Graded Ratifiability

Abstract. How should you act when your actions themselves amount to evidence about what their consequences will be? This paper proposes that one should prefer actions with a greater *degree of ratifiability*. This proposal handles Andy Egan's alleged counterexamples to causal decision theory, as well as the cases that led Egan to reject the potential replacements for it that he considers. And it has more acceptable implications for many-option cases than a related suggestion from Ralph Wedgwood. Although the proposal faces some challenges raised elsewhere for Wedgwood, some of these are less serious than they seem, while others are problems for any view consistent with our intuitions about relevant examples. Perhaps you should not consider the evidential significance of your actions at all, in contrast to these intuitions. But if you should, you should do it as the present proposal instructs.

1. Introduction

Rational agents can be ignorant of matters of fact, including facts about their own motives and character. This does not mean that our knowledge of ourselves is limited to what can be inferred from our behavior, as it arguably is with others. But the non-behavioral access we have to ourselves is incomplete, and it is not clear that even what access we do have is rationally guaranteed, as opposed to being the result of some contingent faculty of inner sense. So just like with others, we can gain evidence about ourselves from our actions. Sometimes we surprise ourselves.

This has proved to be a continuing source of problems for theories of rational decision. The problems stem from the fact that we often care about the motives and character traits that determine our actions, or at least about their consequences. And some otherwise appealing theories yield apparently undesirable results in cases where we do.

The most famous examples are the smoking lesion and Newcomb counterexamples to **evidential decision theory (EDT)**. In these examples, one's actions provide evidence of traits that carry important consequences independently of what one does. Evidential decision theory famously advises us to maximize good news—to act in a way that provides evidence of traits that will yield benefits regardless of how we act. Since one should instead act to maximize good results, EDT must be rejected in favor of **causal decision theory (CDT)**. Or so the usual story goes.

But the problems do not end there. There are also cases where one's actions provide evidence about traits that influence what outcomes one's actions will produce. Andy Egan (2007) provides a recent example, which he attributes to David Braddon-Mitchell:

> **Psychopath Button:** Before you is a button marked "KILL ALL PSYCHOPATHS". You would like to rid the world of psychopaths, but not at the expense of killing yourself. You are confident enough that you are not a psychopath to make pressing rational, if not for one final detail: You are nearly certain that only a psychopath would press the button.

Egan claims, plausibly, that you should refrain from pressing the button. Instead of plowing ahead with your current confidence that you are not a psychopath, you should consider what pressing would reveal about yourself, and thus about the consequences of your actions. Your pressing would be excellent evidence that you are a psychopath. So you can be pretty certain that if you do press, then by doing so you will kill yourself. Intuitively, you should take this into account, and refrain.

Egan also claims, again plausibly, that CDT instead recommends pressing the button as rational. CDT holds that an option B is **rationally preferable** to an option A iff the **causally expected utility** of B-ing is greater than that of A-ing. Where the Ks are **dependence hypotheses**—i.e., maximal hypotheses about how outcomes depend causally on one's actions that form a partition—the causally expected utility of A-ing, or U(A), is defined as follows:

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

CDT gives the right results in Newcomb and smoking lesion cases, because it has the expected utility of one's options depend on unconditional credences about their effects. In contrast, EDT makes the mistake of letting it turn on one's conditional credences about what outcomes occur assuming one performs them, whether the outcome is the effect of the action or not.

Because CDT say one's unconditional credences are what matter, it seems committed to saying you should press. Since your unconditional credence that you are not a psychopath is sufficiently high, your pressing the button maximizes expected utility under CDT. It is only conditional on the assumption that you press that pressing fails to maximize causally expected utility. And CDT says that it is your unconditional credences that matter.

Now pretty much every part of the above will be controversial, especially the claims that you should refrain from pressing the button, and that if so CDT is false.[1] But this paper is not a paper about whether these claims are true, but instead what we should do about it if they are. In particular, I will assume that we should *somehow* take the evidential significance of our actions into account, in a way CDT does not allow. What I'll concern myself with is *how*.

I will start by explaining my proposal as it occurred to me, and how it differs from some that Egan considered but rejected. Then I'll compare my proposal to a related one independently developed earlier by Ralph Wedgwood (2013), and say why I favor my proposal over Wedgwood's.

2. Simple Ratifiability Proposals

Why does pressing seem irrational in Psychopath Button? Here is a rough suggestion that Egan considers: Pressing is **unratifiable**.

[1] See (Cantwell, 2010) and (Ahmed, 2012) for critical discussions of Egan's claim that one should refrain from pressing. And see (Joyce, 2010) for critical discussion of Egan's claim that CDT says otherwise.

I will explicate ratifiability in terms of what I call **conditional expected utility**. Let U(B|A) be the expected utility of B-ing conditional on the assumption that one As, in the following (stipulative) sense:

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

We can now define ratifiability as follows: An option A is ratifiable iff you have no other option O such that U(O|A) > U(A|A).

Egan considers a number of specific proposals as to how ratifiability is related to rational decision. He first considers a proposal which holds that it is always irrational to select unratifiable options. But this proposal faces trouble with cases in which no option is ratifiable, but in which one option nevertheless does seem rationally permissible, and indeed preferable to the other. Psychopath Button is itself plausibly such a case, since on the assumption that you refrain, pressing seems worth the risk. Here are two more:

> **Bottles 1:** You hold a bottle containing a mysterious liquid. You can switch it for a second bottle, or stay. After that, you must drink from whichever of the two bottles you are holding. A malevolent Predictor determined the contents of the bottles in the following way: If you were predicted to switch, then your current bottle contains water, while the other contains a fatal poison. If you were predicted to stay, then the other bottle contains water, while your current bottle contains a mild poison which will make you temporarily ill.
>
> **Lazy Death 1:** You are in Damascus, deciding whether to go to Aleppo. Death is in Aleppo, deciding whether to go to Damascus. Death is an infallible predictor, but a little bit lazy. If you are predicted to go to Aleppo, then Death will definitely wait there for you. If you are predicted to stay in Damascus, then Death will probably go to Damascus, though there is a chance he won't bother.

In Bottles 1, both switching and staying are unratifiable. But it seems at least rationally permissible (and I will later claim obligatory) to stay. Even though staying is conclusive evidence that you are staying with the mild poison, that seems better that switching, which is conclusive evidence that you are switching to the fatal poison. At least, that is what we should say if we think you should refrain in Psychopath Button.

It also seems at least permissible to stay in Death 1, even though both options are unratifiable. Even though staying is substantial evidence that you are staying to meet your death, that seems better than going, which is conclusive evidence that you are going to meet your death. At least, that is what we should say if we think you should refrain from pressing.

But Egan's first proposal says that unratifiable options are never permissible. In light of examples similar to these, Egan rejects it, and I think he was right to.

Egan next considers a proposal which holds that ratifiable options are always preferable to unratifiable options. But this faces trouble with a type of case that Egan attributes to Anil Gupta. Here is my own case of this type:

> **Boxes 1:** You must select one of three boxes, A, B, and C. An infallible Predictor determined their contents according to one of three schemes:
> S1: \$1 in A, \$0 in B, \$0 in C
> S2: \$0 in A, \$2 in B, \$3 in C
> S3: \$0 in A, \$3 in B, \$2 in C
> If you were predicted to select A, then scheme 1 was used, if B, then scheme 2, and if C, then scheme 3.

This case might be controversial, and we discuss many like it in Section 7. But it seems plausible to many, including Egan and myself, that one should prefer B and C to A. And at the very least, it seems that selecting A is not rationally mandatory. And yet, selecting A is the only ratifiable option. So Egan's second proposal must be rejected as well.[2]

Egan agrees that neither forbidding unratifiable actions, nor saying that ratifiable actions are always preferable to them, survives counterexamples. So although he sees sufficient reason to reject CDT, he is unsure what should replace it.

Before setting out my own proposal, I want to consider some further problems for ones like Egan's. Consider:

> **Bottles 2:** You hold a bottle containing a mysterious liquid. You can switch it for a second bottle, or stay. After that, you must drink from whichever of the two bottles you are holding. A benevolent Predictor determined the contents of the bottles in the following way: If you were predicted to switch, then your current bottle contains fatal poison, while the other contains water. If you were predicted to stay, then the other bottle contains fatal poison, while your current bottle contains an elixir granting immortality.
>
> **Lazy Death 2:** You are in Damascus, deciding whether to go to Aleppo. Death is in Aleppo, deciding whether to go to Damascus. But this time, Death is hoping to avoid you. Death is again an infallible predictor, but a little bit lazy. If you are predicted to stay in Damascus, then Death will definitely remain in Aleppo to avoid you. If you are predicted to go to Aleppo, then Death will probably go to Damascus, though there is a chance he won't bother.

In Bottles 2, you should stay. Or at least, that is what we should say if we say that you should stay in Bottles 1. And in Lazy Death 2, you should stay. Or at least, that is what we should say if we say you should stay in Lazy Death 1. But in these cases, all of your options are ratifiable. This raises a problem for any view that explains Psychopath Button merely by badmouthing unratifiable actions. For the preferability of staying in the above cases cannot be due simply to some problem with actions that are unratifiable.

[2] Note that this is consistent with holding that ratifiable options are always preferable to unratifiable options in cases where one has only two options. Indeed, the proposal I outline in Section 3 has this as a consequence.

3. Degrees of Ratifiability

There is a better way to incorporate ratifiability into the theory of rational decision. To get started, let us restrict our attention momentarily to cases in which one has only two options, A-ing and B-ing.

Notice that it follows from the definition of ratifiability above that A-ing is ratifiable iff U(A|A) - U(B|A) ≥ 0. This suggests a natural way of defining a graded conception of ratifiability. Let A-ing's **degree of ratifiability** be defined as U(A|A) - U(B|A). The greater the degree to which A is ratifiable, the greater the degree to which A-ing has greater expected utility than B-ing does, conditional on the assumption that one As. (Unratifiable options will thus have a negative degree of ratifiability.)

I propose that in two-option cases, the rationally preferable option is the one with the higher degree of ratifiability.

Let's see how GRT handles Psychopath Button. Because refraining has much higher causally expected utility than pressing does on the assumption that you press, U(press|press) - U(refrain|press) is not only negative, but far below zero. That is, pressing is not only unratifiable, it is unratifiable to a high degree. Because pressing has only somewhat higher causally expected utility than refraining does on the assumption that you refrain, U(refrain|refrain) - U(press|refrain) is still negative, but less so than for pressing. That is, although refraining is unratifiable, in graded terms it ratifiable to a higher degree than pressing is. So refraining is preferable to pressing.

I take this to capture our intuitions about Psychopath Button. Pressing seems irrational because, assuming one does so, one can expect doing so to have far worse effects than refraining would have had. In contrast, assuming that one refrains, one can expect doing so to have only moderately worse effects than pressing would have had.

And I think a similar story holds for the other cases considered above. In Bottles 1 and Lazy Death 1, switching and staying are both ratifiable to negative degrees, but switching is even more unratifiable than staying is. In Bottles 2 and Lazy Death 2, switching and staying are both ratifiable to a positive degree, but staying is even more ratifiable than switching. Regardless of whether both options are ratifiable or both unratifiable, the difference in degree of ratifiability tracks some natural intuitions about rational preferability.

How should we generalize this proposal to cases in which one has more than two options? One way is to define an option's overall degree of ratifiability in a way that generalizes to many-option cases, and say that one option is preferable to another if it has a greater overall degree of ratifiability. We will consider a recent proposal along these lines from Ralph Wedgwood below. But first, I want to suggest a different approach.

My proposal holds instead that the rational preferability of one option over another is determined by a pairwise comparison. It appeals to the notion of a **comparative degree of ratifiability**. The comparative degree of ratifiability of A-ing relative to B-ing is defined simply as U(A|A) - U(B|A), no matter what other options one has available.

Here is the proposal: B-ing is preferable to A-ing iff B-ing's comparative degree of ratifiability relative to A-ing is higher than A-ing's comparative degree of ratifiability relative to B-ing. That is, B-ing is preferable to A-ing iff U(A|A) - U(B|A) < U(B|B) - U(A|B). Call this **graded ratifiability theory (GRT)**.

GRT correctly says that in Boxes 1, you should prefer B and C to A. Since U(A|A) - U(B|A) = 1 - 0 < U(B|B) - U(A|B) = 2 - 0, B has a greater degree of comparative ratifiability than A. And likewise for C. At the same time, U(B|B) - U(C|B) = 2 - 3 = U(C|C) - U(B|C) = 2 - 3. So GRT correctly says that neither C nor B is preferable to the other.

We will consider a number of many-option cases in Section 7, where I will explain why I have GRT handle these cases as it does.

4. Managing the news

Because GRT has you consider the evidential significance of your actions, some might worry that it recommends an irrational policy of managing the news, as EDT allegedly does. But I think this is a misunderstanding.

GRT agrees with CDT in typical cases, where the dependence hypotheses are probabilistically independent of one's choice. In these cases, the conditional expected utility of an action will equal its unconditional expected utility. Thus U(A|A) - U(B|A) > U(B|B) - U(A|B) iff U(A) - U(B) > U(B) - U(A), which will be so iff U(A) > U(B). Importantly, this is true even in familiar counterexamples to EDT like smoking lesion and Newcomb cases. So GRT does not recommend managing the news in any of these cases.

Yet it might be objected that GRT recommends managing the news in cases where it disagrees with CDT, like those discussed in the preceding sections. It could be worried that our intuitions in these cases are misguided, just as common one-boxing intuitions are arguably misguided in the Newcomb case. In Bottles 1, for example, the objector might reject the intuition that one should stay, arguing as follows: There are two possible scenarios, a bad one where the bottle one drinks from contains a mild poison, and a worse one where the bottle one drinks from contains a fatal poison. And there is nothing one can do to affect which scenario is actual. The intuition that one should stay is just a misguided intuition that one should adopt an action that provides evidence that one is in the good situation already.

But I think this objection misdescribes the matter. Contrast Bottles 1 with the following:

> **Bottles 3:** You hold a bottle containing a mysterious liquid. You can switch it for a second bottle, or stay. After that, you must drink from whichever of the two bottles you are holding. A malevolent Predictor determined the contents of the bottles in the following way: If you were predicted to switch, then both bottles contain a fatal poison. If you were predicted to stay, then both bottles contain a mild poison which will make you temporarily ill.

In Bottles 3, you really don't have any control over what happens. There is a bad scenario where both bottles contain mild poison, and a worse one where they both contain fatal poison—and so nothing you do will affect what you end up drinking. So GRT would be guilty of recommending a policy of managing the news if it said staying was preferable. But it does not say this. It says correctly that neither switching nor staying is preferable to the other, since you know that both actions have the same effects.

The difference in Bottles 1 is that while you do not control which situation you are in, you do control what you drink, because you control which bottle you drink from. For this reason, CDT can itself recommend switching, depending on your unconditional credences about which situation you are in. Like CDT, GRT cares about your credences regarding the effects of your actions. The disagreement concerns which credences are important. CDT says that it is your current, unconditional credences, before you find out which action you will adopt. GRT says it is your conditional credences, which take into account the evidence your actions themselves might provide about your situation, and thus what their effects will be. When considering a highly unratifiable action, you shouldn't act first, and regret it later, when you realize what your selection suggests you have just done. You should prospectively take into account he additional evidence about your situation that your actions will provide.

An additional example will further reinforce the point:

> **Envelopes:** You hold an envelope, and have the option of switching it for another envelope. If you were predicted to stay, your envelope contains $2 while the other contains $0. If you were predicted to switch, then your envelope contains $100, while the other contains $101.

Here GRT says to stay. For staying is the option such that assuming one does it, it can be expected to most improve upon the alternative. But staying is bad news. It is decisive evidence that the envelopes contain little money. So GRT correctly recommends maximizing good results rather than good news.[3]

5. Wedgwood's Benchmark Theory

Ralph Wedgwood (2011) has proposed a theory that he, following Ray Briggs, calls **Benchmark Theory (BT)**. In simple cases and given certain simplifying assumptions, BT delivers the same results as GRT. But relax these assumptions and BT runs into trouble. In this section and the next, I'll discuss the agreement, then in Section 7, the trouble.

Wedgwood's idea is that we evaluate an option's performance under a dependence hypothesis by looking at how it compares to a **benchmark**, which represents the default level of value under that hypothesis. Each option is then assigned a **comparative value** under the dependence hypothesis, which represents how much that action over-performs or under-performs relative to that dependence hypothesis's benchmark. Where $B(K_j)$ is the benchmark for a dependence hypothesis K_j, the comparative value of of A-ing under K_j, or $CV(A,K_j)$, is simply $v(K_jA) - B(K_j)$.

BT says that the rationally optimal action is the one with the highest **evidentially expected comparative value**, which is determined by taking a weighted average of its comparative value under each dependence hypothesis. Importantly, the weighting follows one's conditional credence in the dependence hypotheses given that one performs the relevant action, rather than one's unconditional credences. Thus the evidentially expected comparative value of A-ing, or EECV(A), is:

[3] Cf. Bassett (2015), Section 4.1, which claims a related example as a counterexample to Wedgwood 2013. I think it is not a counterexample, because if one should stay in Envelopes, then one should select box A in Bassett's example, contrary to what Bassett claims.

$$\sum_K Cr(K \mid A)CV(A,K).$$

We are ready to prove that BT and GRT agree about certain cases under certain assumptions. The cases are ones in which an agent has available only two options, A-ing and B-ing. And the assumption is that the benchmarks in such cases should be determined by averaging, such that for any dependence hypothesis K_j,

$$B(K_j) = \frac{v(K_jA) + v(K_jB)}{2}.$$

Since by stipulation $CV(A,K_j) = v(K_jA) - B(K_j)$, it follows that

$$CV(A,K_j) = v(K_jA) - \frac{v(K_jA) + v(K_jB)}{2},$$

and thus

$$CV(A,K_j) = \frac{v(K_jA) - v(K_jB)}{2}.$$

Since BT says to maximize evidentially expected comparative value, it says that B-ing is preferable to A-ing iff

$$\sum_K Cr(K \mid A)\left[\frac{v(KA) - v(KB)}{2}\right] < \sum_K Cr(K \mid B)\left[\frac{v(KB) - v(KA)}{2}\right],$$

which reduces to

$$\sum_K Cr(K \mid A)\left[v(KA) - v(KB)\right] < \sum_K Cr(K \mid B)\left[v(KB) - v(KA)\right].$$

By algebra, this is equivalent to

$$\sum_K Cr(K \mid A)v(KA) - \sum_K Cr(K \mid A)v(KB) < \sum_K Cr(K \mid B)v(KB) - \sum_K Cr(K \mid B)v(KA),$$

which, by the definition of conditional expected utility from Section 2 above, is equivalent to the condition that U(A|A) - U(B|A) < U(B|B) - U(A|B). So in cases involving only two options, and when the benchmarks are set by averaging, Wedgwood's BT is equivalent to GRT's claim that one should prefer the option with the higher degree of ratifiability.

6. Objections to BT and GRT

Because of the partial agreement between BT and GRT, some existing objections to BT apply to GRT.

First, BT and GRT are both incompatible with the **Independence of Irrelevant Alternatives (IIA)**, a principle holding roughly that giving agents new options shouldn't affect their preferences among existing options. Wedgwood thinks this is no problem, but Bassett (2015) disagrees. I am with Wedgwood.

Bassett illustrates the idea behind IIA with an apocryphal story about Sidney Morgenbesser. A waitress offers blueberry and apple pie as desert options, and Morgenbesser orders apple. Then she returns to say that cherry pie is also available. Morgenbesser responds, "In that case, I'll have blueberry." This particular violation of IIA strikes us as irrational. Whether Morgenbesser has cherry pie available shouldn't affect whether he prefers apple to blueberry.

But as Bassett concedes, not all violations of IIA seem intuitively irrational. Sometimes what options one has available can amount to relevant evidence. When invited for tea by an acquaintance, straight-laced Amartya prefers accepting tea to declining. But when invited for tea or cocaine by the same acquaintance, he prefers to decline altogether. This strikes us as rational because the fact that the acquaintance offered cocaine provides Amartya with additional evidence about what taking tea with him would be like.[4]

This shows that IIA must be qualified in some way. But it doesn't yet show the need to qualify it so much that it ends up being consistent with BT and GRT. The acquaintance's offer of cocaine is evidence bearing directly on what effects Amartya's actions will have. Qualifying IIA so it doesn't apply in such cases won't be enough to reconcile it with GRT. But even so, there is a related qualification that will be enough. Consider:

> **Perilous Seat:** Galahad is deciding whether to sit at the Perilous Seat at the Round Table. Galahad would like to sit, but he knows any unworthy knight who does so will die. Galahad is confident enough that he is worthy that he is ready to sit. But just then, he learns that Percival has been captured. Rescuing Percival is the last thing Galahad wants to do today. But he is certain that a worthy knight would drop everything and rescue Percival immediately.

When he gains the new option of rescuing Percival, it is intuitively irrational for Galahad to sit in the Perilous Seat, and rational to prefer to refrain. This is so even though Galahad's having this new option provides no direct evidence regarding the effects of his initial options of sitting or refraining. Instead, it affects the evidential value of his selecting one of those options. Conditional on his selecting either of them, it is now likely that sitting will kill him. This is because he will have passed over the option of rescuing Percival immediately, which he is certain any worthy knight would adopt.

We can even embellish the Moganbesser case to make his change in preferences intuitively rational. Suppose he especially likes the taste of blueberry pie, but orders apple pie instead because he is worried about having a lesion with the following bizarre effects: first, it makes

[4] The example is from Sen 1993.

one highly allergic to blueberries, and second, it makes one irresistibly crave cherries. If so, it is plausible that he can rationally order apple initially, and then rationally change his order to blueberry after getting the additional option of cherry. For his passing over cherry would be strong evidence that he does not have the lesion, and so can safely eat blueberries.

The intuitions I'm pushing regarding these cases are sure to be controversial. But I think they are as plausible as the intuitions that motivate BT and GRT, such as that you should refrain from pressing in Psychopath Button. And qualifying IIA to allow these exceptions plausibly will be enough to reconcile it with GRT, though there are remaining problems for BT discussed in Section 7 below.

Another objection to BT concerns its incompatibility with *prima facie* attractive dominance principles like the following:

> **Weak Dominance:** One should prefer A-ing to B-ing if for every dependence hypothesis K such that Cr(K) > 0, v(KA) ≥ v(KB), and for some K, v(KA) > v(KB).

As Bassett observes, the restriction of Weak Dominance to dependence hypotheses with nonzero credence is crucial. Suppose I am certain that an old lottery ticket did not win, and thus am certain that I have nothing to gain by taking it. Even if I also have nothing to lose, it seems permissible for me not to take the ticket. Even if it is metaphysically possible that taking it will earn me $1,000,000, I am certain that this metaphysically possible scenario does not obtain. This widely accepted restriction on Weak Dominance principles will be important shortly.

Weak Dominance raises a number of problems for BT, but many turn on idiosyncrasies that GRT doesn't share.[5] The general problem for both BT and GRT is the one raised by Ray Briggs (2010, Section 6). Briggs has us consider a principle closely related to Weak Dominance, which Briggs compares to a requirement of Pareto optimality involving votes among one's possible future selves. Here it is in my notation. Suppose that two of your options, A and B, are such that (i) for any option O among those available, U(A|O) ≥ U(B|O), and (ii) for some option O among your options, U(A|O) > U(B|O). Briggs' principle holds that if so, then A-ing is preferable to B-ing. Briggs then argues that any decision theory meeting another apparent desideratum—one which GRT does meet—is inconsistent with the Pareto optimality requirement. Briggs concludes that this is a decision-theoretic paradox.

Briggs proves that the Pareto requirement entails Weak Dominance. Here is a quick proof. If for every K, v(KA) ≥ v(KB), then for every O, U(A|O) ≥ U(B|O). And if for some K, Kj, v(KA) > v(KB), then for some O, U(A|O) > U(B|O), because Kj will have nonzero probability under some O. Thus Weak Dominance follows from the Pareto requirement by an application of strengthening the antecedent.

But despite its *prima facie* appeal, I think Briggs's Pareto requirement should be rejected. Consider:

[5] See Bassett 2015, Sec. 3.5.

> **Boxes 2:** You must select one of three boxes, A, B, and C. An infallible Predictor determined their contents according to one of three schemes:
>
> S1: \$0 in A, \$3 in B, \$2 in C
> S2: \$4 in A, \$3 in B, \$3 in C
>
> If you were predicted to select A, then scheme 1 was used, and if B or C, then scheme 2.

Briggs's Pareto requirement and Weak Dominance both say that selecting B is preferable to selecting C. However, GRT says neither is preferable to the other. Assuming you select B, boxes B and C contain the same amount. And assuming you select C, they contain the same amount. So, U(B|B) - U(C|B) = U(C|C) - U(B|C) = 0.

Anyone who accepts the ratificationist intuitions that motivate GRT, such as that you should refrain in Psychopath Button, should be willing to accept this result, and reject Briggs' Pareto principle and Weak Dominance. The crucial question is this. When comparing B to C, should we be concerned with any dependence hypotheses with nonzero *unconditional* credence, or only with those with nonzero credence *conditional* on one's selecting either B or C? Briggs' principle and Weak Dominance say it is unconditional credences that matter. Since B pays more than C in scheme 1, and since each options must (I assume) have nonzero credence, you should have a nonzero unconditional credence that B will pay more than C. This is why B weakly dominates C. But ratificationists think conditional credences are what matter. Assuming you select either B or C, they certainly contain the same amount. So you can be certain you will not get more money by selecting B than you would have by selecting C. The ratificationist should consider that sufficient for B not to be preferable to C.

7. Objections to BT

We saw in Section 5 that GRT and BT agree in cases with only two options, and when the benchmarks are set by averaging. But the two theories disagree in other ways that I think favor GRT.

No matter how many options are available, under GRT which of any given two options is preferable is determined by a direct comparison between them. All that matters is the outcomes they each are likely produce under the assumption that you select the first option, and the outcomes under the assumption that you select the second. Yet under BT, which of two options is preferable depends instead on how each compares to a benchmark that is determined in part by the outcomes of other options. And this allows extraneous options to exert counterintuitive influence.

Before getting on to my main objection, I will consider a related one that Wedgwood anticipates, which might be called the **dreadful options problem**. At any given time, one has available many courses of action that are "perfectly dreadful," as Wedgwood puts it. I can go to the movie, or to the show, or simply pound my head against a wall all evening. Including the dreadful options among one's options does no harm in standard expected utility theory. But under BT, it can affect the relevant benchmarks, and thus affect which of one's non-dreadful options are permissible. The problem is that it seems unacceptable to allow the availability of dreadful options to affect one's decisions in the way BT seems to license.

Wedgwood's solution is simply to have dreadful options set aside prior to deliberation, and thus not considered in setting benchmarks. This of course raises the difficulty of determining which options are dreadful enough to be set aside. But suppose these difficulties can be handled. Even if so, I think it does not get to the root of the problem. For there is a closely related **redundant options problem**. Compare:

> **Boxes 3:** You must select one of two boxes, A and B. Their contents were determined by an infallible Predictor, according to one of two schemes:
> S1: \$3 in A, \$1 in B
> S2: \$1 in A, \$4 in B
> If you were predicted to select A, then scheme 1 was used. If B, then scheme 2.
>
> **Boxes 4:** You must select one of three boxes, A, B, and C. Their contents were determined by an infallible Predictor, according to one of two schemes:
> S1: \$3 in A, \$1 in B, \$1 in C
> S2: \$1 in A, \$4 in B, \$4 in C
> If you were predicted to select A, then scheme 1 was used. If B or C, then scheme 2.

The difference between these cases is that in Boxes 4, you have the option to select C. But this option is redundant. B and C contain the same amount of money no matter what, and your having C available has no untoward effects on your evidential situation. So intuitively, the presence or absence of C shouldn't affect your preference between A and B.

However, BT says otherwise. Here is why. Anyone sympathetic to ratificationism will say that in Boxes 3 you should select B. And BT has no trouble delivering this verdict, no matter how the benchmarks are set.[6] The problem for BT is Boxes 4. If the benchmarks are set via averaging, then BT says you should select A.[7] This in itself is counterintuitive. But what is worse is the consequence that whether you prefer A to B should depend on whether another option just like B is available. Indeed, even if we think averaging is merely a permissible method for setting benchmarks, it seems wrong to say that selecting A is even permissible in Boxes 4 without being permissible in Boxes 3.

To reinforce the point, consider an embellishment of Boxes 4. You are initially prepared to select A, as BT says is permissible. But just then you realize that Box C is too far away for you to reach, and so you instead must select B. As you prepare to select B, you then realize that it could be reached with either your right hand or your left hand, thus presenting you with two options that involve selecting B. Meanwhile, A is far enough to your left that it can be reached only with your left hand. And so you again feel permitted to select A. Switching your preferences back and forth for these reasons seems irrational, but BT says otherwise.

The problem here is not merely that BT violates a general IIA principle. We saw in Section 6 above that some exceptions to IIA are intuitively plausible. Rather, the problem I am pressing for BT is that it has counterintuitive results in particular cases. The trouble with

[6] E.g., if benchmarks are set via averaging, then $EECV(A) = 1(1) + 0(-1.5) = 1 < EECV(B) = 0(-1) + 1(1.5) = 1.5$.

[7] $EECV(A) = 1(4/3) + 0(-2) = 4/3 > EECV(B) = 0(-2/3) + 1(1) = 1$.

general principles like IIA is that it is hard to anticipate what implications they will have in new and unusual cases. So I'll forgo an attempt at formulating some general principle that BT is incompatible with, and stick to the particular cases.

Perhaps the benchmark theorist could avoid this problem by employing more coarse-grained individuation conditions for options, which treat selecting B or C as a single option, at least for purposes of setting benchmarks. But this solution is insufficiently general. Consider:

> **Boxes 5:** You must select one of three boxes, A, B, and C. Their contents were determined by an infallible Predictor, according to one of four schemes:
> S1: \$3 in A, \$0 in B, \$2 in C
> S2: \$3 in A, \$2 in B, \$0 in C
> S3: \$1 in A, \$3 in B, \$5 in C
> S4: \$1 in A, \$5 in B, \$3 in C
> This time, the Predictor used a chancy method to determine which scheme was used. A fair coin was tossed. If you were predicted to select A, then scheme 1 was used if the coin landed heads, and scheme 2 if tails. If you were predicted to select B or C, then scheme 3 was used if the coin landed heads, and scheme 4 if tails.

If you did not have the option to select C, then BT would say that you should select B.[8] So far, so good. But with the option to select C available, and with benchmarks set via averaging, BT says you should select A.[9] This commitment seems no more plausible than it did regarding Boxes 4 above. And this time, there is no avoiding the commitment by adopting coarse-grained individuation conditions for options. Since you know that selecting B and selecting C will lead to different outcomes, any plausible criteria for individuating options will count them as distinct.

Perhaps the BT theorist could propose instead that benchmarks be set not by averaging over options, but instead over equivalence classes of options. But this reduces BT's explanatory power. Presumably our decision theory should explain why two options are equivalent, such as B and C in Boxes 5. If these options must be grouped together from the outset for BT even to be applied to the case, then it looks like BT presupposes their equivalence instead of explaining it.

Let's try another tack. So far I have been assuming that benchmarks are set via averaging. Wedgwood regards this as a reasonable method, at least for two-option cases. But perhaps it could be dispensed with, in favor of other methods. According to Wedgwood, there are a variety of other reasonable methods for setting benchmarks. One might use the **relief method**, and set the benchmark for a dependence hypothesis at the worst possible outcome under that hypothesis. Or one might use the **regret method**, and set the benchmark at the best possible outcome. Wedgwood ultimately favors a kind of permissivism, under which each of these two methods, as well as intermediate methods including averaging, all are permissible. Wedgwood's only restrictions are that for a given decision the same method be

[8] With C unavailable, and with benchmarks set via averaging, EECV(A) = .5(1.5) + .5(.5) + 0(-1) + 0(-2) = 1 < EECV(B) = 0(-1.5) + 0(-.5) + .5(1) + .5(2) = 1.5.

[9] EECV(A) = .5(4/3) + .5(4/3) + 0(-2) + 0(-2) = 4/3 > EECV(B) = 0(-5/3) + 0(1/3) + .5(0) + .5(2) = 1.

employed regarding each dependence hypothesis, and that the method assign benchmarks no lower than relief and no higher than regret.

But I think the introduction of these additional methods raises additional problems without solving the existing ones. Start with the new problems. As Wedgwood comes close to acknowledging, permissivist BT does not fully accommodate our intuitions about even the cases it is designed to handle. Intuitively, it seems that you should not press in Psychopath Button. And likewise, it is intuitively attractive to say that you should not switch in Bottles 1. But permissivist BT says that these actions are permissible, since benchmarks may be set via relief. Conversely, it seems that one should not switch in Bottles 2. But again, permissivist BT says it is permissible, since benchmarks may be set via regret. Wedgwood acknowledges that these methods are "extreme," but does not consider the actions they license impermissible on that account. Since they make BT unable to fully capture our intuitions about these core cases, this undercuts BT's motivation.

Perhaps the verdicts I'm pushing regarding Psychopath Button and Bottles 1 and 2 could be rejected. Even so, the redundant options problem remains. Start with the regret method, where the benchmark for a dependence hypothesis is set to the best possible outcome under that hypothesis. Consider:

> **Boxes 6:** You must select one of three boxes, A, B, and C. There are three schemes by which money might have been distributed among them:
> S1: \$1 in A, \$0 in B, \$0 in C
> S2: \$0 in A, \$1 in B, \$100 in C
> S3: \$0 in A, \$100 in B, \$1 in C
> The money was distributed by a Predictor in a chancy fashion. If selecting A was predicted, then scheme 1 was probably used. If B, then probably scheme 2. And if C, then probably scheme 3. But the chanciness of the distribution means that even if A was predicted, schemes 2 and 3 each stood a 5% chance of being adopted. And likewise if B or C was predicted.

I say that you should select B or C. But if the benchmark is set via regret, then BT says you should select A.[10] Even if we regard regret as only a permissible way to set benchmarks, BT still is committed to saying that selecting A is permissible. Indeed, Wedgwood defends this commitment regarding a related case from Briggs (2010). What is more, this commitment does not depend on our using a pure regret method of setting benchmarks. Any method assigning any weight to regret, including averaging, will have the same commitments, and face the same problems. (If the weight assigned to regret is low, then the dollar amounts must be adjusted to compensate.)

But it is implausible that selecting A is even permissible. Among other reasons, this faces the redundant options problem. Suppose that just as you prepare to select A, you realize that C is out of reach. Intuitively, this should not matter to your preferences between A and B. But BT again says it does matter. For if A and B were your only options, then the evidentially expected comparative value for selecting A suddenly would be lower than that for B.[11] So BT would say, implausibly, that selecting A suddenly becomes impermissible

[10] EECV(A) = .9(0) + .05(-100) + .05(-100) = -10 > EECV(B) = .05(-1) + .9(-99) + .05(0) = -89.15

[11] EECV(A) = .9(0) + .05(-1) + .05(-100) = -5.05 < EECV(B) = .05(-1) + .9(0) + .05(0) = -.05.

when C is no longer available. To be sure, I agree that selecting A would be impermissible if your options were restricted to A and B. Where BT goes wrong is in saying that in Boxes 6, where C is available, it is permissible to select A.

So far I have been pushing the redundant options problem as the main problem for BT. But the problems don't end there. Consider the **evidential sweetening problem**. The problem is that there are potential changes in your evidential situation that intuitively should sweeten the prospects of selecting A in Boxes 6, but which BT treats instead as reasons to prefer B. Suppose that as you prepare to select A, you learn some new information. This time, it is that the Predictor was unavailable for making predictions today. Contrary to what you had thought, the money was simply distributed via a chancy process where scheme 1 stood a 90% chance of being selected, and schemes 2 and 3 each a 5% chance. All standard expected utility theories, including BT, now say that you should prefer B (as well as C) to A. But this new information seems like a terrible reason to prefer B to A. For its effect is simply to raise some of the relevant conditional and unconditional probabilities that scheme 1 was used. And scheme 1 is the one where A contains more money than B.

A related **outcome sweetening problem** is raised by the following:

> **Boxes 7:** You must select one of three boxes, A, B, and C. There are three schemes by which money might have been distributed among them:
> S1: \$1 in A, \$0 in B, \$0 in C
> S2: \$0 in A, \$1 in B, \$1 in C
> S3: \$0 in A, \$1 in B, \$1 in C[12]
> The money was distributed by a Predictor in the same chancy fashion as in Boxes 6.

BT says—I think plausibly—that you should prefer B (and C) to A.[13] But look what has changed from Boxes 4. In Boxes 4, if you select B, you can regard it as 5% likely that you'll get \$100. In Boxes 5, you are guaranteed to get no more than \$1. This is not a reason in favor of preferring B to A. To be sure, another change is that if you select B in Boxes 6, you should think it likely that you would have been better off selecting C. That may be a reason to prefer C to B, although one that is counterbalanced by opposing reason on the other side. But it is hard to see it as a reason to prefer A to B.

It will not help to switch to a pure relief method, in which the worst possible outcome under a dependence hypothesis is used for the benchmark. Consider:

> **Boxes 8:** You must select one of three boxes, A, B, and C. There are three schemes by which money might have been distributed among them:
> S1: \$100 in A, \$101 in B, \$101 in C
> S2: \$101 in A, \$100 in B, \$0 in C
> S3: \$101 in A, \$0 in B, \$100 in C

[12] I count S2 and S3 as distinct for easier comparison to Boxes 4, but nothing turns on this notational convenience.

[13] With benchmarks set via regret, EECV(A) = .9(0) + .05(-1) + .05(-1) = -.1 < EECV(B) = .05(-1) + .9(0) + . 05(0) = -.05.

> The money was distributed by a Predictor in the same chancy fashion as in Boxes 6 and 7.

This time, I think you should select A. But if the benchmarks are set by relief, then BT says you should select B or C.[14] Even if we are permissivists about how to set benchmarks, so long as the relief method is permissible, BT is committed to saying that selecting B or C is permissible. And this commitment does not depend on our using a pure relief method of setting benchmarks, though I will use this method for simplicity. Any method assigning any weight to relief (including averaging) will have the same commitments, and face the same problems. (If the weight assigned to relief is low, then again the dollar values must be adjusted.)

BT's commitment that selecting B or C is at least permissible raises the usual problems. Consider first the redundant options problem. Suppose that just as you prepare to select B, you realize that C is out of reach. BT has the counterintuitive implication that this should change your preference to A.[15] Now I agree that you should select A were C to become unavailable. But I also think you should select A when C is available. Where BT goes wrong is in denying this, and saying that it is permissible to select B or C in Boxes 8.

Turn now to the evidential sweetening problem. Suppose once more that you are prepared to select B, which BT says is at least permissible. Just before you do, you learn that the Predictor was unavailable today, and the money was simply distributed via a chancy process where scheme 1 stood a 90% chance of being selected, and schemes 2 and 3 each a 5% chance. All standard expected utility theories, including BT, now say that you should select A. But this new information seems like a terrible reason to switch from B to A. For its effect is simply to raise some of the relevant conditional and unconditional probabilities that scheme 1 was used. And scheme 1 is the only one where B contains more money than A.

For the outcome sweetening problem, consider:

> **Boxes 9:** You must select one of three boxes, A, B, and C. There are three schemes by which money might have been distributed among them:
> S1: \$100 in A, \$101 in B, \$101 in C
> S2: \$101 in A, \$100 in B, \$100 in C
> S3: \$101 in A, \$100 in B, \$100 in C[16]
> The money was distributed by a Predictor in a chancy fashion. If selecting A was predicted, then scheme 1 was probably used. If B, then probably scheme 2. And if C, then probably scheme 3. But the chanciness of the distribution means that even if A was predicted, schemes 2 and 3 each stood a 5% chance of being adopted. And likewise if B or C was predicted.

[14] EECV(A) = .9(0) + .05(101) + .05(101) = 10.1 < EECV(B) = EECV(C) = .05(1) + .9(100) + .05(0) = 90.05.

[15] Assuming A and B are your only options, EECV(A) = .9(0) + .05(-1) + .05(-100) = -5.05 < EECV(B) = .05(-1) + .9(0) + .05(0) = -.05.

[16] Again, I list schemes 2 and 3 separately for easier comparison with Boxes 8.

Here BT says that you should select A.[17] This much is plausible. But what is not plausible is that you should select A in Boxes 9 but not Boxes 8. For look at what has changed. In Boxes 8, there is a chance that B contains \$0 rather than \$100. Here in Boxes 9, B definitely contains \$100. This is if anything a further consideration in favor of preferring B to A.

8. Objections to GRT

Arif Ahmed (2012) raises an objection to GRT that is in my view more worrying than some of the others. It makes me worry that the intuitions that motivated GRT in the first place, such as that you should refrain in Psychopath Button, perhaps should be rejected. But assuming we should hang on to them, I think Ahmed's objection gives us little reason to reject GRT in favor of any other view upholding these intuitions.

To start, consider:

> **Psychopath Lever:** Before you is a lever marked "KILL ALL PSYCHOPATHS". Your preferences are the same as in Psychopath Button, and you are as confident as ever that you are not a psychopath. The only differences are that it costs a nickel to pull the lever, and that for whatever reason you consider pulling the lever no evidence at all regarding whether one is a psychopath.

In Psychopath Lever, it is rational to pull. It is stipulated in Psychopath Button that it would be rational to press if not for the evidence pressing would give you of your own psychopathy. Indeed, that's the *whole point* of the example. And I hereby stipulate that the cost of a nickel is too small to make any difference.

But now consider:

> **Psychopath Button and Lever:** Before you are a button and a lever, each marked "KILL ALL PSYCHOPATHS". Your preferences and credences are as before. But this time, there is no refraining. You *must* choose one of them.

You should press the button. You know that pulling and pressing have exactly the same effects, no matter whether you are a psychopath. So you might as well save yourself a nickel.

We have just said that you should prefer pressing the button to pulling the lever when those are your options, and prefer pulling the lever to refraining when those are. But GRT says that you should prefer refraining to pressing the button when these are your options, as in Psychopath Button. Thus GRT says your preferences should be pair-wise intransitive.

Now this alone might be a serious problem for GRT, if we are committed to pair-wise preferences being transitive. But if so, it is a problem for *any* view that says you should refrain in Psychopath Button. Perhaps a commitment to pair-wise transitivity gives us a reason to reject the intuition that you should refrain. But assuming we stick to our intuition, it gives us no reason to give up GRT in favor of some other theory that upholds this

[17] With benchmarks set via relief, $\text{EECV}(A) = .9(0) + .05(1) + .05(1) = .1 > \text{EECV}(B) = .05(1) + .9(0) + .05(0) = .05$.

intuition. For in that case we are bound to give up pair-wise transitivity, no matter which particular theory we accept.

But Ahmed thinks there is a more pressing problem in the vicinity, which is arguably more particular to GRT. Consider:

> **Psychopath Dilemma:** Before you are a button and a lever, each marked "KILL ALL PSYCHOPATHS". Your preferences and credences are as before. But this time, you also have the option to refrain.

Here GRT says that pushing the button is preferable to pulling the lever, which is preferable to refraining, which is preferable to pushing the button. Thus each option has some other option that is preferable to it. And this, Ahmed says, saddles us with the intolerable consequence that no option is rationally permissible. That is, it commits us to the existence of a **rational dilemma**.

In contrast, BT arguably avoids this consequence. Since pressing strongly dominates pulling, pulling gets ruled out of consideration from the outset. And since refraining is preferable to pressing, refraining is rationally permissible.

But if this is what BT says about Psychopath Dilemma then I think it is a bug rather than a feature. Compare:

> **Boxes 10:** You must select one of three boxes, A, B, and C. There are two schemes by which money might have been distributed among them:
> S1: \$3 in A, \$4 in B, \$5 in C
> S2: \$3 in A, \$4 in B, \$0 in C
> If a fair coin landed heads, then S1 was used. If tails, then S2.

Even though selecting B strictly dominates selecting A, there still is a fact of the matter as to which is preferable between A and C—namely, that A is preferable. An adequate decision theory should explain this. But BT, by excluding strictly dominated actions from consideration, seems unable to.

Correspondingly, there is a fact of the matter as to whether pulling the lever is preferable to refraining in Psychopath Dilemma. Pulling is preferable, since the *whole point* of the Psychopath Button example requires it to be. This is uncontroversial in Psychopath Lever. But the reasons for preferring pulling to refraining in Psychopath Dilemma seem no different. The case for pulling over refraining is unaffected by whether one has the further option to press the button. The upshot is that pulling the lever is preferable to refraining in Psychopath Dilemma, and it is a problem for BT if it cannot accommodate this fact.

Moreover, I think it is intuitively appealing to say that in Psychopath Dilemma, each option has another that is preferable to it. We have just seen why pulling the lever is preferable to refraining. And surely pressing the button is preferable to pulling the lever. All that remains is that refraining is preferable to pressing the button. Ahmed thinks we should deny this, and say that pressing is preferable to refraining, both in Psychopath Dilemma and Psychopath Button. But it can be hard to swallow that one should press in Psychopath Button. You know that if you do, then you almost certainly will kill yourself! And this isn't just true in Psychopath Button. It is also true in Psychopath Dilemma. So if we think there is sufficient reason to prefer refraining to pressing in Psychopath Button, we seem stuck saying the same thing in Psychopath Dilemma, when the additional option of pulling the lever is present.

So should we say that Psychopath Dilemma is a genuine rational dilemma? Maybe. I for one agree with Ahmed that it is unappealing to countenance rational dilemmas.[18] But not everyone is as allergic to dilemmas as he and I are. And Psychopath Dilemma is as plausible a case as any of a genuine dilemma.

At the same time, I think we might avoid a commitment to dilemmas by complicating the relationship between rational preferability and permissibility. What does it mean for some option A to be rationally preferable to B? It means at least that one's reasons favor A over B, and that one should adopt a corresponding attitude of preferring (or favoring) A over B. Now if A and B are your only options, then presumably this means among other things that you ought to adopt A, and hence ought not to adopt B. And it is natural to assume more generally that in *any* case where A is preferable to B, then you ought not to adopt B. But that does not follow automatically from the idea that one's reasons favor A over B, or that one should in some sense prefer A to B. For in cases like Psychopath Button, it might be that one's reasons favor B over a third option C, and favor C over A. If so, we *might* claim that one is in a dilemma where one ought not to adopt any one of these options. But we also might claim that one is permitted to throw up one's hands and just pick something. Neither of these claims seems intuitively crazy to me, and I don't think either is forced on us immediately by accepting a distinct claim about what one's preferences should be.

9. Conclusion

Examples like Psychopath Button motivate the thought that when one's actions amount to evidence about what their effects will be, you should not wait around until after acting to consider it. You should take this potential new evidence into account when deciding. For instance, you should not decide to press a button that will kill all psychopaths if you are fairly certain that only a psychopath would press it, even if you otherwise consider yourself unlikely to be a psychopath. This intuition raises difficulties, though I have claimed less serious ones than some critics allege. But in any case, my central claim is the conditional that if we uphold this core intuition, then we should accept GRT. Unlike the absolute ratifiability proposals considered by Egan, GRT has one's preferences follow the *degree* to which an action is ratifiable. This yields more intuitive verdicts about cases like Bottles 1 and 2, Lazy Death 1 and 2, and Boxes 1. And unlike BT, it has one's preferences follow pair-wise comparisons in many-option cases, rather than comparisons with overall benchmarks. This gives us more appealing verdicts about cases like Boxes 2-9. Perhaps you should not consider the evidential significance of your actions at all, in contrast to intuitions about Psychopath Button and related cases. But if you should, you should do it as GRT instructs.[19]

[18] For discussion, see Barnett MS.

[19] For helpful comments and discussion, I wish to thank Arid Ahmed, Ralph Wedgwood, and Jonathan Weisberg.

References

Ahmed, Arif (2012) 'Push the Button' *Philosophy of Science* 79(3): 386-395.

Barnett, David James (MS) 'Internalism, Stored Beliefs, and Forgotten Evidence'

Bassett, Robert (2015) 'A Critique of Benchmark Theory' *Synthese* 192(1): 241-267.

Briggs, Ray (2010) 'Decision-Theoretic Paradoxes as Voting Paradoxes' *Philosophical Review* 119(1): 1-30.

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

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

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

Sen, Amartya (1993) 'Internal Consistency of Choice' *Econometrica* 61(3): 495-521.

Wedgwood, Ralph (2011) 'Gandalf's Solution to the Newcomb Problem' *Synthese* (14): 1-33.