we should endorse here. For our present purposes, this question does not matter since it is clear from the examples and counterexamples which action or option is preferred.

2. The term 'ratifiable' resulted from Jeffrey's idea that the agent can ratify a decision once it has been made. That is, his chosen action should have a maximal expected utility on the assumption that this is the action he is going to choose.

3. The term ratificationism was actually introduced by Jeffrey (1983: 19): 'Ratificationism requires performance of the chosen act, A, to have at least as high an estimated desirability as any of the alternative performances on the hypothesis that one's final decision will be to perform A'. But, to avoid confusion or different interpretations of the notion of ratificationism, I will only focus on the axiomatic notation of ratifiability conveyed by Egan.

4. This example shows us that imposing a ratifiability requirement will not help us to save CDT. It also shows that fans of EDT should take no comfort in the difficulties of this particular example against CDT. What we have here is definitely not an argument for a return to Evidential Decision Theory, since there are enough examples where EDT cannot be saved by RT either (Egan, 2007: 109-112).

5. One could argue that it is always better to have no guideline, than to have a wrong guideline. At least RT principle that counts all unratifiable actions as irrational will not deliver the bad recommendation that we got from the original version of CDT. But according to Egan (2007: 108) this does not do enough. He argues for 'completeness' where the correct theory of rational decision will endorse the rational action. Obviously, RT does not fulfill this requirement.

6. Notice that LRT takes the VAL_{EDT} instead of the value of VAL_{CDT}, which is used in the original RT.

7. One may wonder why this is called Lexical Ratificationism. As far as I know, this refers to the principle according to which entries in a dictionary are ordered, that is, the order depends on the first letter unless these are the same in which case it is the second which decides, and so on. I think this resembles the method where ratifiable actions are always to be preferred over unratifiable ones, but, within the categories, the action with greater VAL_{EDT} is to be preferred.

8. In the world of the Smoking Lesion, smoking is strongly correlated with lung cancer, but this correlation is understood to be the result of a common cause. There is a genetic lesion that tends to cause both smoking and cancer. The Smoking Lesion is an often used counterexample to both EDT and CDT.

9. I have not included a decision table for the Three-Option Smoking Lesion since such a table seems to be only more confusing. (Besides, a lot of spare paper was spilled finding out the right decision table). Instead, I think, a value table is more profitable for understanding the counterexample, so see table 1.

10. I have to admit that I do not find Gupta's counterexample very compelling. It seems peculiar that you would be better off smoking cigars when you are actually smoking cigarettes, and the other way around. I think such an odd thought would not be rational in the first place. However, this is not the focus of my paper. What is important is that the option not to smoke is irrational and should never be endorsed by any decision theorem.

11. Or, as an attentive reader might point out, to make sure that no options – even not option $\neg\psi$ - are ratifiable. He might say that if we un-ratify option $\neg\psi$, then we are at the same desirable point where we can treat all choices alike, and let the value according to EDT decide. But this is a mistake. Remember that all ratification mechanisms advance CDT. So, if there are no ratifiable options on the table, we have to make a decision according to CDT. CDT would tell us not to smoke since non-smokers tend to be best off. And so, CDT does not take into account the conditionality that, if we are to decide between option ψ_1 or ψ_2, we are best off smoking.

12. Note that TORT is principally similar to LRT, except for the fact that it incorporates two orders of ratifiability.

13. Egan (2007: 108-109) gives two constraints for the adequacy of theory of rational decisions: soundness and completeness. A theory is sound if, when it's irrational to perform X, the correct theory of rational decision will not endorse doing X. A theory is complete if it's rational to perform X, the correct theory of rational decision will endorse doing X. My TORT mechanism satisfies both requirements. TORT advises Paul to smoke (which is a rational decision) and does not advice Paul not to smoke (which is irrational in the case of Paul to do).

References

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

Gustafsson, J.E. (2011) 'A Note in Defence of Ratificationism'. In: *Erkenntnis* 75(1), 147-150.

Jeffrey, R.C. (1983) *The Logic of Decision* (2nd ed.). Chicago: University of Chicago Press.

Joyce, J.M. (1999) *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Nozick, R. (1969) 'Newcomb's Problem and Two Principles of Choice'. In: Nicholas Rescher (ed.) *Essays in Honor of Carl G. Hempel,* 114–146. Dordrecht: Reidel.