Graded Ratifiability

Abstract. An action is unratifiable when, on the assumption that one performs it, another option looks better. A common thought is that unratifiable actions are somehow rationally defective. But Andy Egan has brought renewed attention to cases where, although both of one's options are unratifiable, one option nevertheless seems preferable to the other. I claim that to handle such cases, we must invoke a graded notion of ratifiability. One option is preferable to another when it is unratifiable to a lesser degree. After motivating and explaining the proposal, I turn to a related view from Ralph Wedgwood. I claim that what Wedgwood's proposal and mine have in common is defensible against recent objections, but that Wedgwood's view falters in the many-option cases where our proposals disagree.

1. Introduction

We rational agents are part of the world we act on. This means that when we deliberate, the mental states and processes that constitute our deliberation can be causally and evidentially related to the things we value, just like any other states of the world can be. Also like other states of the world, our mental states are not luminous. We can be ignorant of them, just as we can be of other matters of fact. This does not mean that our knowledge of ourselves always is limited to what can be inferred from our actions, 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. Thus when we deliberate about which action to adopt, the same actions we deliberate over can, once performed, amount to evidence not just of their effects but their immediate causes in our own psychology. Sometimes we surprise ourselves.

These facts have been a continuing source of problems for theories of rational decision. Perhaps the best known is the Newcomb problem for **evidential decision theory (EDT)**. You know the story by now.[1] An infallible Predictor has decided to massively reward you for being the kind of person who selects a strictly dominated option in a particular context, so that your doing so amounts to decisive evidence that you will get the reward. Selecting the dominated option does you no good, but it is evidence that your preexisting psychological states will get you the reward regardless of which option you select. Evidential decision theory allegedly advises us to maximize good news, by selecting this dominated option. 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 mental states 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 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 *Cr* is one's credence function, *v* one's value function, and the *K*s **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 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 says 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. For instance, CDT might be defended by rejecting the common intuition that you should refrain from pressing. A lot can be said in favor of such an error theory.[2] But here my tentative working hypothesis is that common intuitions about Psychopath Button and other related cases discussed below are correct. Other defenses of CDT question whether CDT really is committed to recommending pressing.[3] These defenses often assume that relevant facts about your own mind, such as your utility function, degree of rationality, provisional inclinations toward a decision, and the like, are luminous. But I take the interest of cases like Psychopath Button and even the Newcomb case to turn on rejecting luminosity assumptions. Rejecting luminosity is what ensures that the acts of pressing or refraining—the very options you deliberate over—can amount to evidence about their likely consequences; evidence which is not screened off by prior knowledge of your own deliberations. Again, I do not claim the relevant luminosity assumptions are indefensible. Some widely defended theories of self-knowledge take it to be rationally required.[4] But again, my working hypothesis here is that facts about one's own mind, like other matters of fact, are not luminous. You can be perfectly rational without knowing whether you are a psychopath, what your current utility function is, or even which option you currently are leaning toward.[5]

So I plan to simply assume that we should *somehow* take the evidential significance of our actions into account, in a way CDT does not allow. What I will concern myself with is *how*.

This is not a paper about whether Psychopath Button and its cohort are counterexamples to CDT, but about what to do about it if so.

I will start by explaining the 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. Absolute Ratifiability Proposals

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

I will explicate ratifiability in terms of what I call **conditional causally expected utility**. Let U(B|A) be the causally 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 $U(O|A) > U(A|A)$.[6] (Note that the notion under discussion here is *causal* ratifiability, since it is defined in terms of conditional causally expected utility.)

There are a number of ways that ratifiability might be thought to be related to rational permissibility. Let **simple ratificationism** be the view that modifies CDT by imposing ratifiability as a necessary condition on rationality, so that one's rationally permissible options are those with the highest causally expected utility among the ratifiable ones.[7] This proposal arguably fares better than CDT in some cases, such as:

> **Tickets:** A fair coin was tossed, but you do not know how it landed. You can buy a $1 ticket that pays $2 if it was heads, or buy a $1 ticket that pays $2 if it was tails, or refrain from buying a ticket. And infallible Predictor knows which option you will select. In addition to that, the Predictor knows about your dispositions. It turns out that some people have robust dispositions to buy a particular ticket, even though it is not introspectively obvious to them. When they in fact buy a ticket, their buying that ticket rather than the other one would not have varied with minor, rationally irrelevant changes in their circumstances. And even when they in fact buy no ticket, there still is a fact of the matter about which ticket they would have bought if they had bought one. The Predictor, who knows how the coin has landed, informs you that if you were to buy a ticket, you would buy the losing one (even now that you know this).

Only refraining is ratifiable. And it is natural to say that this is your only permissible option, as simple ratificationism holds. This stands in contrast to CDT, which arguably says that at least one ticket can permissibly be bought.[8]

But simple ratificationism faces trouble with cases in which no option is ratifiable. An initial problem concerns the classic case raised by Gibbard and Harper (1978):

> **Death in Damascus:** 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, and will wait for you in Aleppo if you are predicted to go, and will go to meet you in Damascus if you are predicted to stay.

Here no option is ratifiable. So simple ratificationism must say you are caught in a rational dilemma, in which every option is rationally impermissible. This can seem undesirable. You have to do something. So it seems unfair to say of each of your options that you are prohibited from doing it.

Is this a definitive strike against the simple ratificationism? I used to think so, but now I am not so sure. There are many defenders of the possibility of moral dilemmas.[9] And a number of authors have recently argued for the possibility of rational dilemmas in the epistemic domain.[10] Alleged dilemmas raise many vexing issues about the notions of rational permission and obligation, and their connections to reactive attitudes like praise and blame. Plausibly, no matter what you do in Death in Damascus, we should not hold it against you. But perhaps that is compatible with the view that whatever you do, you should not have done it.[11] In any case, Arif Ahmed (2012) has argued forcefully that any view that says you should refrain in Psychopath Button will face similar challenges.[12] It might be that any view that accommodates common intuitions about these cases must accept the possibility of rational dilemmas. So whatever we make of these difficulties, they are not unique to simple ratificationism.

The more troubling cases for simple ratificationism are ones where both one's options are unratifiable, but one nonetheless seems preferable to the other. Psychopath Button is itself plausibly such a case, since neither option is ratifiable, but refraining seems preferable to pressing. Here are two more:[13]

> **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 at the same time, staying seems to have an obvious advantage. 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. For this reason, it is natural to consider staying preferable to switching. At least, that is what we should say if we consider refraining preferable to pressing in Psychopath Button.

In the same way, it is also natural to consider staying preferable to going in Lazy Death 1, even though both options are unratifiable. Even though staying is strong evidence that you are staying to meet an avoidable death, that seems better than going, which is conclusive evidence that you are going to meet an avoidable death. At least, that is what we should say if we think refraining is preferable in Psychopath Button.

Simple ratificationism instead says that Bottles 1 and Lazy Death 1 are rational dilemmas, where no option is permitted. Even if we accept this diagnosis of Death in Damascus, it is harder to accept in these cases, where there seems to be a clear reason to prefer one option to the other. To be sure, both options are unratifiable. If you stay in Lazy Death 1, for example, you will immediately regret it. But if you go, you will regret it even more. This can make staying seem, if not good, then at least less bad than the alternative. This is all I mean when I say that staying is 'preferable' to the alternative. It is natural to take this to entail that staying is permissible, and that the alternative is not. But perhaps a supporter of rational dilemmas could maintain that while both options are impermissible, some impermissible options still can be preferable to others.[14] In that case, the objection to simple ratificationism is not that has false implications in these cases, but instead that it cannot explain all the data.

But the problems for simple ratificationism do not stop there. There are corresponding problems in cases where both options are unratifiable:[15]

> **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, as simple ratificationism does. Since in Bottles 2 and Lazy Death 2 switching and staying are both ratifiable, simple ratificationism does not unambiguously

recommend staying. Instead, it says to prefer the option with higher causally expected utility. This means that what simple ratificationism recommends for these cases depends on one's *unconditional* credences about what one will do. It thus will recommend switching in Bottles 2 and going in Lazy Death 2, if initially you are confident enough that you will stay. While this on its own is a defensible view, what seems less defensible is distinguishing as simple ratificationism does between cases where neither option is ratifiable and cases where both are. If we take your conditional credence to be what matter in Psychopath Button, Bottles 1, and Lazy Death 1, then why not for Bottles 2 and Lazy Death 2?

Related concerns led Egan to reject simple ratificationism. He tentatively proposed instead **lexical ratificationism**, which says an option A is rationally permissible iff either A is ratifiable and has evidentially expected utility no lower than one's other ratifiable options, or else there are no ratifiable options and A has evidentially expected utility no lower than one's other (unratifiable) options. The idea is to have ratifiable options always preferable to unratifiable ones, and then to have evidentially expected utility to settle between options that are either both ratifiable or both unratifiable.

Unfortunately, as Egan himself notes, lexical ratificationism has its own problems. In particular, sometimes unratifiable options seem preferable to ratifiable ones. Here is my version of a case that Egan credits to Anil Gupta:

> **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. At the very least, it seems that selecting A is not rationally mandatory. And yet, selecting A is the only ratifiable option. If so, lexical ratificationism must be rejected as well.

In broad outline, I take the problem to be this. Whether an option is ratifiable depends on the causally expected utility, conditional on one's selecting that option, of *all* one's options. For example, selecting B is unratifiable because conditional on one's selecting B, selecting C instead has higher causally expected utility. Since lexical ratification says ratifiable actions are always preferable to unratifiable ones, this means that selecting A is preferable to selecting B as a result of one's having box C available. If Box C were unavailable, then suddenly lexical ratificationism would say that selecting B is preferable to selecting A. As I discuss in Section 7, it seems implausible that the availability of Box C should in this way affect the preferability of A to B.

Although Egan does not consider this argument, he agrees that selecting A is not required, as lexical ratificationism demands. So although he sees sufficient reason to reject CDT, he is unsure what should replace it.

3. Degrees of Ratifiability

There is a better way to incorporate ratifiability into the theory of rational decision. What gets us into trouble in Psychopath Button and its cohort is that ratifiability is an all-or-nothing notion. We need a notion of ratifiability that comes in degrees.

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 $U(A|A) - U(B|A) \geq 0$. 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.

By appealing to this graded notion of ratifiability, we can capture the intuition that asymmetrical cases like Bottles 1 and Lazy Death 1 are importantly different from symmetrical ones like Death in Damascus, even though in all these cases one's options all are unratifiable.

In Death in Damascus, $U(stay|stay) - U(go|stay)$ is very low, since assuming you stay, it is certain that you are staying to meet an avoidable death. But $U(go|go) - U(stay|go)$ is the same very low value for the same reason. Since both options are ratifiable to the same degree, my proposal says that neither option is preferable to the other.

In Lazy Death 1, by contrast, while your options are both unratifiable, they are unratifiable to different degrees. $U(stay|stay) - U(go|stay)$ is still very low, but it is still higher than $U(go|go) - U(stay|go)$. This is because staying is merely strong evidence that you are causing your death, while going is conclusive evidence that you are. Likewise in Bottles 1, $U(stay|stay) - U(go|stay) > U(stay|stay) - U(stay|stay)$. While staying is conclusive evidence that you cause yourself harm by staying, switching is conclusive evidence that you cause yourself even worse harm by switching.

I take the proposal to similarly capture the common intuition that refraining is preferable in 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. Like in Bottles 1 and Lazy Death 1, both options are unratifiable. Whatever one does, one will regret it. But one option is more regrettable than the other.

Unlike simple ratificationism, the proposal handles cases like Bottles 2 and Lazy Death 2 in a parallel fashion. In these cases, each of one's options is ratifiable. Indeed, one's performing any given option is evidence that the option will produce a better outcome than the alternative would have. But even so, there is an asymmetry. In Lazy Death 2, for example, going to Aleppo is strong evidence that you are going to the city where Death will not be waiting, making $U(go|go) - U(stay|go)$ very high. But staying in Damascus is conclusive

evidence that you are staying where Death will not be, making U(stay|stay) - U(go|stay) even higher.

How should we generalize this proposal to cases in which one has more than two options? One idea 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. As we will see, Wedgwood's proposal faces the same difficulties concerning Boxes 1 and related cases that Egan's lexical ratificationism did.

For this reason, 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)**.

We will consider many-option cases in detail in Section 7. For now, note simply that GRT delivers plausible results for Boxes 1, where lexical ratificationism faltered. The problem with lexical ratificationism arose because an option's ratifiability depends on how it compares to *all* one's options. So which option is preferable among A and B turns on the relevant conditional expected utility for box C, so that the availability of C makes selecting B unratifiable. In contrast, GRT makes the preferability among A and B depend on a pairwise comparison between the two. 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. Thus even though selecting B is unratifiable in an absolute sense, it still is more ratifiable than—and thus preferable to—selecting A.

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 the Newcomb case. So GRT does not recommend managing the news in cases where EDT famously does.

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 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. 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.[16]

5. Wedgwood's Benchmark Theory

Ralph Wedgwood (2013) 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:

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

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 **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.[17] 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.[18]

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 will not 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 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
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 $Cr(K) > 0$, $v(KA) \geq v(KB)$, 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.[19] 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, K_j, v(KA) > v(KB), then for some O, U(A|O) > U(B|O), because K_j 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:

> **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.

This implication of GRT is not obviously correct. But I think it is not obviously incorrect, either. In fact, I think it should be quite plausible to anyone sympathetic to ratificationist intuitions about Psychopath Button and related cases, despite its inconsistency with 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 option (we can stipulate) has nonzero credence, you should have a nonzero unconditional credence that B will pay more than C. This is why B weakly dominates C. But as ratificationists see things, conditional credences are what matter. Assuming you select either B or C, it is certain that they 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 can plausibly 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 likely 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 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.[20] The problem for BT is Boxes 4. If the benchmarks are set via averaging, then BT says you should select A.[21] 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 are obligated to 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 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.[22] So far, so good. But with the option to select C available, and with benchmarks set via averaging, BT says you should select A.[23] 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 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.[24] 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.[25] So BT would say, implausibly, that selecting A suddenly becomes impermissible 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 do not 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[26]
> 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.[27] 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
> 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.[28] 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.[29] 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[30]
> 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.

Here BT says that you should select A.[31] 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. 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. While this intuition is open to challenge, my central claim is the conditional one 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 we should not consider the evidential significance of your actions at all, in contrast to intuitions about Psychopath Button and related cases. But if we should, we should do it the way GRT instructs.[32]

---

[1] If you don't know the story by now, see, e.g., (Weirich 2016, Sec. 2).

[2] See, e.g., (Ahmed, 2012), (Cantwell, 2010), and (Joyce, 2010).

[3] See, e.g., (Briggs 2010, Sec. 7), (Cantwell 2010, pg. 138), and especially (Arntzenius 2008) and (Joyce 2010), which build off prior work by Skyrms (1990) and others. But see also (Hare and Hedden 2016) for further problems for CDT that go along with luminosity assumptions.

[4] For discussion of self-knowledge in a decision theoretic context, see (Sobel 1987). For more general discussion of self-knowledge and rationality, see, e.g., (Gertler 2011 and 2015), (Smithies forthcoming) and (Shoemaker 1996). See *[omitted]* for a partial rebuttal.

[5] Even granting that lacking self-knowledge is irrational, there remains the question of how to act when one is irrationally ignorant. It is not clear CDT gets the right answer, as Hare and Hedden (2016, Sec. 4) discuss.

[6] Cf. (Eells and Harper 1991, pg. 4), (Harper 1986, pg. 33), (Skyrms, 1990, pg. 46), and (Sobel 1990, pg. 62). The notion of ratifiability is originally due to (Jeffrey 1983).

[7] Simple ratificationism is endorsed by (Harper 1986, pg. 33), and opposed by (Eells and Harper 1991, pg. 6) and (Egan 2007).

[8] For at least one ticket t, Cr(buy t | buy some ticket) ≤ 1/2, in which case U(buy t) ≥ 0 = U(buy neither). Of course, this is open to the same objections as corresponding claims about Psychopath Button.

[9] E.g., (Marcus 1980) and (Sinott-Armstrong 1988, Ch. 6).

[10] E.g., (Christensen 2016) and (Pryor 2018). But see *[omitted]* for some worries.

[11] Cf. (Sobel 1990, pg. 71) and (Weirich 1985).

[12] See also Hare and Hedden's (2016) Three Crates case, which is arguably a dilemma if Ahmed's ABZ case is.

[13] Cf. Hare and Hedden's (2016) Asymmetrically Nasty Demon, Harper's (1986, pg. 30) modified Death in Damascus, Richter's (1984) Modified Death Case, Sobel's (1990) Case III, and Skyrms' (1990) Mean Demon.

[14] Christensen (2016) and Pryor (2018) advance related views about epistemic dilemmas.

[15] Cf. Richter's (1984) Button Case and Skyrms' (1990) Nice Demon.

[16] 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.

---

[17] See also Eells and Harper 1991.

[18] The example is from Sen 1993.

[19] See Bassett 2015, Sec. 3.5.

[20] 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.

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

[22] 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.

[23] 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.

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

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

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

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

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

[29] 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.

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

[31] With benchmarks set via relief, EECV(A) = .9(0) + .05(1) + .05(1) = .1 > EECV(B) = .05(1) + .9(0) + .05(0) = .05.

References

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

Arntzenius, Frank (2008) 'No Regrets, or: Edith Piaf Revamps Decision Theory' *Erkenntnis* 68: 277-297.

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.

Christensen, David (2016) 'Conciliation, Uniqueness, and Rational Toxicity' *Noûs* 50(3): 584-603.

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

Eells, Ellery and Harper, William L. (1991) 'Ratifiability, Game Theory, and the Principle of Independence of Irrelevant Alternatives' *Australasian Journal of Philosophy* 69(1): 1-19.

Gertler, Brie (2011) *Self-Knowledge*. London: Routledge.

——— (2015) "Self-Knowledge", *Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2015/entries/self-knowledge/>.

Hare, Caspar and Hedden, Brian (2016) 'Self-Reinforcing and Self-Frustrating Decisions' *Noûs* 50(3): 604-628.

Harper, William L. (1986) 'Mixed Strategies and Ratifiability in Causal Decision Theory' *Erkenntnis* 24: 25-36.

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

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

Richter, Reed (1984) 'Rationality Revisited' *Australasian Journal of Philosophy* 62(4): 392-403.

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

Shoemaker, Sydney (1996) *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Skyrms, Brian (1990) 'Ratifiability and the Logic of Decision' *Midwest Studies in Philosophy* 15: 44-56.

Smithies, Declan (forthcoming) *The Epistemic Role of Consciousness*. OUP.

Sobel, Jordan Howard (1987) 'Self-Doubts and Dutch Strategies' *Australasian Journal of Philosophy* 65(1): 56-81.

———— (1990) 'Maximization, Stability of Decisions, and Actions in Accordance with Reason' *Philosophy of Science* 57: 60-77.

Wedgwood, Ralph (2013) 'Gandalf's Solution to the Newcomb Problem' *Synthese* (14): 1-33.

Weirich, Paul (1985) 'Decision Instability' *Australasian Journal of Philosophy* 63(4): 465-472.

———— (2016) 'Causal Decision Theory' *Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/decision-causal/>.