Graded Ratifiability

ABSTRACT. An action is unratifiable when, on the assumption that one performs it, another option has higher expected utility. Unratifiable actions are often claimed to be somehow rationally defective. But in some cases where multiple options are unratifiable, one unratifiable option can still seem preferable to another. We should respond, I argue, by invoking a graded notion of ratifiability.

We rational agents are part of the world we act on. When we deliberate, the mental states involved in 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. Thus when we deliberate about which action to adopt, the same actions we deliberate over can, once performed, amount to evidence of not just 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. Probably 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 reward you for being the kind of person who takes a strictly dominated option in a particular context, so your doing so is evidence you get the reward. Evidential decision theory allegedly recommends maximizing good news, by taking 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 options will produce. Andy Egan provides a recent example, which he attributes to David Braddon-Mitchell:[2]

> 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. 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 you are a psychopath. So you can be certain that if you press, then you will kill yourself.

Egan also claims, again plausibly, that CDT instead recommends pressing. CDT holds that an option B is *rationally preferable* to an option A iff the *causally expected utility* of B-ing is

[1] If you don't know the story by now, see Paul Weirich, "Causal Decision Theory," *Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), 2016.

[2] Egan, "Some Counterexamples to Causal Decision Theory," *Philosophical Review* CXVI, 1 (2007): 93-114.

greater than that of A-ing. Where *Cr* is one's credence function, *v* one's value function, and the *K*s *dependence hypotheses*—that is, 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:

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

CDT has the expected utility of an option depend on one's unconditional credences about its effects. It gets the right results in Newcomb cases by ignoring one's conditional credences about what is true if one selects the option. But for the same reason, 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.[3] But here my working hypothesis is that common intuitions about Psychopath Button and other related cases discussed below are correct. Other defenses question whether CDT must recommend pressing.[4] But these defenses often require relevant facts about your own mind, such as your utility function, degree of rationality, provisional inclinations toward a decision, and the like, to be luminous. In contrast, I take the interest of cases like Psychopath Button and even the Newcomb case to turn on rejecting luminosity assumptions. Rejecting luminosity ensures that the acts of pressing or refraining—the very options one deliberates over—can amount to evidence about their likely consequences; evidence which is not screened off by prior knowledge of one's deliberations. Again, I do not claim the relevant luminosity assumptions are indefensible. Some widely defended theories of self-knowledge take self-knowledge to be rationally required.[5] But again, my working hypothesis is that facts about one's own mind, like other matters of fact, are not luminous. You can be perfectly rational without knowing

[3] Arif Ahmed, "Push the Button," *Philosophy of Science* LXXIX, 3 (2012): 386-395; John Cantwell, "On an Alleged Counter-Example to Causal Decision Theory," *Synthese* CLXXIII, 2 (2010): 127-152; James M. Joyce, "Regret and Instability in Causal Decision Theory," *Synthese* CLXXXVII, 1 (2012): 123-145.

[4] Ray Briggs, 'Decision-Theoretic Paradoxes as Voting Paradoxes' *Philosophical Review* CXIX, 1 (2010): Sec. 7; Cantwell, "On an Alleged Counter-Example to Causal Decision Theory," *op. cit.*, pg. 138; and especially Frank Arntzenius, "No Regrets, or: Edith Piaf Revamps Decision Theory," *Erkenntnis* LXVIII (2008): 277-297 and Joyce, "Regret and Instability in Causal Decision Theory," *op. cit.*, which build off Brian Skyrms, "Ratifiability and the Logic of Decision," *Midwest Studies in Philosophy* XV (1990): 44-56.

[5] For discussion of self-knowledge in a decision theoretic context, see Jordan Howard Sobel, "Self-Doubts and Dutch Strategies," *Australasian Journal of Philosophy* LXV, 1 (1987): 56-81. For more general discussions linking self-knowledge and rationality, see Gertler, Brie, *Self-Knowledge*, (London: Routledge, 2011) and "Self-Knowledge," *Stanford Encyclopedia of Philosophy* (2015), Edward N. Zalta (ed.); Declan Smithies, *The Epistemic Role of Consciousness*, (Oxford University Press, 2019); Sydney Shoemaker, *The First-Person Perspective and Other Essays* (Cambridge: Cambridge University Press, 1996). For rebuttal, see David James Barnett, "Self-Knowledge Requirements and Moore's Paradox," *Philosophical Review* CXXX, 2; and Timothy Williamson, *Knowledge and Its Limits* (Oxford: OUP, 2000).

whether you are a psychopath, what your current utility function is, or even which option you currently are leaning toward.[6]

So I will mostly just 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 they are.

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 related ones from Ralph Wedgwood and Dmitri Gallow.[7]

I. Absolute Ratifiability

Why does pressing seem irrational in Psychopath Button? A natural suggestion is that pressing is *unratifiable*.

Let U(A|B) be the causally expected utility of A-ing conditional on the assumption that one Bs, such that:

$$U\left(A \mid B\right) = \sum_{K} Cr\left(K \mid B\right) v\left(KA\right).$$

An option A is (causally) ratifiable iff you have no other option O such that U(O|A) > U(A|A).[8] Ratifiability seems potentially normatively significant. But how?

I.1. Simple Ratificationism and Rational Dilemmas. One proposal, *simple ratificationism*, supplements CDT with an absolute prohibition against unratifiable actions, such that one's rationally permissible options are those with the highest causally expected utility among the ratifiable ones.[9] This proposal arguably fares better than CDT in some cases, such as:

[6] Even granting luminosity, there remains the question how to act when one is *irrationally* ignorant. See Caspar Hare and Brian Hedden, "Self-Reinforcing and Self-Frustrating Decisions," *Noûs* L, 3 (2016): Sec. 4.

[7] Wedgwood, "Gandalf's Solution to the Newcomb Problem," *Synthese* XIV (2013): 1-33, and Gallow, "The Causal Decision Theorist's Guide to Managing the News," *Journal of Philosophy*, CXVII, 3 (2020): 117-149.

[8] Cf. Ellery Eells and William L. Harper, "Ratifiability, Game Theory, and the Principle of Independence of Irrelevant Alternatives," *Australasian Journal of Philosophy* LXIX, (1991): 1-19, at pg. 4; William L. Harper, "Mixed Strategies and Ratifiability in Causal Decision Theory," *Erkenntnis* XXIV (1986): 25-36, at pg. 33; Skyrms, "Ratifiability and the Logic of Decision," *op. cit.*, pg. 46; and Jordan Howard Sobel, "Maximization, Stability of Decisions, and Actions in Accordance with Reason," *Philosophy of Science* LVII (1990): 60-77, at pg. 62. The notion of ratifiability is originally due to Richard Jeffrey '*The Logic of Decision*, 2nd ed. (Chicago: University of Chicago Press, 1983).

[9] Simple ratificationism is endorsed by Harper, "Mixed Strategies and Ratifiability in Causal Decision Theory," *op. cit.*, pg. 33, and opposed by Eells and Harper, "Ratifiability, Game Theory, and the Principle of Independence of Irrelevant Alternatives," *op. cit.*, pg. 6 and Egan, "Some Counterexamples to Causal Decision Theory," *op. cit.*

> 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. An infallible (but not omniscient) Predictor had some trouble determining whether you will buy a ticket, but the Predictor still managed to figure out which ticket you would buy if you buy one. The Predictor, who also knows how the coin landed, informs you that the ticket you would buy is the losing one.

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.[10]

But simple ratificationism faces trouble when no option is ratifiable:[11]

> 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.

Since neither option is ratifiable, simple ratificationism must say you are caught in a *rational dilemma*, where no option is permissible. This commitment seems unappealing. You have to do something. If you do not go to Aleppo, then you will stay in Damascus. So it is no fair telling you it is irrational to go, but also irrational to stay.

I used to think this was a decisive strike against simple ratificationism, but now I am not so sure. Some philosophers countenance moral dilemmas.[12] And others admit rational dilemmas in the epistemic domain.[13] We should not prematurely write off the possibility of practical rational dilemmas. And if there are any practical rational dilemmas, among the likeliest candidates are cases like Death in Damascus and perhaps even Psychopath Button, which might be alleged to involve conflicts between deliberative and predictive perspectives on one's actions. The idea that there is some conflict between rational deliberation and prediction is widespread, if controversial. On an extreme view, rational deliberation is incompatible with regarding one's actions as predictable at all, even by others.[14] Various

[10] For at least one ticket t, $Cr(\text{buy } t \mid \text{buy some ticket}) \leq 1/2$, in which case $U(\text{buy } t) \geq 0 = U(\text{buy neither})$. Of course, defenses of CDT regarding Psychopath Button could be deployed here, too.

[11] Alan Gibbard and William L. Harper, "Counterfactuals and Two Kinds of Expected Utility," in *Foundations and Applications of Decision Theory, vol. 1*, A. Hooker, J. J. Leach & E. F. McClennen (eds.), Boston: D. Reidel (1978).

[12] See Ruth Barcan Marcus "Moral Dilemmas and Consistency," *Journal of Philosophy* LXXVII, 3 (1980): 121-136 and Walter Sinnott-Armstrong *Moral Dilemmas.* (Basil Blackwell, 1988), Ch. 6.

[13] David Christensen, "Conciliation, Uniqueness, and Rational Toxicity," *Noûs* L, 3 (2016): 584-603; Nick Hughes, "Dilemmic Epistemology," *Synthese* CXCVI (2019): 4059-4090; and James Pryor "The Merits of Incoherence," *Analytic Philosophy* LIX, 1 (2018): 112-141. But for some worries, see David James Barnett, "Internalism, Stored Beliefs, and Forgotten Evidence," in Stephen Wright and Sanford Goldberg, eds., *Memory and Testimony: New Essays in Epistemology* (Oxford University Press, forthcoming).

[14] Cf. Christine Korsgaard, *The Sources of Normativity* (Cambridge University Press, 1996), pp. 94-96.

weaker views say deliberation merely conflicts with making one's own predictions of one's actions, in the form of certainty about what one will do or on some views even determinate credence.[15] Still other views see a tension between deliberation and one's making the specific prediction that one might not follow through on one's decision.[16] Now I won't here try to offer a theory of how prediction and deliberation might conflict, or how this conflict might generate dilemmas. My point is just that simple ratificationism's commitment to dilemmas in these cases is not unmotivated or arbitrary.

Still, admitting dilemmas raises problems. A first set involve reactive attitudes like praise and blame. Plausibly, no matter what you do in Death in Damascus, we should not hold it against you. But the relationship between permissibility and blamelessness is tricky enough for moral permissibility, and arguably even trickier for rational permissibility. Maybe in dilemmas you can blamelessly do what you should not.[17]

A more serious problem in my view is that rationality is supposed to offer guidance or advice. If an option is rationally impermissible, that is not just one strike against it. It means one is all-things-considered advised against adopting it. In difficult decisions, one might have to balance competing considerations of various strengths. But at the end of the day, one or more options is bound to come out ahead. If those options still can be rationally impermissible, it is as if rationality is advising you not to adopt them, and thus to adopt some worse option instead.

This objection to dilemmas appeals to a kind of comparison between options. Say one option is *rationally preferable* to another if it comes out looking better in this comparison; if it has more to recommend it, or stronger reasons in favor of it. Preferability is so-called because of its apparent relationship to preferences—one plausibly should be in the mental state of preferring one option to another just in case it compares favorably. But I regard this as merely a plausible claim about preferability, rather than a stipulative definition. An eliminativist about preferences could still acknowledge that one option can be recommended over another by one's reasons.

Could it be that staying in Damascus is preferable to going to Aleppo, and also that going is preferable to staying? Simple ratificationists might allege a kind of instability in the comparison of these options. When you consider going to Aleppo, staying in Damascus seems better, and when you consider staying, going does. So depending on which option you start with in making the comparison, each one can come out favored over the other. Maybe that is not the same as both options being stably advised against. But it still means there is no stably advisable resolution to your deliberations.

[15] See Alan Hájek, "Deliberation Welcomes Prediction," *Episteme* XIII, 4 (2016): 507-528; Isaac Levi, "Rationality, Prediction, and Autonomous Choice," *Canadian Journal of Philosophy* XXIII, sup. 1 (1993): 339-363; Yang Liu and Huw Price, "Heart of DARCness," *Australasian Journal of Philosophy* XCVII, 1 (2019): 136-150; Wlodek Rabinowicz, "Does Practical Deliberation Crowd Out Self-Prediction?" *Erkenntnis* 57 (2002): 91-122; and Katia Vavova, "Deliberation and Prediction: It's Complicated," *Episteme* XIII, 4 (2016): 529-538.

[16] Berislav Marušić, *Evidence and Agency: Norms of Belief for Promising and Resolving*, (New York: Oxford University Press, 2015).

[17] Cf. Sobel, "Maximization, Stability of Decisions, and Actions in Accordance with Reason," *op. cit.*, pg. 71 and Paul Weirich, 'Decision Instability' *Australasian Journal of Philosophy* LXIII, 4 (1985): 465-472.

But this proposal also faces a natural objection, which will come back to haunt my own account in Section V.2 below. For it seems possible to take a step back, and see that neither option has anything to recommend it over the other, all things considered. If this holistic comparison is what has ultimate normative significance, we should deny that each option is preferable to the other in Death in Damascus. More generally, for arbitrary options A-ing and B-ing, we should then accept:

> ANTISYMMETRY: If A-ing is preferable to B-ing, then B-ing is not preferable to A-ing.

By Antisymmetry, whenever one has only two options, one must have at least one with no other option that is rationally preferable to it. It will follow that two-option cases like Death in Damascus are not dilemmas by the right-to-left direction of:

> PERMISSIBILITY: An option is rationally permissible iff one has no other option that is rationally preferable to it.

By Antisymmetry and Permissibility, two-option dilemmas are impossible. If simple ratificationists grant Antisymmetry, they must reject Permissibility. But this puts them in an uncomfortable spot. As bad each of one's options are in Death in Damascus, there is nothing to be said against one that cannot just as well be said against the other. But if neither is preferable to the other, then for neither option will it be true that it is advisable to adopt the other option rather than it.

I.2. Simple Ratificationism and Asymmetrical Unratifiability. In Death in Damascus, nothing favors one option over another. But there are further cases where no option is ratifiable, but one nevertheless has something going for it that the other does not. Psychopath Button is plausibly such a case. Here are two more:[18]

> 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 bottle you are holding. A malevolent Predictor determined the contents of the bottles as follows. 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 bit lazy. If you are predicted to go to Aleppo, then Death will definitely wait for you there. If you are predicted to stay in Damascus, then Death will probably go to Damascus, though there is a chance he won't bother.

[18] Cf. Hare and Hedden's Asymmetrically Nasty Demon in "Self-Reinforcing and Self-Frustrating Decisions," *op. cit.*; Harper's modified Death in Damascus, in "Mixed Strategies and Ratifiability in Causal Decision Theory," *op. cit.*; Reed Richter's Modified Death Case from 'Rationality Revisited', *Australasian Journal of Philosophy* LXII, 4 (1984): 392-403; Sobel's Case III, in "Maximization, Stability of Decisions, and Actions in Accordance with Reason," *op. cit.*; and Skyrms's Mean Demon in "Ratifiability and the Logic of Decision," *op. cit.*

In Bottles 1, both switching and staying are unratifiable. Even so, staying has an apparent advantage, which plausibly makes it preferable to switching. Similarly, staying is preferable to going in Lazy Death 1, even though both are unratifiable. At least, that is what we should say if we think refraining is preferable in Psychopath Button.

Simple ratificationism says that Bottles 1 and Lazy Death 1 are rational dilemmas. If so, then rejecting Permissibility does not go far enough. It will have to be possible for staying to be impermissible even when it is preferable to the alternative. But if staying is better than the alternative, then it is the thing to do. Whatever its defects, it has got to be permissible.[19]

Even if we allow an option to be impermissible despite being the best one has available, simple ratificationism has further problems:[20]

> 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 bottle you are holding. A benevolent Predictor determined the contents of the bottles as follows. 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.

Simple ratificationism grants there are no dilemmas here, where all options are ratifiable. So what should you do? If we are still impressed by the importance of ratifiability, it is natural to think staying is again preferable in both cases. But simple ratificationism leaves no room for the asymmetry in ratifiability to make a difference to what is permissible. Instead, it just recommends the option with higher causally expected utility, which will not be staying if you are confident that you will not stay.

I.3. Lexical Ratificationism and Many-Option Cases. 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 settle between equally (un)ratifiable options.

[19] Cf. Christensen, "Conciliation, Uniqueness, and Rational Toxicity," *op. cit.*, Hughes, "Dilemmic Epistemology," *op. cit.*, and Pryor, "The Merits of Incoherence," *op. cit.*

[20] Cf. Richter's *op. cit.* Button Case and Skyrms's Nice Demon in "Ratifiability and the Logic of Decision," *op. cit.*

Unfortunately, as Egan himself notes, lexical ratificationism has problems. 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 set their contents by 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 take A, then scheme 1 was used, if B, then scheme 2, and if C, then scheme 3.

This case might be controversial, and I discuss many like it in Section IV. On a first pass, it seems plausible to many, including Egan and myself, that one should prefer both B and C to A. At the very least, it seems that taking A is not rationally mandatory. And yet, taking A is the only ratifiable option.

In broad outline, I take the problem with lexical ratificationism to be this. Whether an option is ratifiable depends on the causally expected utility, conditional on one's taking that option, of *all* one's options. For example, taking B is unratifiable because conditional on one's taking B, taking C has higher causally expected utility. Since lexical ratification says ratifiable actions are always preferable to unratifiable ones, this means that taking A is preferable to taking B as a result of one's having box C available. If Box C were unavailable, then suddenly lexical ratificationism would say that taking B is preferable to taking A. We will return to this in Section IV.2.

II. 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 treating ratifiability as all-or-nothing. We need a notion of ratifiability that comes in degrees.

Suppose one has two options, A-ing and B-ing. Notice that A-ing will be ratifiable iff $U(A|A) - U(B|A) \geq 0$. That is, A-ing is ratifiable iff its causally expected utility is no lower than the alternative, assuming that one As. 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-ing is ratifiable, the more its expected utility exceeds the alternative, conditional on the assumption that one As.

My proposal for two-option cases is that the rationally preferable option is the one with the higher degree of ratifiability. I take this proposal to offer a natural explanation of the above cases. Just consider the intuitive contrast between asymmetrical cases of unratifiability, like Bottles 1 and Lazy Death 1, and symmetrical ones like Death in Damascus. Even though one's options are all unratifiable, in the symmetrical Death in Damascus, neither option is worse than the other. Assuming you stay, you are certainly much better off going, and assuming you go, you are certainly much better off staying. But both the degree of certainty and the degree of betterness are the same. Since both options are ratifiable to the same degree, neither is preferable.

But in Lazy Death 1, your options are unratifiable to different degrees. Assuming you go, you are certainly much better off staying, but assuming you stay, you are *probably* better off going. And in Bottles 1, assuming you switch, you are certainly much better off switching, but assuming you stay, you are certainly only *somewhat* better off switching.

How does this generalize to many-option cases? One idea is to set an option's degree of ratifiability relative to an overall benchmark, which reflects all one's alternatives. 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.

So I think it is better to have the rational preferability determined by pairwise comparison. Let the *comparative degree of ratifiability* of A-ing relative to B-ing be $U(A|A) - U(B|A)$, no matter what other options one has available. My proposal is that 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 ratificationism (GR).*

In the first instance GR is a theory about preferability; about when one's reasons favor one option over another. But it is natural to accept Permissibility, and say an option is permissible iff no alternative options are preferable. If so, an option A will be permissible iff one has no option B such that $U(A|A) - U(B|A) < U(B|B) - U(A|B)$.

GR agrees with CDT about a lot. In any *stable* decision, where each option's conditional expected utilities equal its unconditional expected utility, $U(A|A) - U(B|A) > U(B|B) - U(A|B)$ iff $U(A) - U(B) > U(B) - U(A)$, which will be so iff $U(A) > U(B)$. But in unstable decisions like those above, GR can recommend an option over a less ratifiable one with higher unconditional expected utility.

Many-option cases raise difficulties for any broadly ratificationist view, and GR is no exception. We return to them in Sections IV and V. For now, note simply that GR 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 taking B unratifiable. In contrast, GR 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 taking B is unratifiable in an absolute sense, it still is more ratifiable than, and thus preferable to, taking A.

That is the basic idea. How it works will become clearer as we consider some implications, objections, and comparisons to competitors.

III. Causal Dominance and Managing the News

The usual knock against EDT is that it favors irrationally managing the news. Does GR? It might appear so, since GR also has you consider the evidential significance of your actions. But the appearance is misleading. Unlike with EDT, what matters for GR is not what your actions tell you about how good a situation you're already in. It's what they tell you about what you can do about it. Contrast Bottles 1 with:

> 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 Predictor determined the contents of the bottles as follows. If you were predicted to switch, your bottle contains a poison that will kill you painfully, while the other bottle contains a poison that would kill you painlessly. If you were predicted to stay, then your bottle contains a mild poison that will make you temporarily ill, while the other contains water.

EDT famously recommends staying in cases like this.[21] For your staying would be great news. It is conclusive evidence that you are in a situation where the worst that could happen is temporary illness, whereas switching is conclusive evidence you are in a situation where death is unavoidable. But despite being good news, staying does not bring good results. It is already settled which situation you are in, and there is no changing it now. By staying, you would merely offer yourself evidence that your situation is already the good one. And this good news comes at a price. While you do not know what is in your bottle, it is certainly worse than whatever is in the other bottle.

III.1. Strict and Weak Causal Dominance. Unlike EDT, CDT recommends switching in Bottles 3. And this time, GR agrees.[22] In recommending this, CDT and GR agree on a paradigmatic instance of:

> STRICT DOMINANCE: A-ing is preferable to B-ing if, for every admissible dependence hypothesis K, v(KA) > v(KB).

A dependence hypothesis is *admissible* if it passes the epistemic threshold for being taken seriously in deliberation. Without the restriction to admissible hypotheses, Strict Dominance would have limited interest. In Bottles 3, for example, it is metaphysically possible for for your bottle to have the more benign contents. But Strict Dominance still says you should switch, since by stipulation you are in an epistemic position to exclude such a possibility from consideration.

So when is a dependence hypothesis K admissible? The usual answer favored by CDT says it is when you cannot rule out K with certainty—that is, when Cr(K) > 0. But GR also upholds Strict Dominance on a higher standard for admissibility, where K is admissible with respect to A-ing and B-ing only if either Cr(K|A) > 0 or Cr(K|B) > 0.[23] By this standard, you do not need to be *unconditionally* certain K is false for it to be inadmissible. It is enough for you to be *conditionally* certain K is false under the assumption you adopt either A or B. Consider:

> BOXES 2: You must select one of three boxes, A, B, and C. An infallible Predictor set their contents by one of two schemes:
> S1: $1 in A, $0 in B, $0 in C

[21] For review, see Weirich, "Causal Decision Theory," *op cit.*

[22] U(stay|stay) - U(switch|stay) < 0 < U(switch|switch) - U(stay|switch).

[23] Quick proof: If v(KA) > v(KB) for every K such that Cr(K|A) > 0, then U(A|A) - U(B|A) > 0. And if v(KA) > v(KB) for every K such that Cr(K|B) > 0, then U(B|B) - U(A|B) < 0.

> S2: \$0 in A, \$1 in B, \$0 in C
>
> If you were predicted to take A or B, scheme 1 was used. If C, then scheme 2.

Since Cr(S2) > 0, scheme 2 is admissible under the usual, CDT-friendly standard. So under that standard, taking B does not violate Strict Dominance. Supporters of CDT should approve, since CDT arguably recommends taking B if you think you probably will take C. But by my ratificationist lights, there is something objectionable about taking B over A. You are sure that if you take either A or B, the money is in A. And that is what GR recommends, since it upholds Strict Dominance even on the higher standard for admissibility. Since Cr(S2|A) = Cr(S2|B) = 0, scheme 2 is not admissible with respect to boxes A and B. So on the higher standard for admissibility, Strict Dominance requires taking A over B.

The higher the standard for admissibility, the stronger Strict Dominance will be. Since GR upholds Strict Dominance on a higher than usual standard, it also upholds it on the usual standard. But that is not so for:

> WEAK DOMINANCE: A-ing is preferable to B-ing if
>
> (i) for every admissible dependence hypothesis K, v(KA) ≥ v(KB), and
>
> (ii) for some admissible K, v(K A) > v(K B).

As with Strict Dominance, GR upholds Weak Dominance on a higher than usual standard for admissibility, where K is admissible only if Cr(K|A) > 0 or Cr(K|B) > 0.[24] This higher standard makes (i) easier to satisfy, and correspondingly strengthens Weak Dominance.[25] But it also makes (ii) harder to satisfy, correspondingly weakening Weak Dominance. As a result, we will see that GR does not uphold Weak Dominance as usually understood, with the CDT-friendly standard for admissibility.

Given this CDT-friendly standard, Weak Dominance is already known to be inconsistent with a ratificationist cousin of GR, Ralph Wedgwood's Benchmark Theory.[26] Some of its issues turn on idiosyncrasies that GR does not share.[27] But I think the ones pressed by Ray Briggs do come down to whether we understand admissibility in a CDT-friendly or ratificationist-friendly way.[28]

Briggs endorses a dominance-like principle that Briggs compares to a requirement of Pareto optimality involving votes among one's possible future selves. In my notation, it says that A-

[24] Quick proof: If (i), then U(A|A) - U(B|A) ≥ 0 and U(B|B) - U(A|B) ≤ 0. If both (i) and there is some K such that Cr(K|A) > 0 and v(KA) > v(KB), then U(A|A) - U(B|A) > 0. If both (i) there is some K such that Cr(K|B) > 0 and v(KA) > v(KB), then U(B|B) - U(A|B) < 0.

[25] Compare Boxes 2 to a case with the following modification: If you were predicted to take A or B, a coin toss determined whether scheme 1 was used, or instead a scheme where \$1 is in A, \$1 in B, and \$0 in C. On the higher standard GR favors, Weak Dominance says, I think plausibly, that A is preferable to B.

[26] Wedgwood, "Gandalf's Solution to the Newcomb Problem," *op. cit.*

[27] See Bassett, "A Critique of Benchmark Theory," *Synthese*, CXCII, 1: 241-267, at Sec. 3.5.

[28] Briggs, "Decision-Theoretic Paradoxes as Voting Paradoxes," *op. cit.*, Sec. 6.

ing is preferable to B-ing if (i) for any option O, U(A|O) ≥ U(B|O), and (ii) for some O, U(A|O) > U(B|O). Briggs offers a proof that this Pareto principle entails Weak Dominance, but the proof assumes the usual, CDT-friendly standard of admissibility.[29] As Briggs notes, BT rejects the version of Weak Dominance entailed by her Pareto principle, and I grant the same goes for GR. But by my lights, this is not a defect of GR or BT. For it remains to be seen whether the alternative ratificationist-friendly standard makes Weak Dominance out to be too weak. And I don't think it does. Consider:

> BOXES 3: You must select one of three boxes, A, B, and C. An infallible Predictor set their contents by one of two schemes:
> S1: $1 in A, $1 in B, $0 in C
> S2: $1 in A, $0 in B, $0 in C
> If you were predicted to take A or B, then scheme 1 was used. If C, then scheme 2. You are not sure what you will do.

Is A preferable to B? Briggs's Pareto principle says so, and so will Weak Dominance on the usual standard for admissibility. Since Cr(S2) > 0, scheme 2 will count as admissible by that standard, and so taking A will weakly dominate taking B. To some this might seem like the right result. But by my ratificationist lights, it is hardly obvious. After all, Cr(S2|A) = Cr(S2|B) = 0. You can be certain that if you take either A or B, then they contain the same amount. GR takes this to mean that neither is preferable to the other.[30] While this result might be debatable, it does not seem to me an independently implausible one that GR had better avoid.

III.2. Alleged Counterexamples to CDT and Strict Dominance. Perhaps GR's partial agreement with CDT makes it vulnerable from the other direction, for agreeing too much with CDT. Ingenious counterexamples have recently been alleged for CDT and even Strict Dominance, by Arif Ahmed, Jack Spencer, and Ian Wells.[31] Here is one adaptation:

> GOLD SWITCH: You will soon decide between two boxes, A and B. A Predictor set their contents by one of two schemes:
> S1: $0 in A, $100 in B
> S2: $100 in A, $0 in B
> If you were predicted to take A, then scheme 1 was used, and if B, then scheme 2. But before deciding, you can pay $1 to flip a gold switch, which will cause you to deliberate unpredictably between boxes. (You might select a

[29] Quick version: Weak Dominance follows from Briggs' Pareto principle (via strengthening the antecedent), given two further assumptions: (1) If for every admissible K, v(KA) ≥ v(KB), then for every O, U(A|O) ≥ U(B|O); and (2) If (1) and if for some admissible K, v(KA) > v(KB), then for some O, U(A|O) > U(B|O). But (1) and (2) are false if admissibility requires Cr(K|A) > 0 or Cr(K|B) > 0.

[30] U(A|A) - U(B|A) = 0 = U(B|B) - U(A|B).

[31] Arif Ahmed, "Dicing With Death," *Analysis* LXXIV, 4 (2014): 587-592; Jack Spencer, "CDT and the Guaranteed Principle," *Analysis* (forthcoming); and Jack Spencer and Ian Wells "Why Take Both Boxes?" *Philosophy and Phenomenological Research* XCIX, 1 (2019): 27-48.

> box by an indeterministic process,[32] or just deliberate using brain regions the Predictor has not scanned.[33])

The Predictor laid a trap for you, by leaving empty whichever box your deliberations are on track to lead you to. By flipping, you might avoid the trap. I agree with Ahmed that you should do it.[34]

Recommending flipping does not commit us to funny backwards causation, however. Flipping cannot retroactively change what the Predictor has done. If your deliberations will be unpredictable, the Predictor already has failed to predict them. But flipping still can change what you do, by changing your future deliberations. Now supposing you will flip anyway, this change offers no expected advantage. Since the Predictor has already failed to lay a trap, you could save a dollar by just picking a box. So the degree of ratifiability of flipping is $U(\text{flip}_{\text{gold}} | \text{flip}_{\text{gold}}) - U(\text{refrain}_{\text{gold}} | \text{flip}_{\text{gold}}) = 49 - 50 = -1$. But supposing you will refrain, the Predictor has set a trap, and you are about to walk into it. If so, flipping could help. By flipping, you could change the course of your deliberations, and perhaps reach a different decision about which box to take. This makes the degree of ratifiability of refraining $U(\text{refrain}_{\text{gold}} | \text{refrain}_{\text{gold}}) - U(\text{flip}_{\text{gold}} | \text{refrain}_{\text{gold}}) = 0 - 49 = -49$.

Thus GR recommends flipping. But again, this is not because flipping can retroactively change what the Predictor has done. It is because it might change what you are about to do. Contrast this with:

> TIN SWITCH: You will soon decide between two boxes, A and B. A Predictor set their contents by one of two schemes:
> S1: \$0 in A, \$100 in B
> S2: \$100 in A, \$0 in B
> If you were predicted to take A, then scheme 1 was used, and if B, then scheme 2. But before deciding, you can pay \$1 to flip a tin switch. The tin switch is not hooked up to anything, and flipping it has no effects. But flipping still amounts to evidence about whether your other deliberations already are predictable, with flipping conclusive evidence of unpredictable deliberations, and refraining conclusive evidence of predictable ones. (Maybe flipping is preferred by erratic people who also select boxes by indeterministic processes, or by people already deliberating with the unscanned brain regions.)

[32] Cf. Ahmed, "Dicing With Death," *op. cit.*

[33] Cf. Spencer and Wells's Semi-Frustrater, in "Why Take Both Boxes?", *op. cit.*

[34] Ahmed, "Dicing With Death," *op. cit.* While I adapt Ahmed's example for continuity with others in this section, GR also agrees about his original counterexample to CDT, in which the randomization option is available simultaneously with the two unratifiable ones.

Flipping would bring good news that the Predictor was unable to lay a trap for you, but it would accomplish nothing. You should not pay $1 for nothing.[35]

Flipping tin might even be prohibited by Strict Dominance. Arguably you in effect have four diachronic options: flip and take A, refrain and take A, flip and take B, and refrain and take B. If so, Strict Dominance prohibits flipping. For wherever the money is, you are worse off flipping and taking A than refraining and taking A. So by Strict Dominance, you should not flip and take A. Likewise, you should not flip and take B. So you should not flip at all.[36]

But whatever we say your options are in Tin Switch, in Gold Switch you face two independent decisions. Any (predictable) decision you reach now between boxes might be reconsidered as a result of flipping the gold switch. If not, flipping gold could only bring good news, not good results. And it is independently plausible that you do not now have ϕ-ing among your options if you know you might not ϕ even if you decide to.[37] So Strict Dominance does not prohibit flipping gold.[38]

Consider one last example:

> SILVER SWITCH: You will soon decide between two boxes, A and B. A Predictor set their contents by one of two schemes:
> S1: $0 in A, $100 in B
> S2: $100+x in A, $0+x in B
> If you were predicted to take A, then scheme 1 was used, and if B, then scheme 2. But before deciding, you can pay $1 to flip a silver switch. Your deliberations over whether to flip are unpredictable, but your unpredictability will wear off before you can decide between boxes. (Perhaps you will switch

[35] Why does GR not entail you should flip the tin switch, as it did for gold? It is because $\text{U}(\text{flip}_{\text{gold}}|\text{refrain}_{\text{gold}}) = 49 \neq -1 = \text{U}(\text{flip}_{\text{tin}}|\text{refrain}_{\text{tin}})$. Whichever switch is offered, you can be sure that, if you refrain from flipping, you are about to select whichever box is empty. So we can partition the dependence hypotheses into $\text{K}_{=}$, those where flipping would cause you to pick a different box from what you will in fact pick, and, $\text{K}_{\neq}$, those where it would not. So $\text{U}(\text{flip}_{\text{gold}}|\text{refrain}_{\text{gold}}) = \text{Cr}(\text{K}_{=}|\text{refrain}_{\text{gold}})\text{v}(\text{K}_{=}\text{flip}_{\text{gold}}) + \text{Cr}(\text{K}_{\neq}\text{refrain}_{\text{gold}})\text{v}(\text{K}_{\neq}\text{flip}_{\text{gold}}) = 1/2(-1) + 1/2(99) = 49$. Meanwhile, $\text{U}(\text{flip}_{\text{tin}}|\text{refrain}_{\text{tin}}) = \text{Cr}(\text{K}_{=}|\text{refrain}_{\text{tin}})\text{v}(\text{K}_{=}\text{flip}_{\text{tin}}) + \text{Cr}(\text{K}_{\neq}\text{refrain}_{\text{tin}})\text{v}(\text{K}_{\neq}\text{flip}_{\text{tin}}) = 1(-1) + 0(99) = -1$. The crucial difference is that you know flipping tin would not change which box you choose, but flipping gold might.

[36] Cf. Spencer and Wells, "Why Take Both Boxes?" *op. cit.*, though the options in their Semi-Frustrater are synchronic, and Spencer, "CDT and the Guaranteed Principle," *op. cit.* rejects diachronic options. For an argument favoring diachronic epistemic options, see Barnett, "Internalism, Stored Beliefs, and Forgotten Evidence," *op cit.*

[37] Brian Hedden, "Options and the Subjective Ought," *Philosophical Studies* CLVIII, 2 (2012): 343-360; and John Pollock, "Rational Choice and Action Omnipotence," *Philosophical Review* CXI, 1 (2002): 1-23. But for problems see Marušić, *Evidence and Agency*, *op. cit.*

[38] Another wrinkle, potentially relevant to Spencer and Wells's Semi-Frustrater: Suppose that *deciding* on a costly action (like flipping a switch, or using a particular hand) causes you to deliberate unpredictability between boxes, but actually performing the action is merely evidence your deliberations were already unpredictable. Is it rational to decide to do it? For discussion of cases where it is advantageous to decide (or intend) to act disadvantageously, see Pamela Hieronymi, "The Wrong Kind of Reason," *Journal of Philosophy* CII, 9 (2005): 437-457 and Gregory S. Kavka, "The Toxin Puzzle," *Analysis* XLIII, 1 (1983): 33-36.

> brain regions.) Flipping destroys box A, forcing you to later take B—even if you would have (and were predicted to) take A.

If you are going to take A when given a choice, then flipping can foil the Predictor's trap. So GR recommends flipping, unless you are confident you would take B anyway. Should you be confident of this? Not necessarily. At least when $x = \$0$, the choice between boxes is arbitrary. Absent further evidence, you might and arguably should regard either box as equally likely to be chosen. And the same goes when $x > \$0$. In that case taking B is good news, but that is no reason to take B—and no reason now to expect yourself to if you expect to choose rationally. So you might rationally flip, no matter the value of x.

This contradicts Spencer's Guaranteed Principle regarding decisions over which options to have available later.[39] Say you *force* an outcome if you leave yourself no option that avoids it, and *guarantee* an outcome if you leave yourself at least one option known to have it. The Guaranteed Principle says guaranteeing a better outcome is always preferable to forcing a worse one. Yet when $x = \$100$, flipping forces \$99, while refraining guarantees \$100. So the Guaranteed Principle prohibits flipping.

But why accept the Guaranteed Principle? It is not because you can always decide now to take the guaranteed option later. If so you could skip the silver switch, and just decide to later take B for free. But the problem is that these decisions are independent, like in Gold Switch. You cannot now (unpredictably) decide to later (predictably) take B. Spencer grants this much, and even thinks decisions at different times always are independent.

Instead, Spencer supports the Guaranteed Principle on the grounds that you should now trust yourself to later select the most choiceworthy option. More specifically, let's say you are an *expected loser* if your (causally) expected utility of having some options is less than that of any one of the options. By forcing a worse outcome rather than guaranteeing a better one, you act like an expected loser, preparing in advance for the bad decision you expect later.

But I think expected losers are more common than Spencer does. He makes allowances for one kind, who expects future irrationality. When Ulysses ties himself to his ship's mast to constrain his later options, it is because he expects an irrational decision. But I think unstable decisions create expected losers too, even when no irrationality is expected. Sometimes when a malicious Predictor is out to trap you, you can even expect to lose no matter what you will do.

Here is why this matters. For any stable decision between two options, GR agrees with CDT that you should prefer the one with higher (causally) expected utility. When one of the options is to later decide between further options A-ing and B-ing, the expected utility of this option is $\Pr(A)U(A|A) + \Pr(B)U(B|B)$, by the probability calculus. What if this further decision is unstable, as in Silver Switch? Theories like GR and CDT can disagree about what is rational to do, and so can disagree about what now to expect if you expect to be rational. If so, they might differ over the expected utility of an unstable decision, by differing over $\Pr(A)$ and $\Pr(B)$. But apart from these disagreements, agents in many unstable decisions are uncontroversially expected losers, just because $U(A|A) < U(A)$ or $U(B|B) < U(B)$. For

[39] Spencer, "CDT and the Guaranteed Principle," *op. cit.* But note GR agrees with Spencer and not CDT regarding the Frustrater. $U(A|A) - U(\text{Envelope}|A) = 0 - 40 < 40 - 50 = U(\text{Envelope}|\text{Envelope}) - U(A|\text{Envelope})$, and similarly for box B.

example when x = 0 in Silver Switch, U(A) = U(B) = $50, while Pr(A)U(A|A) + Pr(B)U(B|B) = $0. Both boxes are permissible, but you still expect to take whichever is empty. Similarly when x = 100, U(B) = $100 > $50 = Pr(A)U(A|A) + Pr(B)U(B|B). While B contains $100, you expect to (permissibly) take A if it contains $0. So you should prefer to force a $99 outcome, despite knowing $100 will be available.

IV. Independence of Irrelevant Alternatives and Benchmark Theory

The above decisions involve two options at a time. But it is plausible that when many options are available simultaneously, the preferability between any two is unaffected by the others. This might be reinforced by the following feature of deliberation. When deciding between many options, it is common to narrow things down, *excluding* some options from consideration to focus on others. Sometimes this might involve deciding against the excluded options, but it might just mean provisionally setting them aside. You might ask yourself "If the decision is down to A-ing and B-ing, which shall it be?" And arguably, your answer should settle which to prefer in your actual decision between many options, so that:

> Independence of Irrelevant Alternatives (IIA): A-ing is preferable to B-ing in a many-option decision just in case A-ing is preferable to B-ing in a corresponding decision excluding other options.

In simple cases, IIA seems appealing. Suppose Sidney orders apple pie when offered apple, blueberry, and cherry. When he then finds out cherry was unavailable anyway, he changes his order to blueberry. Sidney's flip-flopping seems irrational. The availability of cherry should not affect the preferability of apple to blueberry.

Now some ratificationist predecessors of GR have been criticized for violating IIA.[40] And below, I will join with the critics. But I think GR is compatible with IIA, depending on how we understand exclusion.

On one understanding, an option is excluded from a decision only when it is unavailable to the agent. But this leads IIA into trouble when the availability of an option is relevant evidence. Suppose that when invited for tea or cocaine by an acquaintance, strait-laced Amartya prefers to decline. But if the invitation was merely for tea, he might prefer to accept. This is plausibly rational because the offer of cocaine is evidence against the dependence hypothesis that tea would be a dignified affair.[41]

Updating on the availability of cocaine affects Amartya's unconditional credences in the relevant dependence hypothesis. But in other cases it can affect the conditional credences that GR sees as relevant, as in:

> Perilous Seat: Galahad is deciding whether to sit at the Perilous Seat at the Round Table. He would like to sit, but he knows any unworthy knight who does so will die. Galahad is confident enough he is worthy to make sitting rational, if not for one final detail: Percival is captured, and Galahad

[40] For discussion, see Eells and Harper, "Ratifiability, Game Theory, and the Principle of Independence of Irrelevant Alternatives," *op. cit.*, and especially Bassett, "A Critique of Benchmark Theory," *op. cit.*

[41] The example is from Amartya Sen, "Internal Consistency of Choice," *Econometrica* LXI, 3 (1993): 495-521.

> can rescue him. And Galahad is certain a worthy knight would drop everything and rescue Percival immediately.

If not for Percival, sitting in the Perilous Seat would be preferable to refraining. But because of Percival, GR says refraining is preferable. That is not because the availability of rescuing Percival is direct evidence about Galahad's worthiness. Instead, it makes Galahad's sitting amount to evidence he is unworthy. As with Psychopath Button, refraining is arguably preferable.

The trouble with Amartya and Galahad is that when their options change, their credences change too. This might be unavoidable, depending on the connection between having an option and knowing it is available.[42] But even if the available options can change without one's knowing it, IIA would be unsatisfyingly weak if it just said preferability is unaffected by changes like that. So there may be no satisfying way to save IIA if excluding an option means making it unavailable.

But there is a better way to understand exclusion. When you deliberate between many options, narrowing down your options does not require supposing you are in a distinct situation where you cannot select the excluded options, but rather that in your actual situation you will not select them. So for a many-option decision including A-ing and B-ing as options, the corresponding restricted decision is not one where your credences are updated on your having no options besides A-ing and B-ing, but rather on your not selecting the others. Most of the time the difference can be harmlessly ignored, but with Galahad and Amartya it matters. It also matters for the compatibility of IIA and GR. For under GR, the preferability of A-ing to B-ing depends on the expected utilities of these actions conditional on one's A-ing, and conditional on one's B-ing.

IV.1. Benchmark Theory. Ralph Wedgwood's *Benchmark Theory (BT)* is another broadly ratificationist view, motivated by Psychopath Button and the like. But Wedgwood opposes IIA, and his opposition runs deeper than some quibbles about how to understand exclusion. I think this raises problems GR avoids, but before getting to them, I want to highlight our substantial agreements.

Wedgwood's idea is to 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) is the benchmark for a dependence hypothesis K, the comparative value of of A-ing under K, or CV(A,K), is simply v(KA) - B(K).

To a first approximation, BT says that one option is preferable to another iff it has higher *evidentially expected comparative value*, which is determined by taking a weighted average of an option's comparative value under each dependence hypothesis. Importantly, the weighting follows one's conditional credence in the dependence hypotheses given that one selects the relevant option, rather than one's unconditional credences. Thus the evidentially expected comparative value of A-ing, or EECV(A), is:

[42] Cf. Pollock, "Rational Choice and Action Omnipotence," *op. cit.*

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

So how are we supposed to set the benchmarks? One method Wedgwood considers is simply *averaging* the values of each of one's options. Suppose one has two options, A and B. Since CV(A,K) = v(KA) - B(K), setting B(K) by averaging gives us:

$$CV\left(A,K\right) = v\left(KA\right) - \frac{v\left(KA\right) + v\left(KB\right)}{2},$$

and thus,

$$CV\left(A,K\right) = \frac{v\left(KA\right) - v\left(KB\right)}{2}.$$

So in two-option cases, the averaging method has BT agree with GR. Since BT says to maximize evidentially expected comparative value, it says that B-ing is preferable to A-ing iff

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

This reduces to:

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

which is equivalent to GR's condition that U(A|A) - U(B|A) < U(B|B) - U(A|B).

My disagreements with BT concern many-option cases. Roughly speaking, GR has the preferability of one option to another depend on a direct comparison between them. But under BT it depends 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 implausible influence.

IV.2. Redundant Options and IIA. Before getting to my main objections, let's consider a related one that Wedgwood anticipates, the *dreadful options problem*. At any given time, one has many options 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. Including the dreadful options does no harm in orthodox expected utility theory or GR. The problem for BT is that dreadful options can affect the relevant benchmarks, and thus affect which of one's non-dreadful options are preferable.

Wedgwood's solution is to have dreadful options excluded when setting benchmarks. But I think this does not get to the root of the problem. For there is a related *redundant options problem*. Consider:

> BOXES 4: You must select one of three boxes, A, B, and C. An infallible Predictor set their contents by 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 take A, then scheme 1 was used. If B or C, then scheme 2.

Boxes B and C are redundant, in the sense of being certain to have the same effects. So excluding C from the decision should not affect the preferability between B and A. But BT says otherwise. Like GR, it recommends B when C is excluded.[43] But unlike GR, BT recommends A when C is included, at least if benchmarks are set via averaging.[44] This verdict is unattractive on its own. But what is worse is having the preferability of A to B depend on whether another option just like B is included. Indeed, even if we think averaging is merely a permissible method for setting benchmarks, it seems wrong to say that taking A is even permissible with C included but not with C excluded.

The unattractiveness does not depend on my favored reading of IIA, where an option can be excluded from a decision despite still being available to the agent. For unlike with Amartya and Galahad, flip-flopping seems irrational even with known changes in your available options. Suppose in Boxes 4 you are initially prepared to take A, as BT permits. Then you realize C is out of reach, and switch to taking B. As you prepare to take B, you realize it can be reached with either your right hand or left hand, giving you two options that involve taking B. Meanwhile, A is reachable only with your left hand. And so you switch back to A. These changes in preference seem irrational, but BT apparently licenses them.

The problem is not limited to options that are redundant in the sense of being certain to have the same effects. There is a further kind of redundancy exhibited here:

> BOXES 5: You must select one of three boxes, A, B, and C. An infallible Predictor set their contents by 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 take A, then scheme 1 was used if the coin landed heads, and scheme 2 if tails. If you were predicted to take B or C, then scheme 3 was used if the coin landed heads, and scheme 4 if tails.

Excluding C from the decision, BT recommends B.[45] So far, so good. But with C included, and with benchmarks set via averaging, BT recommends A.[46] This means BT cannot avoid the redundant options problem by individuating options in a coarse-grained way, which groups together options that are guaranteed to have the same effects. It also will not do for BT to set benchmarks by averaging over equivalence classes of options, for example by

[43] For example, if benchmarks are set via averaging, then EECV(A) = 1(1) + 0(-1.5) = 1 < EECV(B) = 0(-1) + 1(1.5) = 1.5.

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

[45] 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.

[46] 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.

grouping together B and C in Boxes 5. For if B and C are equivalent here, presumably this is something we want our decision theory to explain, not simply presuppose.

So far I have been assuming that benchmarks are set via averaging. Wedgwood regards this as a reasonable method, but he also proposes others. There is the *relief method*, which sets the benchmark for a dependence hypothesis at the worst possible outcome under that hypothesis. And there is the *regret method*, which sets the benchmark at the best possible outcome. Wedgwood ultimately says an option is permissible if licensed by any of these methods. The only restrictions are that for a given decision the same method is employed regarding each dependence hypothesis, and that the method assigns benchmarks no lower than relief and no higher than regret.

The resulting view is fairly permissive, arguably too much so. Recall that in Bottles 2 and Lazy Death 2, staying seems to have a clear advantage over the alternative. But BT now says any option is permissible, since benchmarks may be set via regret. And recall in Bottles 1 and Lazy Death 1, staying again seems to have an advantage. Again BT says any option is permissible, since benchmarks may be set via relief. Indeed, pressing even gets counted as permissible in Psychopath Button.

Setting aside excessive permissiveness, I think the additional methods raise new problems without solving the old ones. I will illustrate using the regret method, but corresponding points go for relief. The old redundant options problem remains because of cases like:

> BOXES 6: You must select one of three boxes, A, B, and C. An infallible Predictor set their contents by one of three schemes:
> 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
> But the Predictor used a chancy method for picking a scheme. If taking 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.

As before, BT permits taking A, so long as C is included. For with C included, A has the highest evidentially expected comparative value when benchmarks are set via regret.[47] Indeed, Wedgwood defends the permissibility of taking A in a related case from Briggs.[48] Meanwhile, taking A surely would be impermissible if C were unavailable or otherwise excluded. BT gets this right, since A would have lower evidentially expected comparative value than B using any permissible benchmarks.

But it is unappealing to allow the permissibility of taking A over B to depend on the inclusion of C. If A and B were the only options, taking B would have a clear advantage over

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

[48] Briggs, "Decision-Theoretic Paradoxes as Voting Paradoxes," *op. cit.*

A. Adding yet another option with the same advantage over A should not make A more attractive.[49]

IV.3. Further Problems for BT. Setting IIA aside, BT also faces the *evidential sweetening problem*. There are potential changes in your evidential situation that intuitively favor taking A in Boxes 6, but which BT treats instead as reasons to prefer B. Suppose that as you prepare to take A, you learn that the Predictor was unavailable today. So the money was just distributed via a chancy process, with scheme 1 standing a 90% chance of being used, and schemes 2 and 3 each a 5% chance. All of the theories we have considered now say you should prefer B to A, including BT. But this new evidence can only sweeten the prospects of taking A, by raising some conditional and unconditional probabilities that A has more money than B. If you still should take B, as everyone agrees, then you should have in the original version of Boxes 6, contrary to BT.

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

> BOXES 7: You must select one of three boxes, A, B, and C. An infallible Predictor set their contents by one of three schemes:
> 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[50]
> The Predictor used the same chancy method as in Boxes 6.

BT says—and I agree—that you should prefer B to A.[51] But look what has changed from Boxes 6. In Boxes 6, if you take B, you can regard it as 5% likely that you'll get \$100. In Boxes 7, 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 take B in Boxes 6, you should think it likely that you would have been better off taking 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.

V. PREFERABILITY CYCLES AND RATIONAL DILEMMAS

While GR upholds a reasonably strong IIA, this commits it to preferability cycles in certain cases, such as the following from Arif Ahmed:[52]

> PSYCHOPATH CYCLE: Before you are a button and a lever, both marked "KILL ALL PSYCHOPATHS". But above the lever is a further message:

[49] EECV(A) = .9[1-B(S1)] + .05[0-B(S2)] + .05[0-B(S3)] = .9 - .9[B(S1)] - .05[B(S2)] - .05[B(S3)]. EECV(B) = .05[0-B(S1)] + .9[1-B(S2)] + .05[100-B(S3)] = 5.9 - .05[S(B1)] - .9[S(B2)] - .05[S(B3)]. Since B(S1) ≥ 0 and, with C unavailable, B(S2) ≤ 1, this means EECV(A) < EECV(B).

[50] For convenience I count S2 and S3 as distinct, but nothing hangs on it.

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

[52] Ahmed, "Push the Button," *op. cit.* See also Hare and Hedden's Three Crates in "Self-Reinforcing and Self-Frustrating Decisions," *op. cit.*, and Gallow's Improvement Cycle, in "The Causal Theorist's Guide to Managing the News," *op. cit.*

> "WARNING: MILD ELECTRICAL SHOCK FOR PULLING". You can press the button, pull the lever, or refrain altogether. You are still sure that only a psychopath would press the button, and for whatever reason you regard pulling the lever as no evidence of psychopathy. (Perhaps psychopaths are averse to electrical shocks.)

Which option is most preferable in Psychopath Cycle? Whatever else we say, pulling the lever must be preferable to refraining. It would be rational to press in Psychopath Button if not for the evidence pressing would give you of your own psychopathy. That's the *whole point* of the example. And I hereby stipulate that the electrical shock is too mild to make a difference.

Moreover, pressing the button must be preferable to pulling the lever, since pressing strictly dominates pulling. Whether you press or pull has no effect on your psychopathy, so if you pull, you are shocking yourself just to avoid bad news.

It will follow that refraining is not preferable to pressing if:

> ACYCLICITY: If A-ing is preferable to B-ing, and B-ing is preferable to C-ing, then C-ing is not preferable to A-ing.

But GR says that refraining is preferable to pressing, as in Psychopath Button. So either Acyclicity must go, or else GR.

Call me crazy, but I think it should be Acyclicity that goes. If refraining is preferable to pressing in Psychopath Button, it is in Psychopath Cycle, too. For your preference between pressing the button and refraining should not flip-flop, depending on whether the lever is available. The availability of the lever is no evidence against your psychopathy, nor does it weaken the evidence pressing provides for psychopathy. So compared to refraining, pressing does not look any better on account of the lever.

Now Acyclicity will be upheld by any theory that has preferability follow linearly ordered expected utilities, like EDT and CDT. But there are independent reasons for doubting that (rational or moral) preferability must linearly order one's options, stemming from cases of incomparable values and supererogation.[53] And while I am myself less confident about them, there are more direct motivations for preferability cycles arising from the puzzle of the self-torturer.[54] These cases raise issues far removed from those here, but they illustrate the unobviousness of the usual view of preferability as linearly ordering one's options. If we are going to reject GR for violating Acyclicity, we need a more specific reason.

V.1. Money Pumps. One such reason is that preferability cycles allegedly render you exploitable, as in:

[53] Ruth Chang, "The Possibility of Parity," *Ethics* CXII (2002): 659-688; Daniel Muñoz "Three Paradoxes of Supererogation," *Noûs* (forthcoming); and Wlodek Rabinowicz, "Money Pump with Foresight," *Imperceptible Harms and Benefits*, Ed. Michael J. Almeida (Springer, 2000).

[54] For discussion, see Toby Handfield, "Rational Choice and the Transitivity of Betterness," *Philosophy and Phenomenological Research* LXXXIX, 3 (2014): 584-604.

> PSYCHOPATH MONEY PUMP: Before you are a button and a lever, each marked "KILL ALL PSYCHOPATHS". Initially you have no plans to pull or press. At Stage 1, you are offered to switch plans for one penny to pulling the lever. You know that iff you switch, you will at Stage 2 be allowed to pay another penny to switch plans to pressing the button. And if you switch, at Stage 3 you will again be allowed to pay a penny to switch to refraining. You are certain that only a psychopath would at any stage plan to press the button, and that most psychopaths are bloodthirsty enough to stick to a plan to press, even if it means killing themselves.

If preferability is cyclic in Psychopath Cycle, the objection goes, then it will be here, too. But that would mean you ought to pay three cents to end up back where you started. And you shouldn't do that.

There is a standard response to money pump arguments like this. It assumes that if you are rational you will have *foresight*—that is, that you will be aware of your cyclic preferences, and thus foresee at early stages what awaits later on.[55] In particular, you will foresee at Stage 1 that if you switched to pulling, you would not stick to it. Instead, you would switch again at Stage 2, and then again at Stage 3. So at Stage 1, refraining can be preferable to *planning* to pull, even if not to pulling.

But it is doubtful the standard defense ultimately succeeds even when you have foresight.[56] And even if it does, it is irrelevant to many cases that concern us here. If we could always assume foresight, GR would be of little interest. For GR disagrees with CDT only when your conditional causally expected utilities differ from your unconditional ones. This means GR departs from CDT by recommending cyclic preferences only when you cannot know enough about yourself to foresee your future choices.

So I grant that GR yields preferability cycles at Stages 1 and 2, and also that it licenses being exploited. What I deny is that it licenses being exploited *because* it yields preferability cycles. Your exploitability is ensured instead by some distinct commitments that it shares with competing views including CDT.[57] The relevant commitments are not about preferability cycles at any one stage, but rather across the three stages. And this commitment is unavoidable if we allow failures of foresight.

If you lack foresight, any plausible view recommends switching at Stage 1. Pulling is preferable to refraining, so you should switch to pulling if you mistakenly think you will stick with it. Switching also is preferable at Stage 2 if you lack foresight. Pressing strictly dominates pulling, so GR and CDT agree you should switch to pressing if you think you will stick. What about at Stage 3, when you can switch from pressing to refraining? GR controversially claims refraining was preferable all along, even when you doubted your psychopathy. But by now, you should realize you are a psychopath, since only psychopaths would find themselves at Stage 3 with the option to switch to refraining. And of course you

[55] See Rabinowicz, "Money Pump with Foresight," *op. cit.* for review.

[56] See *ibid.* and "A Centipede for Intransitive Preferrers," *Studia Logica* LXVII (2001): 167-178.

[57] See also Dmitri Gallow, "Escaping the Cycle," (MS).

should switch if you know you are a psychopath. This uncontroversial claim is what makes you exploitable, not GR's controversial one.[58]

Could the real lesson be that rational agents must foresee their future choices after all?[59] They might if the credences and values that will motivate one's future choices are luminous, but I won't rehash the the arguments against luminosity here.[60] Instead, I will say only that we opponents of luminosity should not be troubled by exploitability. Rational agents can lose bets due to factual ignorance, and Psychopath Money Pump just dresses this up in the trappings of exploitation. If you are an unwitting but rational psychopath, you will place a losing bet at Stage 1, and set yourself up to die. If things stopped there it would be a terrible outcome, but one rooted in ignorance, not irrationality. And it is no further sign of irrationality if in later stages you manage to partly recover your losses, after acquiring new evidence about your psychopathy.

V.2. Rational Dilemmas. A second problem is that preferability cycles generate rational dilemmas, at least given Permissibility. By its left-to-right direction, an option is permissible only if one has no other options preferable to it. Since in a cycle every option has another preferable to it, none will be permissible.

As we saw in Section I, admitting rational dilemmas seems tantamount to making rationality out to offer inconsistent guidance. And yet, as we also saw, it has plausibly been held that morality does just that,[61] and that rationality does too, in the epistemic domain.[62] And furthermore, if rationality offers inconsistent guidance anywhere, cases like these, where one's practical deliberations collide with predictions of one's actions, are among the likeliest candidates.[63] So despite some misgivings, I think we should be open to dilemmas.

More worrisome to my mind are apparently asymmetrical cycles, like:[64]

[58] To be sure, GR but not CDT would have you exploited if Stage 3 came first, though CDT still leaves you exploitable in many other cases. See, for example, Gallow "Escaping the Cycle," *op. cit.*, and Spencer, "CDT and the Guaranteed Principle," *op. cit.*

[59] Cf. Briggs, "Decision-Theoretic Paradoxes as Voting Paradoxes," *op. cit.*, Sec. 7 and Roy Sorensen, "Anti-Expertise, Instability, and Rational Choice," *Australasian Journal of Philosophy* LXV, 3 (1987): 309. And see also Sobel, "Self-Doubts and Dutch Strategies," *op. cit.*, pg. 69.

[60] For a classic discussion, see Williamson, *Knowledge and Its Limits*, *op. cit.*, Ch. 4. And for my take, see Barnett, "Self-Knowledge Requirements and Moore's Paradox," *op. cit.*

[61] Marcus, "Moral Dilemmas and Consistency," *op. cit.* and Sinnott-Armstrong, *Moral Dilemmas*, *op. cit.*

[62] Christensen, "Conciliation, Uniqueness, and Rational Toxicity," *op. cit.*, Hughes, "Dilemmic Epistemology," *op. cit.*, and Pryor, "The Merits of Incoherence," *op. cit.*

[63] For general discussion of prediction and deliberations, see Hájek, "Deliberation Welcomes Prediction," *op. cit*, Korsgaard, *The Sources of Normativity*, *op. cit.*, pp. 94-96, Levi, "Rationality, Prediction, and Autonomous Choice," *op. cit.*, Liu and Price, "Heart of DARCness," *op. cit.*, Marušić, *Evidence and Agency*, *op. cit.*, Rabinowicz, "Does Practical Deliberation Crowd Out Self-Prediction?" *op. cit.*, and Vavova, "Deliberation and Prediction," *op. cit.*

[64] Thanks to an anonymous referee for pressing this. For further discussion, see Gallow, "A Causal Decision Theorist's Guide to Managing the News," *op. cit.*

> BOXES 8: You must select one of three boxes, A, B, and C. An infallible Predictor set their contents by one of three schemes:
> S1: $10 in A, $100 in B, $0 in C
> S2: $0 in A, $10 in B, $100 in C
> S3: $100 in A, $0 in B, $10 in C
> If you were predicted to take A, then S1 was used. If B, then S2. And if C, then S3. Before you decide, a visible $3 bonus is placed on top of A, a $2 bonus on B, and $1 on C.

GR says preferability is cyclic, and in particular that B is preferable to A. For you know that if you will take either A or B, then despite A's larger bonus there is more money in B.

But it might seem myopic to let this settle the preferability of B to A. You also know that if you will take B or C, then C has more money, and that if you will take A or C, then A does. So for each option, there is another one sure to have more money if you will take either of them. Considered in this broader context, it might seem that all options in effect have the same strike against them, and that taking A has an advantage on account of its larger bonus. The problem for GR is that it might seem unable to capture this advantage.

Now it is true that GR does not allow A's larger bonus to affect which options are permissible, nor which are preferable to which. But this does not prevent the bonus from making it more attractive to take A. For comparison:

> BAD NEGOTIATOR: You ought to buy a car for an offered price, but while you deliberate, the salesperson preemptively offers the same car for a reduced price.

The price reduction does not affect which option is preferable or permissible, but it still makes buying *more* preferable to refraining that it was already. Likewise in the dilemmic Boxes 8, the GR theorist can say the bonus makes taking A more preferable to C, and less dispreferable to B—presumably by affecting the differences in comparative ratifiability.

Still, it might be objected this does not go far enough. The bonuses do not just make A less dispreferable to B than it was before, the objector might claim. They make it the best option you've got. So it must be preferable to your other options, including B. If B still comes out ahead in a direct comparison to A, the objection goes, that just shows direct pairwise comparisons are not the whole story about preferability. There must be some further procedure for translating these pairwise comparisons into an ultimate linear order—one that advantages options like A for being less dispreferable in the pairwise comparisons.

But long story short, to avoid obvious problems, the translation procedure has got to be awfully complicated. Dmitri Gallow has devised a more elegant proposal than anything I could come up with, but it's still a doozy.[65] And while it avoids obvious problems, it still has acknowledged implications that I see as serious drawbacks. For example, while it upholds Strict Dominance if restricted to two-option cases, it rejects it for many-option cases.

The rejection of Strict Dominance is not a dispensable part of Gallow's proposal. It is a prerequisite of any anti-dilemmic view about cyclic cases like Boxes 8. Each option is strictly

[65] *Ibid.*

dominated by another, at least given the conditional standard of admissibility discussed in Section III.1. For any box you consider, there is another which certainly offers more money if you will take either of them. This is a rather extreme situation to be in, and it is not obvious what we should say about it. But it is at least uncomfortable to straightforwardly recommend taking any given box.

It also means allowing unappealing violations of IIA. Unless we give up entirely on ratificationist intuitions in Psychopath Button and other cases, B surely is preferable to A if C is excluded from the decision. For then you could be sure B has more money. There is something awkward about letting this change when C is included. For you still know B offers more than A if you take either of them. It is arguably even more awkward if our theoretical goal is to linearly order your options. For surely if they are linearly ordered, C is the worst. So the preferability between B and A would have to flip once you narrow down your options to the two best.

While nothing is obvious about extraordinary cases like these, I think the best course is to admit the impossible position they put you in. In a cycle, even an asymmetrical one, there is no balancing the reasons available into a stable recommendation for action. For any option you consider, you have sufficient reason to take another one over it.

VI. Conclusion

In unstable decisions like Psychopath Button, one's actions amount to evidence about what their effects will be. A natural thought is that if so, you should not wait around until after acting to consider this evidence. For instance, you should not decide to press a button that will kill all psychopaths if you think only a psychopath would press it. It may be this natural thought is mistaken, and you can simply ignore evidential significance of your actions. But if not, GR offers the best theory of how to accommodate such evidence.[66]

[66] For valuable feedback, I would like to thank Arif Ahmed, Brendan de Kenessey, J. Dmitri Gallow, Ralph Wedgwood, Jonathan Weisberg, and three anonymous referees.