'ALONG AN IMPERFECTLY-LIGHTED PATH':

PRACTICAL RATIONALITY AND NORMATIVE UNCERTAINTY

by

ANDREW CHRISTOPHER SEPIELLI

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Philosophy

written under the direction of

Ruth Chang

and approved by

____

____

____

____

New Brunswick, New Jersey

January, 2010

ABSTRACT OF THE DISSERTATION

'Along an Imperfectly-Lighted Path': Practical Rationality and Normative Uncertainty

by ANDREW CHRISTOPHER SEPIELLI

Dissertation Director:

Ruth Chang

Nobody's going to object to the advice “Do the right thing”, but that doesn't mean everyone's always going to follow it. Sometimes this is because of our volitional limitations; we cannot always bring ourselves to make the sacrifices that right action requires. But sometimes this is because of our cognitive limitations; we cannot always be sure of what is right. Sometimes we can't be sure of what's right because we don't know the non-normative facts. But sometimes, even if we were to know all of the non-normative facts, we'd still not be sure about what's right, because we're uncertain about the normative reasons those facts give us. In this dissertation, I attempt to answer the question of what we're to do when we must act under fundamentally normative uncertainty.

It's tempting to think that, in such circumstances, we should do what we regard as most probably right. I argue that this view is mistaken, for it is insensitive to how degrees of actions' values compare across different normative hypotheses; if an action is probably right, but, if wrong, is terribly, terribly, wrong, it may be rational not to do that action. A better answer is that we should do the action with the highest expected value. I spend the first part of the dissertation providing arguments for and rebutting arguments against this view of action under normative uncertainty. I spend the next part of the dissertation explaining what degrees of value are, and showing how they can be compared across normative hypotheses. In the remaining parts of the dissertation, I consider two questions related to our primary question – first, what is one required, or obligated, to do under normative uncertainty; and second, what is it rational for one to do when one is not only normatively uncertain in the way we've been discussing, but also uncertain about what it is rational to do under this sort of normative uncertainty.

> “...many plain honest men really do think that they always know what their duty is – at any rate, if they take care not to confuse their moral sense by bad philosophy. In my opinion such persons are, to some extent, under an illusion, and really know less than they think....It is not to plain men of this type that our appeal is made, but rather to those whose reflection has made them aware that in their individual efforts after right living they have often to grope and stumble along an imperfectly-lighted path; whose experience has shown them uncertainty, confusion, and contradiction in the current ideal of what is right, and has thus led them to surmise that it may be liable to limitations and imperfections, even when it appears clear and definite.”
>
> Henry Sidgwick, “My Station and Its Duties”

Acknowledgments

Let me begin with some personal acknowledgments. Thanks to the baristas and regulars at the George Street Starbucks for making my writing environment so pleasant. Thanks to Mercedes Diaz and Susan Viola for indulging my frequent and only occasionally necessary visits to your offices. Thanks to my friends, especially Sophie Ban and Dan Lee, for your loyalty, and for your role in making the last few years the best ones of my life so far. Finally, thanks to my brother Matthew, and my mom, Jeannie, for supporting me unconditionally as I've groped and stumbled along the imperfectly-lighted path that's led to this point. I love you both very much.

Now for the more properly philosophical stuff. Some of the material in the dissertation is based on work I've previously published. The discussion of Ted Lockhart in Chapter 4 is based on Sepielli (2006), which appeared in Ethics. Parts of the “taxonomy” section of Chapter 1, as well as the second half of Chapter 4, are based on Sepielli (2009), which appeared in Oxford Studies in Metaethics, Vol. 4.

I received helpful comments on this dissertation from the following students and faculty at Rutgers: Geoff Anders, Saba Bazargan, Nick Beckstead, Tim Campbell, Pavel Davydov, Preston Greene, Jonathan Ichikawa, Alex Jackson, Michael Johnson, Ben Levinstein, Martin Lin, Barry Loewer, Jenny Nado, Josh Orozco, Derek Parfit, Jacob Ross, Ernest Sosa, Larry Temkin, Jonathan Weisberg, Dennis Whitcomb, and Evan Williams; and the following others elsewhere: Lara Buchak, Richard Chappell, Richard Fumerton, Desmond Hogan, Tom Hurka, Rae Langton, Toby Ord, Philip Pettit, Wlodek Rabinowicz, Peter Railton, Mark van Roojen, George Sher, and Jennifer Whiting.

I must extend special thanks to the members of my committee. Michael Smith generously stepped in at the last moment to serve as my external examiner, and raised some very stimulating questions at my defense. Brian Weatherson played a major role at two stages of the dissertation's development. He led my favorite class at Rutgers, the Spring 2008 Dissertation Seminar, and then agreed to serve on this committee. In both of these capacities, he presented me with deep and thoughtful challenges that led to substantial improvements of the dissertation. I was fortunate that Holly Smith returned to her roots as a full-time philosopher very shortly after I began the dissertation, fortunate as well that she agreed to join my committee, and still more fortunate that her interests intersected so nicely with mine. The arguments you're about to read would have much looser and more superficial but for Holly's extraordinarily detailed comments on preliminary drafts, and our discussions regarding our respective manuscripts-in-progress.

It would be difficult to exaggerate my debts to Ruth Chang. Before I settled on Ruth as an advisor, I had it in mind that the topic of this dissertation would form one chapter of a five-chapter dissertation. Three years later, I could see writing an entire dissertation on the ideas that consume a few pages of the monograph you have in front of you. Over the course of what must have been hundreds of meetings, phone conversations, and lengthy e-mails, Ruth helped me to see that real progress in philosophy must be earned the hard way, and that questions worth considering can rarely be treated in isolation from one another. She's given me more guidance and encouragement than I had any right to expect, and I hope she's been able to sense, even during my most stubborn and adversarial moments, how much I appreciate it.

Table of Contents

List of Tables

List of Illustrations

INTRODUCTION

You and I are imperfect beings, and must therefore make our decisions under uncertainty. There are two types of uncertainty with which we must contend. One is non-normative uncertainty – uncertainty about matters of non-normative fact. Non-normative facts may include everything from the age of the universe to the gross domestic product of India to the health effects of drinking four gallons of Mountain Dew in one night. The other is normative uncertainty – uncertainty about the reasons those facts give us. For example, someone might be uncertain about the permissibility of abortion, even if she were certain about the science of fetal development, the kind of life a child born to her would lead, and so on. Similarly, someone may be uncertain whether the reasons to support a tax increase outweigh the reasons to oppose it, even if she is sure about what the economic and social effects of the measure would be. At a more theoretical level, someone may be uncertain whether utilitarianism or Kantianism or contractualism or some other comprehensive account of morality is correct.

A good deal has been written on the issue of what we should do when we're non-normatively uncertain. Most of decision theory concerns rationality under non-normative uncertainty, and it's typically seen as incumbent upon ethicists to develop theories capable of guiding agents who are uncertain about the non-normative facts. By contrast, shockingly little has been written on the question of what we should do when we're normatively uncertain.[1] The explanation for this can't be that normative uncertainty is

[1] The only recent publications to address the issue are Hudson (1989), Oddie (1995), Lockhart (2000), Ross (2006), Guerrero (2007), Sepielli (2009), and Zimmerman (2009). A very similar debate – about so-called "Reflex Principles" – occupied a central

rare. It's common among ordinary people, and I would hope, pervasive among moral philosophers, who spend their lives engaging with sophisticated arguments and brilliant thinkers on all sides of moral questions. Perhaps the explanation is that one answer to the question strikes us as obviously correct – and so obviously correct that the question of whether it is correct does not even arise for most of us. This answer is: When you must act under normative uncertainty, you should simply do what you think is most probably best. If you think it's probably better to give than to receive, then you should give. If you think that the consequentialist solution to a moral problem is more probable than the deontological solution, then you should implement the consequentialist solution.

But I've come to think that this answer is mistaken – and so obviously mistaken that the only explanation for its persistence is that the question of whether it is mistaken does not even arise for most of us. The aim of this dissertation is to develop and defend an alternative: When you must act under normative uncertainty, you should do the act that has the highest expected value. The expected value of an act is the probability-weighted sum of its values given each of the different ways the world could be, respectively. The act with the highest expected value is not always, and not typically, even, the act that most probably has the highest value. We'll soon see how very different these two criteria are.

In the course of defending my view, I'll address some issues of general interest in practical philosophy – how to most perspicuously think about normative phenomena like

place in Early Modern Catholic moral theology. The most notable contributors to this debate were Bartolomé de Medina (1577), Blaise Pascal (1853), and St. Alphonsus Liguori (1755). The various positions are helpfully summarized in Prümmer (1957), The Catholic Encyclopedia (1913), and The New Catholic Encyclopedia (2002). I discuss this debate at the end of Chapter 1, and intermittently throughout the dissertation.

obligation, permission, strength of reasons, subjective vs. objective rightness, value incomparability, etc.; what it is for a normative theory to be “action-guiding”; the relationship between reasons and rationality, and between the different notions of rationality; the structural features of normative theories; theories about the semantics of normative claims – most notably expressivism and conceptual role semantics; and the ability of austere formal decision theory to represent the wild and wooly world of the normative. Here's a preview of what's to come:

In Chapter 1, I'm going to do three things. First, I'm going to say more precisely what normative uncertainty is, and demonstrate as conclusively as I can that there is such a thing. Second, I'm going to draw three distinctions – between objective and belief-relative normative features; between the strength of reasons for action and deontic statuses of actions like “required”, “supererogatory”, and so forth; and between absolute and comparative normative features – and explain the structure of the dissertation's arguments in terms of these distinctions. Third, once we have this structure in hand, I'll briefly review other philosophers' work on normative uncertainty, and explain how this dissertation contributes to that literature.

In Chapter 2, I'm going to present positive arguments for the view that we should maximize expected value when acting under normative uncertainty. First, I'll argue that views that are more or less like expected value maximization, which I'll call “comparativist” theories, are superior to views that are more or less like the one that we should do what's most probably best, which I'll call “non-comparativist” theories. Second, I'll argue that expected value maximization is better than all of the other comparativist theories. I'll conclude this chapter by considering cases where the expected values of actions are indeterminate. This cases arise when agents have some credence that actions are not related by any of the three standard value relations – better than, worse than, and equal to.

In Chapter 3, I'm going to confront a bevy of objections to expected value maximization. First, I'll consider the objection that it's too demanding a theory of rationality under normative uncertainty. Second, I'll consider the objection that it often asks us to behave in ways that are not true to ourselves. Third, I'll consider the objection that it's inconsistent with leading a life that's narratively coherent. Fourth, I'll consider objections that, because the expected values of actions are influenced disproportionately by extreme or absolutist normative hypotheses about their values, my view gives too much "say" to such hypotheses. Fifth, I'll consider the objection that it offers us no way of explaining why it's sometimes rational to engage in further normative deliberation when one is uncertain. Sixth, I'll consider the objection that acting in accordance with my view requires that one act from objectionable motivations. Seventh, I'll consider the objection that it doesn't even make sense to speak of maximizing expected value, because there's no univocal notion of value the expectation of which might be maximized. Eighth, I'll consider the objection that formal views like mine place fewer constraints on our behavior than one might have supposed.

All of these objections are to be taken seriously, but none provides the most daunting challenge to my view. The most daunting challenge is that maximizing expected value may not even be possible. Here's the gist of the problem: Suppose you're uncertain whether utilitarianism or Kantianism is true. You're deciding between two actions, one of which utilitarianism recommends, and the other of which Kantianism recommends. To find out which action has the higher expected value, you will have to know how the difference in value between the first and the second, if utilitarianism is correct, compares to the difference between the second and the first, if Kantianism is correct. The problem is that neither theory provides the resources necessary to make such intertheoretic comparisons, and it's not clear where such resources might be found. I call this the Problem of Value Difference Comparisons.

In Chapter 4 and Chapter 5, I'll try my hand at solving it. I'll begin Chapter 4 by discussing three other solutions to the problem – one by Ted Lockhart and two by Jacob Ross – and explaining why they do not succeed. I'll then provide my own solution, which involves doing two things to the various rankings of actions – cardinalizing them, and normalizing them. To cardinalize a ranking is to give it cardinal structure, which it has if it contains information not only about the ordering of actions, but about the ratios of the differences in value between them. To normalize rankings is to "put them on the same scale", such that differences between actions according to one ranking can be compared with differences according to the other rankings. I'll provide several possible methods by which we might cardinalize and normalize, respectively.

Cardinalization in particular is bound to prompt worries. First, what does it even mean to say that the difference in value between two actions stands in some ratio to the difference in value between two other actions? And second, is it really plausible to suppose that we have such cardinal rankings in our heads? Chapter 5 is devoted to answering the first question, and in the process, answering the second as well. It provides a theory of cardinalization that should fit comfortably with a variety of normative views.

That will mark the conclusion of my defense of expected value maximization.

Expected value maximization is not, however, the whole story of rational action under normative uncertainty. For I may be uncertain not only about the values of actions – about what's better or worse than what – but also about their statuses – about what's required, forbidden, supererogatory, or permitted. In Chapter 6, I'll use my view about uncertainty regarding value as the foundation upon which we might develop views about uncertainty regarding statuses. I'll begin the chapter by categorizing the different accounts of what determines the statuses of actions. On one account, an action's status is determined by its value plus some feature that is not an evaluative feature of the action, like whether an agent would be blameworthy for performing or failing to perform the action; on another account, an action's status is a function of the degrees of the different kinds of value it has; on a third account, an action's status is a complex function of its overall value. I'll then show how my theory of rationality under uncertainty about value leads us to theories of rationality under uncertainty about statuses that correspond to each of these accounts of statuses, respectively.

In Chapter 7, I'll address a problem up to which everyone providing a theory of rationality under normative uncertainty, whatever her view thereof, must face: If I can be uncertain about first-order normative hypotheses, then presumably I can also be uncertain about what it's rational to do under uncertainty about first-order hypotheses. And if I can be uncertain about what it's rational to do under uncertainty about first-order hypotheses, then presumably I can also be uncertain about what it's rational to do under uncertainty about what it's rational to do under uncertainty about first-order hypotheses. Abstracting from our cognitive limitations, this uncertainty might continue ad infinitum. Several philosophers have suggested that this possibility makes trouble for the very project of giving a theory of rationality under normative uncertainty. After explaining the two specific problems that this possibility gives rise to, I'll offer a solution to one of these problems, and a dissolution of the other.

I'll conclude the dissertation by discussing some issues related to normative uncertainty that merit further consideration. In the course of so doing, I'll provide an informal summary of the dissertation's main arguments.

CHAPTER ONE: UNCERTAINTY, TAXONOMY, AND HISTORY

This chapter is the setup chapter, and as such is devoted to three preliminary tasks. The first task is to explain what normative uncertainty is and to convince you that it exists. This will require, among other things, diffusing philosophical arguments from on high for the skeptical claim that there's no such thing as normative uncertainty. The second task is to provide you, the reader, with the conceptual distinctions necessary to understand the argumentative structure of the dissertation, along with an outline of that structure. The final task is to review the history of work on normative uncertainty, and explain, in light of that history, how this dissertation contributes to it.

Task #1

Let's say that an agent is normatively uncertain just in case a) her degrees of belief (or "credences", or "subjective probabilities") are divided between at least two mutually exclusive normative propositions, and b) this division in her degrees of belief is not entirely due to non-normative uncertainty. Consider a commander-in-chief deciding whether to go to war. If she has some credence in the proposition the balance of reasons favors going to war rather than not going to war and some credence in the proposition the balance of reasons favors not going to war rather than going to war, and this is not fully explained by her uncertainty regarding the non-normative facts, then the commander-in-chief is normatively uncertain.[2]

[2] These two conditions can be summarized and stated formally as follows: For at least two normative propositions, Norm_1 and Norm_2, and at least one complete non-normative description of the world, Comp, $p(\text{Norm}_1|\text{Comp}) > 0$, $p(\text{Norm}_2|\text{Comp}) > 0$,

Someone may be normatively uncertain, and indeed, uncertain generally, even if his credences are imprecise, or as they're sometimes called, “mushy”. I may be uncertain, for example, whether prioritarianism or egalitarianism is correct without having a credence of exactly .215, or .732, or .416 in prioritarianism. I might instead have a credence of “around .6” in prioritarianism, or a credence that I might might express by saying that prioritarianism is “pretty damned likely to be true”. At the limits, I might have a credence that I express by saying “it's more likely than not”, or “there's some chance or other...” that prioritarianism is right. There are different ways of formally representing imprecise credences – one involving credence intervals, and another involving families of credence assignments.[3] Because I don't have anything new to say about reasoning with imprecise credences, and because such credences don't seem to suggest any theoretical issues when their contents are normative that don't arise when their contents are non-normative, I'm going to ignore them in this dissertation. Instead, to ease exposition, I'll speak as though credences are precise. I raise the matter here simply to warn you off of the tempting but mistaken thought that normative uncertainty, as I'm understanding it, requires precise credences.

Now, this whole project will strike you as rather pointless if you deny, on philosophical or other grounds, that there's such a thing as normative uncertainty. So let me spend a little while showing why you'd be wrong to think that. My positive case will consist of what I regard as clear examples of normative uncertainty. I'll then try to diffuse

p(Comp) > 0, and $p(Norm_1) + p(Norm_2) = p(Norm_1 \text{ OR } Norm_2)$, where p(q) is the subjective probability of q.

[3] See Levi (1974) and (1982) and Gärdenfors and Sahlin (1982) for discussions of the available options.

skeptical arguments against the existence of normative uncertainty.

I began the introduction with some situations in which many people might be normatively uncertain. People are uncertain about the permissibility of abortion, the justice of war, and which, if any, comprehensive moral theory is correct. But we can provide even more decisive examples than these. There are two “recipes” for producing such examples:

First, there are cases of one-way known entailment. I know that preference utilitarianism entails utilitarianism, but not the other way around; that utilitarianism entails distribution-insensitive consequentialism, but not the other way around; that distribution-insensitive consequentialism entails consequentialism, but not the other way around. Given one-way known entailment, there are two possibilities for someone's doxastic states regarding normative propositions. One is that, for any normative proposition, P1, that is known to be entailed by each of P2, P3, P4...Pn, which are mutually exclusive, my credence in one of P2...Pn is equal to my credence in P1, and my credence in each of the others of P2...Pn is zero. The other is that my total credence in all of P2...Pn equals my credence in P1, but my credence in two or more of P2...Pn is greater than zero. If the second possibility ever obtains, then there is normative uncertainty, so the foe of normative uncertainty must insist that the first possibility always obtains (and must insist further that my credence in P1 is 1). Otherwise, there will be normative uncertainty at some level of specificity. Consider the example at the beginning of this paragraph. A thinker is sure that utilitarianism is right, and he learns that there are all these ways of spelling out what “utility” consists in – preference satisfaction, satisfaction of the preferences one would have if one were idealized in this way or that, this qualitative feel, that qualitative feel, this measure of objective goods, that measure of objective goods, etc. The foe of normative uncertainty must claim that, upon realizing all of these varieties of utilitarianism, the thinker must immediately fix on one of them and be as certain of it as he was in utilitarianism in general. This does not seem psychologically realistic. Recognition of new possibilities within normative theory almost always gives rise to uncertainty among those possibilities.

Second, there are cases involving gradable normatively relevant features. Suppose that a thinker's credence is 1 that it's okay to kill 1 to save a trillion, but that her credence is zero that it's okay to kill 1 to save 1. But what about killing 1 to save 2? 3? 4? 1 million? He who denies that there is normative uncertainty must claim that there is some number, N, such that the thinker's credence is zero that it's okay to kill 1 to save N, but 1 that it's okay to kill 1 to save N+1. But again, this seems psychologically implausible as it regards most of us. It's much more likely that our thinker's credence is intermediately valued for at least some numbers of people saved.

Those who deny that there's normative uncertainty aren't going to get very far, then, by denying the existence of putative examples thereof. Rather, they'll have to undercut the force of these examples on more general grounds. I'll consider three types of normative uncertainty-skeptics: First, the skeptic who says there's no normative uncertainty, since there's no uncertainty of any sort; second, the skeptic who says that, even if there is uncertainty in general, there is a conceptual bar against normative uncertainty; and third, the skeptic who says that, because non-cognitivism is the correct theory of normative judgment, there can't be normative uncertainty or anything like it. Now, again, I described these skeptics as presenting arguments from “on high” that there's no normative uncertainty. None of them is presenting empirical research that suggests it doesn't exist. (And for what it's worth, if the examples I've produced are convincing, and if the skeptics' theoretical arguments are erroneous, I rather doubt that empirical research could show that people are not normatively uncertain; it could at most tell us how normative uncertainty is instantiated.) That doesn't mean that they aren't arguing from “different heights”, as it were, or that my counter-arguments aren't, in turn, pitched at different levels to each respectively. For example, I'll argue against the first skeptic by saying we can't explain certain bits of behavior without appealing to uncertainty, but I'll argue against the last skeptic by invoking the need for non-cognitivism to solve the “Frege-Geach Problem” – very different kinds of claims. But just as it's crucial to see that both the “birther” and the compositional nihilist are skeptics about Barack Obama's birth certificate, it's crucial to see that all of the characters I'll be addressing are skeptics about whether people are ever normatively uncertain. So let's have a look at each character in turn:

One might say that there's no normative uncertainty just because there's no uncertainty, period. There is just belief, disbelief, and maybe suspension of judgment. To really say something satisfactory against this general position would require some deep work in philosophical psychology that I'm not going to do here. All the same, a cursory examination suggests that this is an untenable position. I'll bet you a trillion dollars to your one that the gravitational constant won't change in the next minute; I'll bet you at most five dollars to your one that Roger Federer will win another Grand Slam tournament. It is, prima facie, very difficult to explain this set of behaviors without appealing to degrees of belief.

"But wait," one might reply, "Can't we explain these same behaviors with non-personal probabilities, rather than credences/subjective probabilities?" Instead of saying that my degree of belief regarding the gravitational constant must be higher than my degree of belief regarding Federer's winning another Grand Slam, we can say instead that I have a full belief that there's a very, very, very, high probability of the constant's staying the same, and a full belief that there's a merely very high probability of Federer's winning another Grand Slam.

The problem with this tack is that some people don't have all of the beliefs-about-probabilities that would be required to explain the behavior in question, and so we're left with an explanatory gap that it's most natural to fill with credences. Some people don't believe in non-personal probabilities other than zero and 1. Bruno DeFinetti is the writer most often associated with this position.[4] A more common view is that there are non-personal probabilities other than zero and 1, but that they don't apply to certain types of hypotheses. Many will accept such probabilities regarding the future ("It's probable that Barack Obama will win a second term"), but not about the past ("It's probable that George Washington won a second term"); or about concreta ("The cat is on the mat"), but not about abstracta ("It's unlikely that 94 times 95 is 465"); or about propositions thought to be contingent ("There's a good chance that are more cattle than bison in North America"), but not about those thought to be necessary ("The chance that heat is mean molecular motion is .95", "It's likely that one has reason not to harm others"). But imagine someone who denies non-personal probabilities other than zero and 1 with

[4] See DeFinetti (1937).

regard to the past. Such a person may still bet more on Washington's having won two terms than on Zachary Taylor's having done so, and more on Taylor's having done so than on Jimmy Carter's having done so. This seems inexplicable without degrees of belief.

This argument does not go so far as to establish that there's normative uncertainty. But if indeed there's no normative uncertainty, this can't simply be because there's no uncertainty at all.

Another anti-normative uncertainty position is that, while there may be uncertainty regarding some matters, there is no normative uncertainty because we are always certain regarding the truth/falsity of normative propositions. Now again, in light of examples like the ones I've introduced, this view seems to defy common sense. What philosophical arguments might serve, then, to dethrone common sense? The only one I can see is that being certain of normative propositions is constitutive of possessing normative concepts, and so it's impossible to be uncertain regarding those propositions. By way of analogy, suppose you claim to be certain that Bob is a bachelor, but uncertain whether he's unmarried. It's reasonable to reply that your claim can't be right. To have that set of mental states, you must employ both the constituent concepts UNMARRIED and BACHELOR. But it's a condition on having the concept BACHELOR that you are certain that someone is unmarried when you're certain that he's a bachelor. So you can't really be thinking what you say you're thinking. (Maybe you're using some other concept BACHELOR*, and the only belief constitutive of possessing BACHELOR* is that you must be certain that someone is slovenly if you're certain that he's a bachelor*.) Similarly, perhaps it's a condition on having the concept REASON that you're certain that one has reason not to kill innocent adults. It would be impossible in that case to be uncertain whether there are reasons not to kill innocent adults.

This position seems to undergird at least some denials of normative uncertainty,[5] but it's extraordinarily weak. First, even if having certain doxastic attitudes involving non-normative concepts is constitutive of possessing them, I can see no reason why these attitudes need to be certainty. Why not instead say that the condition on possessing the concept REASON is that you have a relatively high credence that one has reason not to kill innocent adults? Second, if my opponent's aim is really to cut my project off at the root, he must show not only that there are some normative propositions about which one must be certain, but that all normative propositions are like that. But it's just nuts to think, for example, that your possession of the concept WRONG depends on your being certain that abortion is wrong.[6] Sometimes, if not all the time, you can go either way regarding a normative question without jeopardizing your ability to think about the question in the first place.

There is one more argument against normative uncertainty. It moves from non-cognitivism – the claim that normative judgments are not belief states – to the conclusion that normative judgments therefore can't come in degrees. Now, non-cognitivism is a controversial view. If it's indeed false, then of course any argument against normative uncertainty that takes it as a premise is going to be unsound. I think it's more interesting, though, to suppose that non-cognitivism is true and see if it can be reconciled with

[5] See, e.g., Foot (1958), Adams (1995).

[6] This is just enough discussion of normative concepts, I think, to give you a feel for my opponent's views and how we might deal with them. I have more to say about the matter in my “Apriority, Analyticity, and Normativity”, which is part of a much broader project about the metaphysics, semantics, and epistemology of the normative.

something like normative uncertainty. I say "something like" normative uncertainty because it's definitionally impossible for there to be true normative uncertainty if non-cognitivism is true. Uncertainty is a matter of degrees of belief, and non-cognitivism says that normative judgments aren't belief states. But perhaps there can be a state of divided normative judgment that behaves just as normative uncertainty would have if there were normative uncertainty.

This may strike you as very easy to show. After all, beliefs are certainly not the only attitudes that come in degrees. I have a desire of some strength to eat a sandwich; I have a desire of greater strength to visit Budapest. So maybe whatever can be said about normative uncertainty can be cross-applied to divided normative judgments even if these judgments are desires. Then the non-cognitivist can simply read this dissertation and replaces all instances of "credences" with "degrees of desire".

However, Michael Smith has shown that such a simplistic substitution of degrees of desire for degrees of belief won't work.[7] Smith's claim is that non-cognitive states do not have enough structural features to map onto the features of normative judgment that are relevant in explaining action. One feature of normative judgment is what Smith calls "Certitude".[8] Just as I can be more certain that humans and apes share a common ancestor than the Copenhagen interpretation of quantum mechanics is correct, I can be more certain that gratuitous torture is wrong than that lying on one's resume is. Another feature of normative judgment is not a feature of the attitude as such, but rather a feature of the world as represented in the attitude's content. It is what Smith calls "Importance". I can judge that I have slightly more reason to vote for Nicolas Sarkozy than for Ségolène

[7] Smith (2002), p. 345.
[8] Ibid., p. 346-347.

Royal; I can also judge that I have significantly more reason to vote for Sarkozy than for Jean-Marie Le Pen.

Both of these features are relevant in explaining action, at least for the practically rational agent. The more certain I am that I have reason to do something, the more I'll be motivated to do it. And the more important I think doing something is, the more I'll be motivated to do it. Obviously, cognitivism can accommodate both of these action-explaining features. "Certitude" is just degree of belief; "Importance" is just strength of reasons represented in belief. But Smith alleges that the non-cognitivist cannot account for both of these features. Consider this toy non-cognitivist theory: to judge that I have reason to do A is to desire to do A. All we have is the motivational force of the desire, which may correspond to either Certitude or Importance, but not both.[9]

There is a good explanation for why non-cognitivism faces this problem. The evaluative "oomph" of normative judgment that cognitivism assigns to the content of a mental state – belief – non-cognitivism assigns to the mental state itself. And so the Importance that the cognitivist can account for in terms of gradable properties represented in the content of a judgment, the non-cognitivist can account for only in terms of the judgment itself's being gradable. But the gradability of the judgment was supposed to correspond to Certitude, not Importance, as indeed it does on the cognitivist picture. So cognitivism is a more satisfactory view than non-cognitivism when it comes to representing gradable judgments about gradable properties.

It's worth noting that Smith himself does not mean to argue against the existence of normative uncertainty on the grounds that the non-cognitivist can present us with

[9] Ibid., p. 354-355.

nothing that plays its role. Quite the opposite: he’s arguing against non-cognitivism on the grounds that it can’t accommodate anything like normative uncertainty. But as they say, one person's *modus ponens* is another's *modus tollens*; a dyed-in-the-wool non-cognitivist might wield Smith's arguments against normative uncertainty. I want to take away the non-cognitivist's ability to do this by presenting a kind of non-cognitivism that is compatible with something like normative uncertainty.

An opponent might say, “That doesn't really diffuse the threat very much. Just because there's *one kind* of non-cognitivism that's amenable to your picture doesn't mean that all, or most, or even any other kinds of non-cognitivism are.”

The opponent is correct that there are other sorts of non-cognitivism out there that don't have enough structure to meet Smith's challenge. The simple “normative judgments as desires” theory was one of them. However, I deny that these represent a threat that needs to be diffused. They represent a threat only insofar as they are otherwise plausible. But theories without the requisite structure to survive Smith's attack are not otherwise plausible. For the very structure required for this end is also, I shall argue, required to overcome the “Frege-Geach Problem” and its close cousins, which arise when the non-cognitivist tries to give us a theory of normative language. The Frege-Geach Problem is of such importance that any account of normative thought that does not permit a solution to it must be considered untenable.

Let's first see how this problem, which is in the first place a problem about normative *language*, is also a problem for non-cognitivism, which is, after all, a theory of normative *thought*. Consider what the the non-cognitivist might say about normative language. She might say that, because normative sentences are used to express non-cognitive states, they are meaningless.[10] But this is an unappealing option because the meaning of normative terms is essential to explaining so much about linguistic behavior involving them – for example, why someone who says “Murder is wrong” and someone who says “Murder is not wrong” count as disagreeing, and why it's a mistake to assert “Murder is wrong”, and “If murder is wrong, then abortion is wrong”, but to also assert “Abortion is not wrong.” She might say, on the other hand, that normative sentences stand for propositions. This is, I think, the correct thing to say, but it doesn't fit well with non-cognitivism. For the natural picture of language and thought is that, in uttering a sentence, I'm expressing a mental state whose content is the proposition that sentence stands for. But the notion that normative judgments have normative propositional contents can have no truck with non-cognitivism, on which the mark of normative judgments is attitude-type, not content.

Contemporary non-cognitivists tend to adopt another semantic theory – expressivism. This is the view that the semantic features of normative sentences are a function of the mental states they're used to express, not of the propositions they stand for or of the constituents of those propositions.[11] It's a view that divorces the explanation of normative language from “the world”, and aims to account for it solely in terms of “the mind”. (This does not mean, of course, that the features that determine the truth/falsity of normative claims are all mental features. Keep that in mind for when we get to about page 27 of this chapter.)

This is where the Frege-Geach Problem rears its head. The problem has been

[10] This was the position of the earliest non-cognitivists. See Ayer (1952), Chapter 6.

[11] The most lucid comprehensive treatments of expressivism are, in my view, Gibbard (2003) and Schroeder (2008).

stated many different ways in many different places, but I take its essence to be this: Without invoking specifically normative contents of mental states, there just aren't enough differences among such states to systematically explain the semantic differences among normative sentences, which in turn are supposed to explain which arguments are good, which conflicts count as disagreements, and so forth. In other words, the semantics of normative language is very complex; normative thought on the non-cognitivist picture is much less so, and so good luck accounting for the former in terms of the latter.

It's worth pointing out that to criticize a non-cognitivist theory for failing to solve the Frege-Geach Problem is to presuppose a certain methodology. Call it the “semantics-first” methodology: we start with what we take to be semantic facts, and work backwards to conclusions about the features that normative judgments must have in order to explain these semantic facts. This may be contrasted with the “psychology-first” methodology, on which we start from what we take to be facts about the features of normative judgment, and then work forwards to conclusions about what normative semantics must look like. In arguing that there's something like normative uncertainty on all plausible versions of non-cognitivism, and setting as a criterion of plausibility the ability to solve the Frege-Geach Problem, I'm taking sides with the semantics-first methodology.

The semantics-first approach will look more congenial the fewer non-semantic constraints we place on what may count as a normative judgment. For example, if we say that an utterance may express a judgment only if that utterance was immediately caused by that judgment, then it may turn out that the judgments that cause normative utterances are non-cognitive states that can't support our intuitions about normative semantics. This may be added to the other reasons we have to reject this simple causal account of expression, but you get the point. At any rate, while I think the the semantics-first approach is right, I'm not going to defend it here. Philosophical arguments for claims about psychology have to start somewhere, and semantics doesn't strike me as an unusually bad place to start.

Let's get into the argument by looking at a specific case.[12] We will want to explain why “Murder is wrong” is inconsistent with “Murder is not wrong”. The cognitivist can explain this in terms of “murder”, “is”, and “wrong” standing for the same propositional constituents in the former as they do in the latter, and of course, the semantic value of “not”. The non-cognitivist, however, must explain this inconsistency in terms of the attitudes these sentences express. A toy expressivist semantics for these sentences might say that “Murder is wrong” expresses disapproval of murdering, and that “Murder is not wrong” expresses disapproval of not murdering.

There are two problems with this toy theory. The first is that, if disapproval of murdering is expressed by “Murder is wrong”, then disapproval of not murdering should be expressed by “Not murdering is wrong”, rather than “Murdering is not wrong”. (These last two sentences have different meanings of course. If not murdering is wrong, than murdering is obligatory, but if murdering is not wrong, then murdering is merely permitted.) This naturally raises a second problem: How do we explain the semantic value of “Murdering is not wrong”?

Perhaps we can simply say that “Murdering is not wrong” is expressing an attitude towards murder that it's inconsistent to hold along with the attitude expressed by

[12] This bit is borrowed almost wholesale from Schroeder (2008), p. 39-56.

“Murder is wrong.” This gives us the theory that “Murder is wrong” expresses disapproval of murdering and “Murder is not wrong” expresses (let's call it) “tolerance” of murdering.

The problem with this approach is that is assumes precisely what needs to be explained – namely, that these two attitudes really are inconsistent. The cognitivist can say that the attitude expressed by the former is inconsistent with the attitude expressed by the latter because they’re the very same attitude – belief – toward logically inconsistent contents. The expressivist, though, must explain everything in terms of the kind of state. She must say that that there are just these two states, “disapproval” and “tolerance” as we’re calling them here, that, while sui generis and not interdefinable, are nonetheless not okay to hold together. So yes, while the expressivist on this approach can solve the problem as presented by throwing in an additional attitude, this does no explanatory work.

You might think, “Oh, that's only two attitudes – disapproval and tolerance. Not worth getting in a huff about!” But that's a mistake. The expressivist needs more than two attitudes. She will need to explain the semantic value of every logical form by appealing to a different attitude. One for “If murder is wrong, then abortion is wrong”. Another for “Murder is not not wrong”. Another for “All murders are wrong”. Another for “If all murders are wrong, then saving someone's life is not wrong”. This is not what we demanded earlier, which was a systematic explanation of the semantics of normative sentences.

Mark Schroeder quite rightly says that there's only one way out for the non-cognitivist: “add structure”.[13] The non-cognitivist can't countenance normative propositional contents of beliefs, but she can say that normative sentences express mental states with features that play the roles of state and content in the explanation of normative semantics. Here is Schroeder's suggestion for how such theory might look:

> “We will have to say that it is some kind of very general non-cognitive attitude. Let’s give it a name, and call it 'being for'. The solution is to say, just as all descriptive predicates correspond to belief plus some property that is contributed by the predicate, that all normative predicates correspond to being for plus some relation that is contributed by the predicate. For each predicate, F, there is a relation, RF, so that ‘F(a)’ expresses FOR(bearing RF to a). So, for example, to borrow a proposal from Gibbard (1990), we might say that ‘wrong’ corresponds to being for blaming for , so that ‘murder is wrong’ expresses FOR(blaming for murder). Similarly, we might say that ‘better than’ corresponds to being for preferring, so that ‘a is better than b’ expresses FOR(preferring a to b).”[14]

So instead of just an attitude towards, say, murder, we have an attitude towards bearing some relation to murder. “Murder is not wrong”, then, expresses FOR(not blaming for murdering), and “Not murdering is wrong” (read: “Murdering is obligatory”) expresses FOR(blaming for not murdering). The role that content played for the cognitivist, this extra relation can play for the non-cognitivist. And indeed, it must play this role, unless she going to leave the semantic properties of normative sentences unexplained.

Now let's get back to normative uncertainty. This is ammunition for the non-cognitivist to use against Smith, for she can exploit the very same extra structure that she

[13] Ibid., p. 61.

[14] Ibid., p. 58. Schroeder is not concerned to defend this theory in particular. Maybe some relation other than blaming (for “wrong”) or preferring (for “better than”) should be slotted in instead. He's simply trying to get at the nut of the Frege-Geach problem, and provide a schema for a non-cognitivist psychology/expressivist semantics that gets around it. So it's important not to get too hung up on the details.

needed to solve the Frege-Geach Problem in order to answer Smith's challenge.[15] To stick with Schroeder's scheme, we might want to say that degrees of Being For are for the non-cognitivist what degrees of belief are for the cognitivist. Mutatis mutandis for degrees of blame and strength of reasons, respectively. For example, being very sure that interrupting someone is only slightly wrong would be replaced with being very "For" slightly blaming someone for interrupting.

At this point, it's natural to object that Being For differs from credence in that the latter satisfies the Normalization axiom, while the former does not. That is to say, the highest degree of belief that I might have in a proposition is 1, but no matter how For anything I am, I could always be more For it. The measure of Being For, then, is unbounded, just as the measure of length or duration is. Moreover, even if answering the Frege-Geach challenge requires the non-cognitivist to posit a structural element that might play the degree of belief role, and another that might play the degree of value role, it does not seem to require him to say that the element that plays the degree of belief role must obey Normalization.[16]

I should say: even if this difference is genuine, it's not clear how relevant it is for

[15] What's interesting, too, is that certain ways of adding structure that fail to solve the Frege-Geach Problem also fail to solve Smith's problem. Simon Blackburn's (1984) approach uses higher-order versions of the attitudes that are also supposed to serve as moral judgments to represent different logical constructions. As van Roojen (1996) shows, however, this is unsatisfactory. There is a difference in kind between judgments about morality and judgments about how attitudes may or must be combined. Similarly, as Smith argues, there is difference between certainty and importance that the higher-order attitudes approach fails to capture. Smith puts the objection in terms of it's being "arbitrary" which order of attitude is taken to represent Certitude, and which is taken to represent Importance. p. 356.

[16] Thanks to Ruth Chang for emphasizing this to me.

the purposes of my project. I see no reason to think that the rationality of expected value maximization or any other response to normative uncertainty depends on credences' satisfying Normalization. The theory of rationality applicable when my credence in H1 is .2 and my credence in H2 is .8 seems just as applicable if my credence in H1 were 20 and my credence in H2 80.

I grant that it's a bit difficult to evaluate this response, seeing as it depends on conceiving of credences as exceeding 1 – no easy mental feat. However, I think there is a better response available. Whatever the nature of Being For, there's a way to assign numerical values to degrees of it such that these never exceed 1; and as I'll show in a moment, assigning values in this way is, in fact, required in order to solve a problem very much like the Frege-Geach Problem. Let's begin by considering how the Normalization axiom is often formulated: $P(\Omega) = 1$, where Ω is a “universal set” whose members are all possible events. In other words, it's certain that something will happen. Nor can it be more likely that any particular event will happen than that some event or other will happen. This is due to the Additivity axiom, which states that the probability of a union of mutually exclusive elements is equal to the sum of their individual probabilities. So the probability of some event or other happening is just the sum of the individual probabilities of the mutually exclusive events happening, and positive value cannot be greater than the sum of itself and another positive value.

We might formulate a Normalization axiom for Being For along similar lines: $\text{FOR}(\Omega) = 1$, where Ω is a “universal set” whose members are all possible relations an agent might bear to an action. In other words, insofar as I am For bearing some relation or other to some action or other – either blaming for everything, or not blaming for anything, or blaming for cruel acts only, and so on, and so on – I must be For it to degree 1. There can also be an Additivity axiom for Being For, which would yield the result that I can be no more For bearing a particular relation to a particular action than I would be For bearing some relation or other to some act or other; therefore, the greatest degree I can be For bearing any relation to any action is 1.

Because of Normalization of credence, increases in the probabilities of events must coincide with decreases in the probabilities of events with which they are mutually exclusive. Evidence that raises the probability that it will rain simultaneously decreases the probability that it will not rain. Similarly, because of the Normalization of Being For, the degree to which I can be For bearing some relation to an action can increase only if the degree to which I am For bearing some logically incompatible relation to an action decreases. For example, if I become more For blaming for stealing, I must become less For not blaming for stealing. Two metaphors may be helpful: Given the Normalization axiom, we shouldn't think of either belief- or For-revision as adding more and more sand onto spots in the possibility space; we should think of it as shifting a given amount of sand around that space. Nor should we think of For-revision as altering the total magnitude of the vectors that determine the direction of motivation; we should think of it as altering only that direction – as pushing the arrow around the compass, we might say.

"That might be how Being For works," an objector could reply, "But it doesn't have to work that way. It could be that it's not additive; consequently, I could be more For blaming for stealing, say, than I am for For bearing some relation or other to some act or other. I mean, why not? Also, it could be that the non-Normalized way of assigning values to Being For is correct, such that I could have degrees of Being For that don't sum to 1. Again, why not?"

I grant that these are possibilities. The Additivity and Normalization of Being For are perhaps not as intuitively obvious as the Additivity and Normalization of credence are. So Being For doesn't have to work the way I say. But look: whatever non-cognitive state normative judgment is doesn't <u>have to</u> work in such a way that the non-cognitivist can solve the Frege-Geach Problem, either. Insofar as we demand that the non-cognitivist solve this problem, though, we're demanding that he give us a theory of normative judgment on which the inferences that we intuitively think of as good come out as such. In demanding that Being For obey Normalization and Additivity, I'm simply making the same sort of demand, but with an eye towards probabilistic inference rather than deductive inference. If Being For doesn't obey Additivity, then the following could be a perfectly good inference:

> Premise: "There's a .4 probability that stealing is wrong."
>
> Conclusion: "There's a .39 probability that either stealing is wrong or dancing is wrong."

The premise expresses Being For to degree .4 blaming for stealing; the conclusion expresses Being For to degree .39 either blaming for stealing or blaming for dancing.

Of course, this is not a good inference, and I contend that any non-cognitivist theory that can't explain this is implausible. So Being For has to obey Additivity if a theory built around it is to be plausible, and once again, our concern is only whether there can be something like normative uncertainty on a plausible non-cognitivist theory, not whether there can be such a thing on any non-cognitivist theory whatsoever.

Similarly, if Being For doesn't obey Normalization, then I could coherently say the following without a change of mind:

> “There's a .8 probability that stealing is wrong,” and
>
> “There's a .8 probability that stealing is not wrong.”

The first claim expresses Being For to degree .8 blaming for stealing; the second claim expresses Being For to degree .8 not blaming for stealing.

If we're just looking at the attitude Being For, in isolation from the semantics it supports, it might seem perfectly okay to have these Beings For together. But what drives both the Frege-Geach challenge and the present argument is that we have, prior to any view about the psychology of normative judgment, views about notions like coherence and good inference that our psychological theories had better not falsify. And so once we see that a psychological view about normative judgment commits us to regarding as acceptable sets of claims that clearly are not acceptable, we must reject this psychological view. Only if Being For obeys Normalization can the non-cognitivist theory built on it count as otherwise plausible.

The bottom line, then, is that the non-cognitivist attitude must have an element that we can say corresponds to degree of belief, and a different element that we can say corresponds to degree of value represented in the belief, if non-cognitivism to solve the Frege-Geach Problem. Schroeder helps us to see this. But on top of this, accommodating other basic intuitions about inference and coherence requires that the element that corresponds to degree of belief must obey the Normalization and Additivity axioms. It must have the two aspects of structure to explain, among other things, non-probabilistic inference; the aspect that corresponds to degree of belief must obey the axioms if it's to explain probabilistic inference.

As a coda, I should mention that the non-cognitivist would probably also want to replace blaming with some other relation, for two reasons. First of all, there are differences in strength of reasons that correspond to differences in praiseworthiness but not blameworthiness. For example, I am blameworthy neither for eating a sandwich nor for saving a drowning child, but I am more praiseworthy for the latter act than for the former. Secondly, there are differences in strength of reasons that don't seem to correspond either to differences in praiseworthiness or blameworthiness. This is the case when the reasons in question are prudential. I have stronger reasons to attend a superior Ph.D. program than I have to attend an inferior one, but neither choice is a fitting object of praise or blame. But it's not important right now to specify exactly the relation(s) that might serve as the object(s) of Being For (or whatever you'd like to call the attitude). All that matters for my defense of normative uncertainty is that only those non-cognitivist theories on which the relevant attitudes can mimic states of normative uncertainty can do the explanatory work required to solve the Frege-Geach Problem and its relatives.

To summarize all of this: There are what seem to me very clear examples of normative uncertainty, and it is up to the skeptic to undercut the force of these. We imagined three ways in which he might try to do that. In reverse order of presentation, these were: a) arguing that, since normative judgments aren't beliefs, there can't be something like normative uncertainty, b) admitting that there may be uncertainty, but denying that there is normative uncertainty, and c) denying that there is, in the strict sense, uncertainty. None of these strategies was successful, so we should let stand the commonsensical thought that we can be uncertain about the normative.

Task #2

We won't be able to frame the questions of this dissertation without having in hand the right distinctions: the distinction between objective and various sorts of belief-relative normative notions; the distinction between rankings of actions in terms of the strength of reasons that support them, and the deontic statuses of actions; and the distinction between absolute and comparative normative notions. In this section, I'll first explain all of these distinctions, and, particularly in the case of the first, explain why they are important. Then I'll provide a “guided tour” – a “where all your questions are answered” section – of the dissertation that employs, and indeed, is impossible to provide without, the distinctions here.

Objectivity vs. Belief-Relativity

What should you do when you're uncertain about what you should do? There's a way of answering this question flatfootedly: You should do what you should do. If utilitarianism is right, you should act in accordance with utilitarianism, and if it's better to give than to receive, you should give, uncertainty be damned. On one reading of the question, this sort of answer is the only correct one. But I shall be giving a different sort of answer throughout the dissertation, which means I must intend a different reading of the question. The different reading depends on there being different senses of “should”, and more generally, different senses of all of the normative notions.[17] I'll explain all of these senses by talking specifically about the value of an action, by which I simply mean the strength of reasons to perform it.

Start with what I shall call Objective Value. It's tempting to contrast objective normative notions with subjective ones, and say that the mark of the former is that they're mind- or belief-independent – that they depend, rather, on features of the extra-mental world. But this is a mistake. For example, we'll want to allow that an action's being a lie may be an objective reason not to do it, but of course whether an action is a lie depends on the beliefs of the agent. And as a general matter, we will not want to rule out by stipulation that the objective value of an action may depend on any old feature of the world whatsoever, an agent's beliefs included. So let us place no conceptual bar on the features upon which objective value depends, and instead define other sorts of value in terms of objective value.

Belief-Relative Value (BRV), we shall say, depends on the agent's beliefs about the features upon which objective values depend. So suppose you have an objective reason not to touch the stove when the stove is hot. Then you have a belief-relative reason not to touch the stove if you believe the stove is hot. Now consider a case where, we'll assume for argument's sake, your objective reasons depend on your beliefs: you have an objective reason to take a picture of the Eiffel Tower just in case you believe the Eiffel Tower is beautiful. Then you have a belief-relative reason to take the picture if you believe that you believe the Eiffel Tower is beautiful. To put it spatial-metaphorically:

[17] This section benefited substantially from my discussions with Holly Smith.

take all the determinants of objective value, whatever they are, and pack them into a box. It’s beliefs about things in that box that determine BRV. A consequence of defining things in this way is that I can't tell you what BRV depends on until we settle what objective value depends on. But this strikes me as precisely the tack we'd want to take if we're concerned not to label too many substantive positions as conceptual errors.

Now, the cases above were ones in which my beliefs concerned non-normative, rather than normative, facts. But the formulation I've given also allows for value that depends on my beliefs about the normative. After all, if objective value is a function of *anything*, it's a function of itself and other normative features. So my beliefs about objective value/objective reasons/the objective “ought” will also count as beliefs about what I labelled “the determinants of objective value”. For example, whether punching you in the face is something I have objective reason not to do depends upon whether doing so will cause you pain; it also depends, in the most direct way possible, on whether doing what causes others pain is something I have reasons not to do. So if I believe that I have reason not to cause you pain, then I have belief-relative reason not to punch you in the face.

Let's distinguish now between the kinds of belief-relative value. There is a notion of value that is relative to the agent’s *beliefs about* the non-normative determinants of objective value, but relative to the *actual normative determinants of* objective value. Call this *Non-Normative Belief-Relative Value (N-NBRV)*. This is what people are usually talking about when they talk about belief-relative value, or subjective value. But there are other notions of belief-relative value. There's one that’s relative to the agent's *beliefs about* the normative determinants of objective value, but relative to the *actual non-*

normative determinants of objective value. Call this Normative Belief-Relative Value (NBRV). Finally, there is the most belief-relative kind of value of all – relative to the agent's beliefs about both the normative and the non-normative determinants of objective value. Call this Rational Value, and the general type of normativity to which it belongs Rationality.

Now for a distinction within rationality itself. This is the distinction between two ways of assessing rationality – globally and locally. The global rational value of an action depends on all of that agent's mental states. The local rational value of an action depends on only a subset of that agent's mental states. But while it makes sense to speak of global rationality simpliciter, it doesn't make sense to speak of locally rational simpliciter. We first have to specify which subset of the agent's mental states we're talking about. So evaluations of local rationality will always be evaluations of an action's rational value relative to this or that subset of an agent's mental states.[18] Most norms of rationality we talk about are local. We say, for example, that it's irrational for someone who believes P and P → Q to maintain those beliefs and also to form the belief that ~Q. But in saying this, we tend to ignore other of the agent's beliefs – for example, his beliefs that R and R → ~Q, that his evidence strongly supports ~Q over Q, so on. Other norms of rationality – regarding conditionalization on evidence, akrasia, intending the means to one's ends, and not believing P if you believe that the evidence supports ~P, for example – also pertain to small subsets of beliefs in isolation, and as such are local rather than global.

These do not exhaust the types of value that will be relevant. There is another sort of value that's belief-relative in some sense, but that is crucially different from BRV. I call

[18] For a defense of the view that local rationality is explanatorily prior to global rationality, see Kolodny (2005), p. 516, especially fn. 8.

it Epistemic Probability-Relative Value (EPRV), and it will take a bit of extra apparatus to explain.

Start by considering how you'd express a full belief that P – not report that you have it, mind you, but express it. You'd simply say "P". If there's a way to express a full belief, then surely there must be some way to express a partial belief, or credence between zero and 1, that P. But what is it? It can't be by saying, "My credence is X that P." That's how you would report your credence in P, but expression is not reporting. It also can't be by saying, "There's an X objective probability that P." That's how you would express a full belief with objective-probabilistic content, not a degree of belief. Instead, you may express your credence of X that P, I claim, by employing the language of Epistemic Probability (EP), or at least, language that stands in for it.

For example, one might express one's .3 degree of belief that the Lakers will win the NBA Championship and .7 degree of belief that the Cavaliers will win the NBA Championship by saying, "There's a .3 EP that the Lakers will win the Championship, and a .7 EP that the Cavaliers will win the Championship." Of course, that's a very stilted way to talk. Normally we use shorthand – "Eh, it's probably gonna rain tomorrow", or "I guess Kila will just meet us at the Tilt-a-Whirl", or "There's a decent chance Merv will be at the party" – and allow context to make it clear that what we're doing is expressing our credences.

The main difference between expressing full beliefs and expressing credences, then, is that in expressing full beliefs, we use sentences that stand for the propositions that are the contents of those beliefs. Not quite so in expressing credences. The content of my credence of .3 that the Lakers will win the NBA Championship is The Lakers will *win the NBA Championship*, and the sentence that stands for this is “The Lakers will win the NBA Championship”. But the sentence that I use to express this attitude *reflects its degree*: “There's a .3 EP that the Lakers will win the NBA Championship”. This is an odd feature, but I don't see how it can be avoided. What I say to express a credence of .3 must differ from what I say to express a credence of .5, and this can't be the case unless the difference in credence is manifested as a sentential difference. Let me emphasize again, though, that epistemic probabilities are not subjective probabilities; if they were, then expressing a credence would be the same as reporting it, which it's clearly not.

So then what *are* epistemic probabilities? Of particular concern is whether they are features of the world to which we commit ourselves whenever we express our credences. That'd be strange, particularly because we don't commit ourselves to anything of the sort in expressing full belief or certainty. We escape this result by assigning semantic values to EP-statements, in part, *expressivistically*. Rather than give the semantic value of statements of the form, “There’s such-and-such an epistemic probability that...” by appeal to EP-properties in the world, we give their semantic values by the mental states they’re used to express – namely, credences of less than 1. EP-expressivism, then, works just like moral expressivism. I think *some* version of this strategy must be right; otherwise, I don’t see how we could manage to express our uncertainty while at the same time saying no more about the world than we say to express our certainty or full belief.

Epistemic probability statements lack truthmakers traditionally understood, but they’re still capable of being true or false, and their truth/falsity does not depend, as a conceptual matter, only upon features of the agent’s mind. (That would be subjectivism about epistemic probability, which is absurd.) For example, the truth of the statement that there's a decent chance that Merv will be at the party depends on how far the party is from Merv's house, whether Merv's a partyin' kinda guy, and so forth. This is exactly parallel to how moral expressivism works. The moral expressivist says that the semantic values of normative terms are to be explained by the mental states they're used to express, but it is not part of his position that, say, murder is wrong if and only if I disapprove of murder. (That would be subjectivism about morality, which is absurd.) Instead, the truth of the statement that murder is wrong depends on murder's causing pain, murder's violating the autonomy of the victim – that sort of thing. Anyway, that's a sketch of a meta-semantic theory about epistemic probability statements.[19] It's a marriage of a view about how credences may be expressed, with a view about how such expression is related to the semantics of epistemic probability.

Now that that's on the table, we can talk about a kind of value that's relative to epistemic probability, in the same way that BRV may be relative to subjective probability or credence. This is what we'd called "EPRV". Just as the (non-normative) BRV of bringing Merv's favorite beer to the party increases as my credence that Merv will be at the party increases, the EPRV of bringing Merv's favorite beer to the party increases as the EP that Merv will be at the party increases; and this EP increases when, for example, Merv finds out his crush will be at the party, and decreases when, for example, Merv's car breaks down on his way to the party.

There are other sorts of value we might define up. We could have Evidence-

[19] An expressivist meta-semantic treatment of epistemic probability statements is also defended in Yalcin (forthcoming). Yalcin also gives a rather detailed semantics for such statements. I won't do the same here, but I urge those interested in the issue to consult his impressive paper.

Relative Value, which is relative to the evidence concerning the determinants of objective value, or Idealized BRV, which is relative to the credences regarding the determinants of objective value a thinker would have if she were completely theoretically rational. These are interesting notions, and they may indeed play important practical roles, but I won't spend any more time on them. Understanding the structure of the normative uncertainty debate depends only on grasping the other normative notions I've explained so far. Once such an understanding is in place, applying what I've said to questions about these other sorts of value should be simple.

Rankings and Statuses

This brings us to our second distinction: the one between rankings of actions and their deontic statuses. A ranking of an action is an ordering of actions in terms of the strength of reasons to do them. I'm rendering a ranking judgment when I say, “The balance of reasons favors A over B, B over C, and C over D.” Throughout the dissertation, as I did above, I will sometimes use the language of value (and sometimes of an action's being better than/worse than/equal to another) as shorthand for the language of reasons. So I might equally well express the judgment above by saying, “A has more value than B, B more than C, and C more than D,” or “A is better than B, which is better than C, which is better than D.” This may not be how “value” and “better than (etc.)” are always used in the contemporary literature. These are perhaps more often used to express what writers like David Wiggins and Judith Thomson have called evaluative notions, rather than what they've labelled *directive* ones.[20] But I shall be using them to express directive notions here; since the focus of the dissertation will be focused entirely on directive, and not evaluative, notions, this shouldn't prove too confusing.

Rankings can be more or less structured. An *ordinal* ranking of actions contains information about which actions are better than/worse than/equal to other actions, and that's it. It doesn't contain any information about the relative sizes of the differences in value between actions – about whether the gap between, say, murder and theft is greater than the gap between theft and dancing. An *ordinal difference* ranking contains ordinal information about which differences in value between actions are greater than/less than/equal to other such differences, as well as information about which actions are better than/worse than/equal to other actions. Such a ranking might say, first, that murder is worse than theft, which is worse than dancing; and second, that the difference between the first two is greater than the difference between the second two. However, an ordinal difference ranking doesn't contain information about the *ratios* of differences between actions. A *cardinal* ranking, by contrast, does. It might say that murder is worse than theft, which is worse than dancing, and also that the difference between the first two is 3 times the difference between the second two. An *absolute cardinal* ranking contains the most information of all – cardinal information, but also information about the absolute values of actions. Because such information also determines ratios of actions' values (not just the ratios of *differences between* actions' values), it is often called a “ratio” ranking. Such a ranking might say that murder has a disvalue of -1100, theft a disvalue of of -100, and dancing a positive value of 100; these numbers stand in for non-numerical absolute

[20] “Evaluative”/”Directive” comes from Wiggins (1979) and Thomson (2008).

assessments of actions. -1100 is “very bad”, perhaps; 100 is “pretty good”.

Ordinal and cardinal rankings will play major roles in this dissertation – roles that I'll spell out soon. Ordinal difference rankings will not. Given the theory of rationality under uncertainty I espouse – expected objective value maximization – the extra information provided by ordinal difference rankings will rarely be helpful in determining which actions are most rational. This will become apparent once we see what expected objective value maximization amounts to, and how it applies in concrete cases. Nor will absolute cardinal rankings play a major role. This is because absolute, as opposed to comparative, normative features will in general be consigned to the sidelines. I'll explain why at the end of this section.

Sometimes ranking actions is not such a clean and simple affair. First, it's conceivable that some pairs of actions are on a par. Neither is better than the other, nor are they equal; rather, they stand to one another in a fourth positive value relation – parity.[21] Second, it's conceivable that some actions are incomparable with one another: they stand to one another in no positive value relation whatsoever.[22] Third, it may be that some actions are “absolutely” better than others, such that it seems natural to represent their relations by invoking infinite value or disvalue.[23]

In addition to rankings, this dissertation will be concerned with the statuses of actions: required, supererogatory, permitted, suberogatory, and forbidden. There are synonyms for these terms. For example, we might call a required action “obligatory”

[21] See Chang (2001) and (2002).

[22] See, e.g., Raz (1988), Stocker (1990), and Anderson (1995), and several of the papers collected in Chang (1997)

[23] See Jackson and Smith (2006).

instead, or a forbidden action “prohibited”, or a permitted action “okay to do”. There are different views about how statuses relate to rankings. On one view, an action's status is a function of its value and whether one would be blameworthy for doing it or not doing it, provided certain other conditions obtain. So, for instance, it would be obligatory to do A from a field of A, B, C, and D if A had the highest value and one would be blameworthy for not doing A; it would be supererogatory to do A from a field of A, B, C, and D if A had the highest value but one would not be blameworthy for not doing A. There are many other theories besides this one, which I'll discuss in Chapter 6, when I address uncertainty regarding statuses in particular.

That the theory above is a live theory means that it's not a conceptual truth that one is required to do the highest-ranking action in every situation. This is a substantive thesis that must be argued for, presumably only after one settles on an account, of the sort I'll survey in Chapter 6, of what statuses are. But rankings nonetheless constrain statuses and statuses constrain rankings, to some degree. It is a conceptual error, for example, to say that A is better than B, but that one is required to do B and forbidden to do A when they are in the same field. In general, though, pairwise comparisons between actions tell us very little about the statuses of those actions. Insofar as statuses depend on rankings, they depend mostly on actions' rankings vis a vis all of the action in the field. A might be better than B, even though both are forbidden in a field of A, B, C, and D.[24]

My own view is that rankings are conceptually prior to statuses – that we understand statuses in terms of, among other features, rankings, rather than the other way around. You needn't agree with this view, however, to accept the distinction between

[24] Thanks to Martin Lin and Jenny Nado for helping me to get straight on this.

rankings and statuses as I've presented it here. It is possible to say how different concepts are related to one another without relying on or implying a view about priority.

Absolute vs. Comparative

Comparative normative features are, as a conceptual matter, dependent upon the normative features of other actions – typically of other actions possible in the same situation. Absolute features are not. Once again, because we will not want to rule out the possibility that the normative features of an action depend on any old feature of the world whatsoever, it is wrong to say that absolute features are independent of the normative features of other actions. The normative features of other actions are, after all, features of the world. But that this type of dependence obtains is a conceptual truth for comparative normative features, and not a conceptual truth for absolute normative features.

The comparative/absolute distinction cuts across the objective/belief-relative distinction, and, it's important to see, across the ranking/status distinction. There are comparative ranking notions – better than, more reason to, more valuable than; there can be absolute ranking notions – good, strong reason to, valuable. There can be comparative status notions: “It's obligatory to do A rather than B, C, D, etc.”; there can be absolute status notions: “It's obligatory to do A.”

I introduced the previous two sets of distinctions because each will play an important role in the dissertation, as you'll see in the “where your questions will be answered” section that's to come. I introduce this distinction to tell you in advance that it won't play a major role. The entire dissertation will focus on comparative, rather than absolute, features of actions.

One reason I won't discuss such absolute features is that part of me thinks there aren't any. We shouldn't think there are absolute normative features, because such things don't seem to play any role in our practices. Consider absolute ranking features first. When I'm deciding what to do, I (hopefully) opt for what I have most reason to do; I don't avoid doing so because I think that that action is absolutely bad, or because I think, say, the second-best action is absolutely good. However, maybe that's a bit simplistic. For don't I lament cases where all of my options are absolutely bad, or rejoice when all of my options are absolutely good, and aren't lamenting and rejoicing “our practices” as well? But it seems more appropriate to say that I lament cases in which all of my options compare unfavorably with a salient set of options – for example, the options that I'd encounter on a typical day. Mutatis mutandis for rejoicing.

I also want to deny the existence of absolute statuses, but it will be tough to give you a persuasive argument for this denial without saying more about what statuses might be, which I'm not going to do until later. Here's an example, though: Suppose the mark of the forbidden is blameworthiness. Your action has the status of being forbidden only if it's appropriate to blame you for doing it. I can see, then, how there could be comparative statuses on this conception. I can be blameworthy for doing A when the other options are B, C, and D if A is much worse than these. But how can it be appropriate to blame me for doing A, whatever the other available actions are? If I “make the best of a bad situation”, aren't I deserving of, if anything, praise rather than blame? (Of course, it's consistent with this that I may be blameworthy for creating the situation in the first place.) I suggest that there will be similar problems if we adopt other theories of statuses, too.

Of course, it's a very strong claim that there are no such things as absolute normative features. It's important to acknowledge, further, that my reasons for denying their existence stem from views about when lamentation, blame, etc. are appropriate. If you hold opposing views – for example, if you believe it's appropriate to blame someone for doing the best it's possible to do – then you shouldn't yet be persuaded that there are no such things as absolute normative features.

However, I think I'm justified in ignoring uncertainty about absolute features, even if there are such things, on the grounds that my concern in this dissertation is with decision-making under uncertainty, not with lamentation or blame. Even if absolute features of actions play a role in determining lament-worthiness or blameworthiness in a way that comparative features do not, they do not play such a unique role in guiding action. In determining action for rational beings, what matters is how actions compare normatively to each other, not how they fare normatively in some absolute sense.

The Plan of the Dissertation

Now that we have the crucial distinctions on the table, I can tell you in those terms how the dissertation will progress.

We can state the question of this dissertation most broadly as, "Given an agent's credence distribution over propositions about the objective normative features of an action, what are the local rational normative features of that action?" So we'll be taking credences regarding objective normative features as inputs, and spitting out verdicts about rational features as outputs.

On both the input and output side, the focus will be on rankings. Chapters 2, 3, 4, and 5 will constitute my attempt to answer the question, “Given an agent's credence distribution over propositions about the objective rankings of actions, what are the local rational rankings of those actions?” Credences in objective rankings on the input side, then, and rational rankings on the output side.

In Chapter 6, I will shift the focus to statuses, and at least sketch an answer to the question, “Given an agent's credence distribution over propositions about the objective statuses of actions, what are the local rational statuses of those actions?” So credences regarding objective statuses as the inputs, and rational statuses as the outputs. There is a very closely related question that I'll also address in Chapter 6: “Given an agent's credence distribution over propositions about the determinants of objective statuses – objective rankings, blameworthiness, or whatever – what are the local rational statuses of actions?” So credences regarding the determinants of objective statuses as inputs, and rational statuses as outputs. It won't be until that chapter that we'll develop the conceptual resources necessary to pry these two questions apart, so if the contrast doesn't make sense right now, stay tuned.

That's it for the very general normative features – rankings and statuses. But I'll also devote some attention to cases in which rankings go awry. Later in Chapter 2, I'll see what sorts of outputs we get if the agent has some credence in the parity and/or incomparability of actions. Midway through Chapter 3, I'll consider agents who have some credence in absolutist normative theories, and see what ends up being rational for them.

EP-relative normative features won't come into play until the very end of the dissertation – Chapter 7 – when they'll show up as one of the candidate types of norms by which we might guide our actions. Also in that chapter, we'll lean heavily on the notion of local rationality, as we struggle with the problems raised by agents who are uncertain about the rules of rationality under normative uncertainty.

Task #3

I'll be discussing others' work on normative uncertainty intermittently throughout the dissertation, but since we've just now reviewed the structure of the thing, it'd be worthwhile to see how other general treatments of the issue map onto this structure. It's a topic with a gappy history. There's been a decent amount of work on normative uncertainty in the last decade or so, but before that, the issue had only been seriously addressed by Catholic moral theologians, mostly in the late Medieval to Early Modern period.

Contemporary Work on Normative Uncertainty

Some of the contemporary work has been focused solely on very specific questions related to normative uncertainty. In applied ethics, Graham Oddie (1995) has written on moral uncertainty and human embryo experimentation, Dan Moller (ms) has written on moral uncertainty and abortion, and Alex Guerrero (2007) has written on moral uncertainty and killing in general. John Broome has recently begun considering the implications of normative uncertainty for the ethics of climate change.

There have been two comprehensive treatments of normative uncertainty in recent years: Ted Lockhart's Normative Uncertainty and its Consequences and Jacob Ross's dissertation Acceptance and Practical Reason. Lockhart was trying to answer more or less the same questions I am, although he fails to distinguish clearly between rankings and statuses, and for this reason courts confusion in ways we'll see later. Ross was trying to answer a slightly different question – not “What is it rational to do under normative uncertainty?” but “Which normative hypothesis is it rational to accept under normative uncertainty?”, where acceptance is something like “treating as true for the purposes of reasoning”. Much of what Ross says about rational acceptance is applicable to our discussion about rational action, but some of it is not. To take a very obvious dissimilarity, he devotes a chapter of his dissertation to the rationality of optimism. Insofar as optimism is an attitude and not a set of behaviors, it wouldn't make sense for me to tackle the same issue.

One of the main differences between this work and Lockhart's and Ross's is one of breadth. Lockhart discusses, in separate chapters, the morality of having an abortion, the morality of the Roe v. Wade decision, and the morality of professional confidentiality. His book also includes a fascinating chapter about whether the proper normative focal points are single actions or successions of actions. Finally, he arrives at his preferred view of rationality under normative uncertainty only after canvassing a series of alternatives, applying them to well-described examples, and showing that they deliver counterintuitive verdicts. By contrast, I consider “applied” cases like abortion and so forth only to introduce theoretical points; I consider the evaluation of successions of actions only in response to an objection; and I've already told you which view of rational action under normative uncertainty I favor.

Ross spends considerable time applying his view of rational acceptance to the issues of: whether we ought to accept theories that are, to various degrees, “deflationary”;

whether optimism is rational; whether we ought to accept subjectivist theories of moral reasons; whether we ought to accept skeptical hypotheses in epistemology; and whether we ought to accept the same moral theory over time or switch between different ones. These applications are fascinating, but I don't pursue any of them specifically – again, partly because they don't arise when our focus is on actions to perform rather than on hypotheses to accept.

On the other hand, this dissertation considers certain topics in much greater depth than either of Lockhart's or Ross's works. There is a problem called the Problem of Value Difference Comparisons that it's incumbent on each of us to solve, given the theory of rationality to which we all subscribe. Lockhart spends one-half of a chapter suggesting a solution to this problem; Ross spends a few pages. I spend two lengthy chapters on this problem, and explain why Lockhart's and Ross's purported solutions are unsatisfactory. Lockhart and Ross each defend expected value maximization through a few examples designed to show only that this theory is superior to some alternatives that are much, much different and more extreme. They also spend very little time dealing with philosophical objections to the theory. I show more comprehensively how the preferred theory of rationality is better than other theories, even some theories that are very similar to it and not obviously objectionable. I also consider more and deeper objections to the preferred theory. Finally, I consider theoretical wrinkles that Lockhart and Ross either ignore or treat cursorily: incomparability, parity, and absolute prohibitions, as well as the crucial difference between rankings and statuses.

Normative Uncertainty Throughout History[25]

Of the figures that most contemporary philosophers would place in the "canon", only two addressed the topic of action under normative uncertainty. These two are Pascal and Hegel, both of whom adopted what has been called a "Rigorist" position in the Catholic moral-theological debate. Pascal's contribution to the debate consisted of some of his earlier Provincial Letters, which tended to take the form of suggestive anti-Jesuit polemic than of serious moral philosophy or theology.[26] At any rate, the Provincial Letters were condemned by the Church shortly after their publication, and for this reason had less of an impact on the course of this debate than they otherwise would've. Hegel's contribution consisted of a brief passage in his Philosophy of Right that suggests that he failed to grasp the distinction between objective and belief-relative normative notions.[27]

[25] I would not have known about the body of work discussed in this section were it not for a meeting with Desmond Hogan.

[26] See Pascal (1853), letters 4, 5, and 6.

[27] Hegel writes:

> "But since the discrimination between good and evil is made to depend on the various good reasons, including also theological authorities, despite the fact that they are so numerous and contradictory, the implication is that it is not this objectivity of the thing, but subjectivity, which has the last word. This means that caprice and arbitrary will are made the arbiters of good and evil, and the result is that ethical life, as well as religious feeling, is undermined. But the fact that it is one's own subjectivity to which the decision falls is one which probabilism [a Catholic view about what to do under moral uncertainty; see the discussion below] does not openly avow as its principle; on the contrary...it gives out that it is some reason or other which is decisive, and probabilism is to that extend still a form of hypocrisy." p. 141.

Given the taxonomy of normative concepts I'm employing, the obvious rejoinder is that the "objectivity of the thing" has the last word on what we objectively ought to do, while subjectivity has the last word on what we subjectively ought to do. Hypocrisy averted. Nor am I being unfair in criticizing Hegel for missing a distinction he couldn't have possibly known about. As we'll soon see, the moral theologians to whom he was responding made a substantially similar distinction crystal clear.

The more important contributors to the debate were not philosophers of the secular canon, but theologians whose intellectual influence within Catholicism far outstripped their influence outside of it.

To see how their work bears on the questions of this dissertation, it's necessary to get a bit clearer on the conceptual scheme that undergirded their debate, and compare it with the one that undergirds ours. The easiest difference to see is that, while I'm concerned with the rationality of actions performed under uncertainty, the theologians were concerned with the probability-relative value of actions performed under conditions of divided non-personal probability.[28] If we wanted to speak very loosely, we might say that my focus is on uncertainty in the subjective sense, whereas theirs was on uncertainty in the objective sense. (This is loose because "uncertainty in the objective sense" is not really uncertainty; it is simply objective probability.) Given their focus on non-personal probability, it was of course incumbent upon the theologians to say which features of the world can affect the probability of a normative hypothesis. It was assumed in the debate that two features, in particular, were relevant – the number of Church fathers who supported some hypothesis, and the authority of those Church fathers. These dimensions could be "traded off" against one another; for example, the word of one especially authoritative scholar like St. Alphonsus Liguori – the Father of serious work on normative uncertainty, and perhaps not coincidentally, the patron saint of those suffering from scrupulosity – was worth the word of several scholars of lesser reputation.[29] This difference between uncertainty and objective probability is less important than it might otherwise be, however, since the Catholics to whom this debate was addressed probably

[28] See The Catholic Encyclopedia (1913), entry on "Probabilism".

[29] Ibid.

did, as a matter of psychological fact, apportion their credences roughly to the numbers and authority of scholars on either side of a question.

Another, more fraught distinction between our debate and the Catholic ones concerns what I'd earlier called “inputs” and “outputs”. The principles of rationality I have in mind take probabilities regarding objective normative features as inputs, and yield judgments about rational normative features as outputs. By contrast, the principles at issue in the Catholic debate, which they called “Reflex Principles”, take as inputs probabilities regarding the material sinfulness of actions, and deliver as outputs judgments about the formal sinfulness of actions.[30]

Might we think of the material/formal distinction as simply the objective/subjective distinction described in other terms? This would be inappropriate, for it's not clear that a formal sin is a mere defect in rationality. A formal sin is defined as a “sin of the conscience”.[31] Sins of the conscience are failures to do what's rationally required, to be sure, but they are more than that. They're failures to do what's rationally required that owe part of their explanations to the agent's insufficient concern for, or desire to do, what's objectively required. By contrast, acting in a risk-averse way out of sufficient concern for doing what's objectively valuable is not a defect of conscience. It's a run-of-the-mill failure of rationality, or so I claim. There is room to doubt, therefore, that to merely act on the wrong Reflex Principle is to commit a formal sin. Doing the former is more plausibly regarded as only a necessary condition on doing the latter.

The foregoing can be no more than a quick-and-dirty assessment of the conceptual

[30] Ibid., entry on “Sin”.

[31] Ibid.

scheme within which the Reflex Principle debate was prosecuted. The scheme of normative and related concepts that late Medieval and Early Modern Catholic theologians employed is so different from the scheme contemporary secular moral philosophers employ that one can't simply look at bits and pieces of their scheme in isolation from the rest and offer a theory about how these match up to bits and pieces of our scheme. Concepts are more holistically determined than that.

Let's press on, though, and consider how the positions in the Reflex Principle debate might fare as views about rationality under normative uncertainty. The first thing to say is that none of these is put forward as what I'd consider a comprehensive position. The focus in the Catholic debate is, as far as I can tell, exclusively on situations in which there's some chance that doing an act is a material sin, and no chance that not doing it is a material sin. In such cases, there's risk of material sin only on one side; the question is what probability of material sin, on that side, gives rise to formal sin. A comprehensive position would also address cases in which one risked material sin whatever one did.

Rigorism is the view that if there's any chance that an act is a material sin, then it's a formal sin. At the opposite pole, Laxism is the view that if there's any chance that an act is not a material sin, then it's not a formal sin. Both of these views are extreme, and both were formally condemned by the Church.

Probabiliorism is the view that only if it's more probable than not that an act is not a material sin, then it's not a formal sin. Equiprobabilism is ever-so-slightly more lax. It's the view that an act is not a formal sin only if it's more probable than not that it's not a material sin, or if the probabilities of it's being a material sin it's not are equal. Probabilism – the dominant view in this debate – is more lax still. It′s the view that an act is not a formal sin so long as there's a reasonable probability that it's not a material sin, even if it's more probable that it is a material sin.[32]

As I'll suggest in Chapter 6, these views are all too coarse-grained to be right. For not only do probabilities matter; the severity of potential material sins also matters, and none of these positions takes that into account. It might be countered that all sins are equally severe, but this strikes me as so unreasonable that I'd be surprised if anyone sincerely believed it. Certainly, it's not consistent with Catholic moral theology, which distinguishes between less severe venial sins, and more severe mortal sins, and more generally, between sins for which the cost of repentance is higher, and those for which it is lower.

This fact has not gone unobserved in the Catholic tradition, and a position called Compensationism has gained currency as a result. This is the view that whether an act is a formal sin depends not only on the probability that it's a material sin, but also on a) how severe a material sin it would be if it were one, and b) how much value might be gained by risking material sin.[33] This strikes me as an eminently sensible position, and in Chapter 6, I'll be defending views about action under uncertainty about statuses that are contemporary variants on Compensationism. More loosely, Compensationism is allied with what, in the next chapter, I'll call comparativist theories of rationality under uncertainty – ones that take into account degrees of value, and not just degrees of belief. The other Reflex Principles canvassed so far are closer to what I'll call non-comparativist theories of rationality under uncertainty, which don't take into account degrees of value,

[32] See The Catholic Encyclopedia (1913), The New Catholic Encyclopedia (2002), and Prümmer (1957). Probabilism is defended in Liguori (1755).

[33] Ibids., defended in Prummer (1957).

only probabilities of hypotheses and ordinal rankings on those hypotheses.

With all that said, this is a dissertation the aim of which is to solve a set of problems regarding action under normative uncertainty; it's not a dissertation in the history of ethics. That's one reason why I'll return to the Catholic debate only occasionally. But I do think that there's gold to be mined here in the future. Consider this Probabilist argument against Equiprobabilism and Probabiliorism: Probabilism is more probable than either of these other views, since the vast majority of theologians accept it. Therefore, Equiprobabilists and Probabiliorists are committed by their own principles to give up these principles and switch to Probabilism instead.[34] Fascinating, huh? For my own part, I think this argument fails, and some of what I say in Chapter 7 will go towards showing this, but seeing why it fails involves addressing issues that have been all but ignored by the great secular moral philosophers. So I'll conclude by marking this historical topic as an avenue for future research, and press on to a positive defense of my own contemporary version of a Reflex Principle.

[34] The Catholic Encyclopedia (1913), entry on “Probabilism”.

CHAPTER TWO: RATIONALITY UNDER NORMATIVE UNCERTAINTY

Introduction

What is it most locally rational to do under normative uncertainty, relative to your credences in objective rankings of actions? Perhaps the most natural answer to this question is: It’s most rational to act in accordance with the ranking in which you has the highest credence. So if your degree of belief is highest that Action A is better than Action B, then it's more rational to do A than to do B. We might offer a similar answer in the case of uncertainty about normative theories: If your degree of belief is highest in Negative Utilitarianism, it's most rational for you to do whatever Negative Utilitarianism says is best in any given situation. That this answer seems so natural is, I suspect, one reason why so little attention has been paid to the question of what to do under normative uncertainty.

We should be leery of this approach, though, because some similar courses of action under non-normative uncertainty seem so clearly mistaken. Suppose that I am deciding whether to drink a cup of coffee. I have a degree of belief of .2 that the coffee is mixed with a deadly poison, and a degree of belief of .8 that it’s perfectly safe. If I act on the hypothesis in which I have the highest credence, I’ll drink the coffee. But this seems like a bad call. A good chance of coffee isn’t worth such a significant risk of death – at least, not if I assign commonsensical values to coffee and death, respectively.

Similarly, suppose there’s some chance that A is objectively better than B, and an ever-so-slightly greater chance that B is better than A. I also believe that, if A is better than B, then A is saintly and B is abominable; but if B is better than A, then A is slightly nasty and B is merely okay. Despite the fact that my credence is higher that B is better than A, it still seems as though it's rational to do A instead, since A's 'normative upside' is so much higher than B's, and its 'normative downside' not nearly as low.

Here, then, is a more promising answer: It's most rational to perform the action with the highest Expected Objective Value (EOV). We get the EOV of an action by multiplying the subjective probability that some ranking is true by the objective value of that action if it is true, doing the same for all of the other rankings, and adding up the results.[35] This strategy is sensitive not only to degrees of belief, but also to degrees of value – the relative sizes of the upsides and downsides of actions. We do not need to know the absolute values of actions, if there even are such things, in order to determine how actions rank in terms of EOV. All we need to know are the ratios of value differences between actions on different normative hypotheses. If the difference between A and B, if A is objectively better, is 4 times the difference between B and A, if B is objectively better, this is enough information, in combination with my credences in the two rankings, to tell me how A and B rank in terms of EOV.

Let's consider a concrete example. Suppose that my credence is .7 that the balance of reasons supports eating meat over not eating meat, and .3 that the balance of reasons goes the other way. If the "natural" theory of rationality under normative uncertainty is correct, then it's easy to see which is more rational: eating meat. But EOV maximization may yield a different result. For it's plausible to think that if "meat is murder", then by eating meat I'm an after-the-fact accessory to murder, which is pretty bad; but if eating

[35] More formally, EOV of A = $\sum_i$ p(Ranking$_i$)·v(A given Ranking$_i$), where p(PC$_i$) is the subjective probability that Ranking$_i$ is true, and v(A given Ranking$_i$) is the objective value of the action A given that Ranking$_i$ is true.

meat is morally innocuous, then by eating meat I'm eating meals that are, on the whole, only slightly more tasty than their vegetarian alternatives. If that's true, then the gulf between eating meat and not if the former is better is arguably much smaller than the gulf if the latter is better, and so EOV maximization may support not eating meat. This is a simplified case, of course, but it gives you some idea of how the theory functions in application.

In this chapter and the next, I'll argue that the the most rational action under uncertainty is the one with the highest EOV. More generally, I'll argue that A is more rational than B just in case A's EOV is higher than B's, that A and B are equally rational just in case their EOV's are equal, and that A and B are either rationally on a par or rationally incomparable just in case it's indeterminate which has a higher EOV.

In this chapter I'll argue in favor of EOV maximization and against rival theories. In the next chapter, I'll consider some objections to EOV maximization and offer some replies. EOV maximization is a type of comparativist theory of practical rationality under normative uncertainty. Comparativist theories all say that whether it's more rational to do A or to do B under normative uncertainty depends on how the difference in value between A and B on one or more normative hypotheses compares to the difference in value between B and A on other normative hypotheses. Comparativist theories, then, are ones according to which the relative sizes of value differences matter. By contrast, non-comparativist theories give no role to the relative sizes of value differences in determining which actions are rational to perform under normative uncertainty. Non-comparativist theories are ones according to which the relative sizes of value differences don't matter.

I'm going to defend EOV maximization incrementally. First I'm going to argue that comparativist theories are better than non-comparativist theories. Then I'm going to argue that EOV is better than all of the other comparativist theories. I'll close by discussing how EOV can be indeterminate, how this is constitutive of either "rational parity" or "rational incomparability", and how we should behave in the face of rationality parity and rational incomparability.

Comparativist vs. Non-Comparativist Theories

The "natural view" with which we began the chapter is a kind of non-comparativist theory. We can call this theory Credence Pluralitarianism (CPlur), because it says that it's most rational to do what's objectively best by the plurality of one's credence. It's non-comparativist because it gives us a rationality-ranking of actions based only on the agent's credences in normative hypotheses, and the ordinal rankings of actions implied by those hypotheses. No comparisons of value differences are needed. Here are some other non-comparativist theories:

> Credence Majoritarianism (CMaj): It's more rational to do A than to do B just in case your credence is at least .5 that is objectively better to do A than to do B.

The difference between this theory and CPlur is that this theory deems A more rational only if it's being objectively better than B enjoys the majority of your credence, whereas the previous theory deems A more rational so long as its being objectively better than B enjoys more of your credence than B's being objectively better than A. So if my credence distribution were as follows:

.4 that A is better than B

.3 that B is better than A

.3 that A and B are equal

...then CPlur will count A as more rational, while CMaj will not.

Some other non-comparativist theories:

Credence Paretianism (CPar): It's more rational to do A than to do B just in case you have some credence that A is better than B, and no credence that B is better than A.

Credence Supermajoritarianism (CSuper): It's more rational to do A than to do B just in case your credence is at least [insert any number greater than .5 and less that or equal to 1.0] that it is better to do A than to do B.

Credence Absolutism (CAbs)*:* It's more rational to do A than to do B just in case your credence is higher that A is at least the Nth best thing to do in a situation than that B is at least the Nth best thing to do in that situation, for some value of N.

CAbs may be illustrated by an example. Suppose that we divide the possibility space in some situation into four actions – A, B, C, and D. And suppose further that your credence distribution over rankings is as follows:

> .25 that A is better than B, B is better than C, and C is better than D
>
> .25 that C is better than A, A is better than B, and B is better than D
>
> .25 that D is better than A, A is better than B, and B is better than C
>
> .25 that D is better than C, C is better than B, and B is better than A

Now, it's clear that the probability of A's being better than B is .75. But suppose we're credence absolutists, and the value our "N" is 3. Then in determining which of A and B is more rational, we care which one is more likely to be at least the 3rd best thing to do in this situation. And that action is B, for B is certain to be at least the 3rd best thing to do, while A has only a .75 of being at least the 3rd best thing to do.

There are, of course, different versions of CAbs, each corresponding to a different value for N. The one that most naturally comes to mind gives N a value of 1:

> Credence Optimalism (COpt): It's more rational to do A than to do B just in case your credence is higher that A is the best thing to do in a situation than that B is the best thing to do in that situation.

There are other non-comparativist views as well. This is just a sample. Again, what they have in common is their insensitivity to the relative sizes of value differences according to hypotheses. We have good reason to think that this insensitivity dooms these theories to failure.

For one thing, it is responsible for non-comparativist theories yielding highly counterintuitive sets of judgments about particular cases. Suppose that a magistrate must decide whether to convict an innocent man in order to avert a riot, or to acquit the man and thereby prompt the riot. Suppose that the magistrate's credence is .6 that it's better to acquit, and .4 that it's better to convict. On one way of filling out the scenario, the bad consequences of the riot will be trifling – a few broken windows here and there, some traffic stalls, and some other minor inconveniences. Then the obvious choice seems to be acquittal; it's more probably the better thing to do, and the factors that might render conviction better are of such piddling significance. And indeed, this is what a theory like CPlur says. But on another way of filling out the scenario, the consequences of the riot will be very, very bad; thousands will die, homes and businesses will burn, and so forth. Since the magistrate's credence is still higher that acquittal is better, CPlur will still demand acquittal. But now this seems like a less obvious option.

The real problem with CPlur is not that it delivers a counterintuitive judgment about the second version of this case in particular. It's that it delivers the same judgment, no matter how bad we make the consequences of the riot, so long as we hold the credences constant. For that matter, it delivers the same judgment, whether the agent disvalues convicting the innocent a lot or only a little, so long as we hold the credences constant. More generally, it delivers the same judgment, no matter how the difference between acquittal and conviction, if the former is better, compares to the difference between acquittal and conviction, if the latter is better. Nor is this uniquely a feature of

CPlur. It's a feature of every non-comparativist theory, for the verdict of every such theory about cases like this one will be independent of how the aforementioned value differences compare. By consequence, it will be independent of many of the concrete features of the case that determine those value differences, such as the gravity of the consequences associated with the riot. This insensitivity to how the normative "stakes" on one hypothesis compare to the "stakes" on the opposing hypothesis represents a kind of normative blindness. Insofar as it's possible to compare value differences across normative hypotheses – and we'll take up this question later – the way these differences compare should play some role in determining what it's rational to do.

Furthermore, nearly every non-comparativist theory has one or the other of two unwelcome features. Either it delivers the consequence that the *more rational than* relation is intransitive, or it leads to violations of a version of the *Independence of Irrelevant Alternatives (IIA)*.[36] To see this dilemma, start by dividing non-comparativist theories into two general types – *relative-placement* non-comparativist theories, and *absolute-placement* ones. According to relative-placement theories, whether A or B is more rational will depend on how A and B rank *vis a vis* one another according to the rankings in which the agent has credence. CPlur, CMaj, CPar, and CSuper are all relative-placement theories; they "care" about how various rankings rate A and B against each other. By contrast, on absolute-placement theories, whether A or B is more rational will depend on how A and B rank *vis a vis* some larger set of actions available in the situation, according to the different rankings. CAbs is an absolute-placement theory. It "cares"

[36] That intransitivity and IIA-violation represent a kind of Scylla and Charybdis for ordinalist selection procedures in general is noted in Luce and Raiffa (1957).

about which of A or B is most likely to be the best, or the second-best, or not-the-worst, or what have you, and not, about how A and B compare to each other, as such.

Relative-placement non-comparativist theories typically yield an intransitive more rational than relation. Consider the following case: There are three possible actions – A, B, and C. The agent has credence of 1/3 that A is better than B, and B is better than C, 1/3 that C is better than A, and A is better than B, and 1/3 that B is better than C, and C is better than A.

On CPlur, CMaj, and versions of CSuper where the "magic number" is no greater than 2/3, it's more rational for the agent to do A than to do B, since there's a 2/3 chance that A is better than B, and only a 1/3 chance that B is better than A. It's also more rational to do B than to do C, since there's a 2/3 chance that B is better than C, and a 1/3 chance that C is better than B. Yet it's also more rational to do C than to do A, since there's a 2/3 chance that C is better, and only a 1/3 chance that A is better. Since A is more rational than B, B is more rational than C, and C is more rational than A, we have intransitive rationality-rankings. This result generalizes to versions of CSuper with higher "magic numbers". We simply need more actions and more possible rankings of those actions. There are two relative-placement theories that do not yield this result – CPar, and CSuper where the magic number is 1. But they escape intransitivity only at a high cost: the range of cases in which they say that one action is more rational than another is very, very small. Consider a case in which your credence is .99 that A is better than B, and .01 that B is better than A, and in which you believe that the difference between A and B on the former hypothesis is 100 times the difference between B and A on the latter hypothesis. Even then, neither of these theories will deliver the result that A is more rational than B.[37]

Now, the intuitive thing to say about cases like the one represented in Diagram 2.1 is that A, B, and C are all equally rational. They each appear first in one ranking, second in another, and third in another, and the rankings are equiprobable. And absolute-placement theories can capture this intuition. For example, COpt will not say that any of A, B or C is more rational than any of the others, for the agent's credence is the same in each's being the best action in the situation. Versions of CAbs with other values of N will say the same.

But there is a problem with absolute-placement theories; they lead to the violation of a version of the Independence of Irrelevant Alternatives (IIA)[38]:

> A is more rational than B when actions (1...n) are the alternatives just in case A is more rational than B whatever the alternatives.

For suppose the agent's credence is .4 that A is better than B, and B is better than C, .35 that C is better than A, and A is better than B, and .25 that B is better than C, and C is better than A.

On COpt, we get the result that it's more rational to do A than to do C, since there's a .4 chance that A is the best action in the situation, and a .35 chance that C is. But suppose that we remove B from consideration. Whatever case one might make that B is

[37] Thus, they violate an analogue of what Arrow (1951) termed the "unlimited domain" requirement.

[38] Ibid.

more rational than either of these two, its presence or absence should not affect the rationality-ranking of A and C. And yet it does. If B is removed, then the new rankings look like this: There's a .4 chance that A is better than C, and a .6 (.35 + .25) chance that C is better than A.

Since there's a .6 chance that C is better than A, and only a .4 chance that A is better than C, we get the result that C is more rational than A, according to COpt. And this kind of argument generalizes. Whenever the rationality-ranking of two actions depends on the absolute places of those actions in the various objective value rankings, this rationality-ranking will depend on which other alternative actions are available. This is because the absolute position of an action in a ranking depends on which other actions are included in the ranking.

To complete this argument against non-comparativist theories, we'd have to show that intransitivity of the more rational than relation and violation of IIA are undesirable. This is not something I'm going to undertake here. There is already a well-developed literature about both transitivity and IIA, and I have no grand insights to add.[39] Suffice it to say, I think the cases against intransitivity and IIA violation are strong enough that, in showing that various non-comparativist theories yield these features, I've put at least a dent in the armor of these theories.

I do, however, want to close this section with a reminder about the debates over transitivity and IIA. The most popular way of arguing against transitivity and IIA is through the use of examples about particular cases. Here are some instances of that form of argument:

[39] See Arrow (1951), Chernoff (1954), Luce and Raiffa (1957), Sen (1970), Temkin (1987), (1995), and (forthcoming), and Broome (1995) and (2004).

> Against Transitivity of Evaluative Relations: It's better to visit Rome than to go mountaineering, better to stay home than to visit Rome, but not better to stay home than to go mountaineering. That's because mountaineering is scary, so it's better to visit Rome; sightseeing is boring, so it's better to stay at home; but staying at home rather than going mountaineering is cowardly, so it's not better to stay home.[40]

> Against IIA for Evaluative Relations: When the only two alternatives are visiting Rome and staying home, it's better to stay home, since sightseeing is boring. But when the alternatives are visiting Rome, staying home, and mountaineering, then it's better to visit Rome than to stay home; to do the latter would be cowardly in a way that visiting Rome would not.

Let's just assume arguendo that these counterexamples show the transitivity requirement and IIA to be false. It would be a mistake to conclude that the non-comparativist theories we've been discussing are no worse off for yielding violations of these requirements. For while these counterexamples show that these purported requirements are not genuine, they do so by demonstrating that they fail in particular cases, due to the substantive properties of the objects of choice in those cases. But non-comparativist theories of rationality yield violations of one or the other of these purported requirements for reasons that have nothing to do with substantive properties of actions.

[40] This example is adapted almost verbatim from Broome (1995), p. 101.

They yield such violations in virtue of an agent's having credences distributed in a certain way over ordinal rankings of actions. Look at it this way: we didn't say anything about the substantive features of the actions under consideration – features which might explain failures of transitivity and IIA. We just called the actions "A", "B", "C", and so forth, and *still* we got violations of these requirements. That shouldn't happen.

Here's substantially the same point, presented in scientific argot: Counterexamples analogous to those above might disconfirm the hypothesis that the *more rational than* relation is transitive. But they do nothing to disconfirm the hypothesis that the *more rational than* relation is transitive, *except when the related actions have whichever substantive features render the relation intransitive with respect to these actions*. Now, that second hypothesis is obviously wishy-washy, and not all that interesting to defend. But give it at least this much – it's as intuitive as you can get in the domain of practical philosophy. And if any of a certain class of non-comparativist theories of rationality is right, then this hypothesis is wrong. That tells against theories of that class. *Mutatis mutandis* for the Independence of Irrelevant Alternatives.

EOV Maximization vs. Other Comparativist Theories

Comparativist theories of rationality under uncertainty are sensitive to the relative sizes of value differences across hypotheses, so all of them are in this way more similar to EOV maximization than any of the non-comparativist theories were. Since they all share this attractive feature with EOV maximization, it will be more difficult to score a decisive victory against them. I should say that I don't find this fact especially troubling; one of my major goals in writing this dissertation was to clear the way for and show the advantages of comparativist theories in general, and if my arguments to this point have been successful, then I've accomplished that goal. Still, I persist in thinking that EOV maximization is the best of the comparativist lot, and in this section, I'll present to you the considerations that I think most powerfully support this position.

We can see in highest relief the contrast between EOV maximization and its comparativist cousins by stating the former a bit more formally, and then showing in these terms how it differs from the latter. Expected Objective Value is given by the following formula:

$$\Sigma_i\ p(S_i) \cdot v(A \text{ given } S_i)$$

Where $p(S_i)$ is the probability of a state of the world, S_i, obtaining, and $v(A$ given $S_i)$ is the value of the action under consideration, given that state of the world. A state of the world includes all facts about that world, both normative and non-normative.[41] Since our concern is uncertainty specifically about the normative, we'll want a version of this formula that somehow "separates out" the normative aspects from the non-normative aspects of states of the world. There are three equivalent ways of doing this. First:

$$\Sigma_{i,j}\ p(N_i \wedge NN_j) \cdot v(A \text{ given } (N_i \wedge NN_j))$$

$p(N_i \wedge NN_j)$ is the probability of the state of the world characterized by normative facts N_i and non-normative facts NN_j, and $v(A$ given $(N_i \wedge NN_j))$ is the value of the action

[41] I use "normative fact" in its minimally committing sense: It is a normative fact that murder is wrong just in case murder is wrong.

under consideration, given both sets of facts. Basically, we're just dividing the world into normative and non-normative facts.

The second, and as we'll see, more helpful way of separating out the two types of facts is this one:

$$\Sigma_{i,j}\ p(N_i|NN_j) \cdot p(NN_j) \cdot v(A \text{ given } (N_i \wedge NN_j))$$[42]

The three multiplicanda are the probability of normative facts N_i conditional on non-normative facts NN_j, the probability of NN_j, and again, the value of A given all of the facts that characterize some state of the world. So the EOV of an action is a multiplicative function of the value of that action in a world, the probability of a world with those non-normative features, and the probability of a world with those normative features, conditional on it's also being a world with those non-normative features.

And let me introduce one last formulation, which will be useful in contrasting the theory of rationality I'll be defending with its close competitors:

$$\Sigma_{i,j}\ F1(p(N_i|NN_j)) \cdot F2(p(NN_j)) \cdot F3(v(A \text{ given } (N_i \wedge NN_j)))$$

Where F1(*), F2(*), and F(*) are functions of the two probabilities and the value, respectively, and F1(x) = x, F2(x) = x, and F3(x) = ax + b, where a is a positive number.

[42] These last two formulations can be shown to be equivalent by the definition of conditional probability, $p(A|B) \cdot p(B) = p(A \wedge B)$:

Σi,j p(Ni ^ NNj) · v(A given (Ni ^ NNj))

Σi,j **p(Ni|NNj) · p(NNj)** · v(A given (Ni ^ NNj)).

Here's the idea: F1(*) and F2(*) are "probability-weighting" functions, and F3(*) is a "value-weighting" function. If F1(p(A)) = p(A) and F2(p(B)) = p(B) then the theory is not a probability-weighting theory. It just takes probability values as inputs, and spits out the very same values as outputs. If F3(v(A given (N_i ^ NN_j))) = a(v(A given (N_i ^ NN_j))) + b, then the theory is not a value-weighting theory, for the value of F3(*) will simply be a linear function of the value of A. EOV maximization is unique in being the only non-probability-weighting, non-value-weighting theory of the general form above. Every other theory of this general form is either probability- or value-weighting. (Question: why can a theory count as non-value-weighting if the weighting function is linear, but count as non-probability-weighting only if the weighting function is the identity function? This is because probability values must be between zero and 1, inclusive, and some linear F1(*) and F2(*) functions will produce violations of this. Other than that, there is no problem, from the point of view of expected utility theory, of making them linear, but not identity, functions.)

Both probability- and value-weighting theories of this general form are sometimes called risk-sensitive. But the more sophisticated ways of representing risk-sensitivity usually involve a different form. Lara Buchak has defended a risk-sensitive theory of rationality on which the following quantity is maximized:

v(A given S_2) + F(p(S_1)) · (v(A given S_1) – v(A given S_2)); where F(*) does not equal (*) and v(A given S_1) is greater than v(A given S_2)[43]

[43] Buchak (ms #1), p. 11.

This is like EOV maximization, probability-weighting, and value-weighting in that it depends on probabilities of states of affairs, and the values of actions given those states of affairs. It is different in that it involves multiplying the probabilities of states not by the values of actions given those states, but rather by the differences between values of actions given different states: by v(A given S_1) – v(A given S_2), not by v(A given S_1), say. (This quantity is different from the others in that it involves the value of A in only two states of affairs, S_1 and S_2, while the other quantities involve the value of A in any number of states of affairs. The general, sigma-notational form is unnecessarily complicated for our purposes, but may be found and explained in the appendix of Buchak (ms #1).)

All of the views of rationality just discussed – EOV maximization, the weighting theories, and risk-sensitivity properly so-called, are comparativist, insofar as they make what it's rational to do depend on the relative sizes of value differences across different hypotheses. The task now is to show that good old EOV maximization is the best of the lot.

Before I start in on that, I want very quickly to canvas the debate about expected value maximization in general. In many quarters, it’s taken to be obvious that the rational thing to do under uncertainty is whatever maximizes expected value. This is the state of play, for example, in most of decision theory, philosophy of science, and formal epistemology. Neither Ross's nor Lockhart's monographs, despite their impressiveness in other respects, contained much defense of expected value maximization as against other comparativist theories.[44] Other philosophers have gone so far as to say that expected value maximization is by definition rational or that it's impossible to provide arguments for it.

I think there's not much to be said for this ostrich-like stance. That so many ethicists and decision theorists, who are presumably competent speakers of languages containing some or other translation of "rational", regard it as open to debate whether it's rational to maximize expected value suggests that this is not a definitional truth. And if it's not, then this position cries out for substantive arguments in its favor.

So if we have to argue for EOV maximization, how do we do it? There is one kind of argument that, if successful, would seem to seal the victory for EOV maximization. This is the Dutch Book Argument. Dutch Book arguments purport to show that agents who violate some putative rule of rationality will sometimes behave under uncertainty in such a way that guarantees a loss. One kind of Dutch Book argument has been pursued against agents who are risk-sensitive.[45]

Suppose an agent values money linearly. That is, her difference in objective value between \$1 and \$2 is the same as her difference in objective value between \$5 and \$6, and so on. If she is risk-neutral, the value of a gamble with an expected payoff of \$1 is the same as the value of a guaranteed \$1. So, for example, she will be indifferent between \$1, and a gamble that yields \$2 if some event E happens, where the probability of E is .5, and nothing of E doesn't happen, where the probability of not-E is also .5. But if an agent is risk-sensitive, say, the value of a gamble with an expected payoff of \$N will not always

[44] Lockhart (2000) and Ross (2006) and (ms).

[45] See Resnik (1987).

be the value of a guaranteed $N. She may be indifferent between $.95 and the $2/$0 gamble if she is risk-averse; she may be indifferent between $1.05 and that gamble if she is risk-seeking.

Problems arise for the risk-sensitive agent when she confronts two or more non-independent gambles. For example, the risk-averse agent will take $.97 rather than a gamble that yields $2 if E, and $0 if not-E. She will also take $.97 rather than a gamble that yields $0 if E, and $2 if not-E. If she is offered these gambles at the same time, she will take the $.97 twice over, and end up with $1.94. If she had taken the riskier option both times, however, she would have ended up with $2 ($2 + $0 if E, and $0 + $2 if not-E). Her risk aversion cost her $.06. Not an "expected" $.06. Six guaranteed cents – something of value to the agent regardless of her attitudes towards risk.

As others have pointed out,[46] this scenario is too bad to be true. Here are two possibilities: 1) The agent does not know that the gambles both depend, in different ways, on E, and 2) The agent does know that the gambles both depend on E. If the agent does not know about the non-independence of the gambles, she does not know that she is choosing $1.94 rather than $2, and so her doing so is no strike against her rationality (although it is obviously unfortunate for her). If, on the other hand, she agent does know about the non-independence of the gambles, then the situation is misleadingly described. The agent is really choosing between a guaranteed $1.94 (i.e. $1.94 if E, $1.94 if not-E) and a guaranteed $2 (i.e. $2 if E, $2 if not-E). Since no risk inheres in either option, and the reward is greater with the second option, nothing in her attitudes towards risk compels her to choose the first option. The problem this version of the Dutch Book

[46] See, e.g., Kyburg (1978), Schick (1986), Maher (1993), Buchak (ms #1).

argument, then, is this: It presumes that an agent will make an obviously foolish choice simply because she will make each of two separate choices whose possible outcomes are, together, the outcome of the foolish choice. There are ways of modifying the scenario so that the agent never faces a single choice between $1.94 and $2. We can present gambles one after the other, for example. But if the agent knows that the second gamble is coming, then she can coordinate her choices over time so that she ends up with $2 rather than $1.94.[47,48]

Obviously there's more to say about the merits of Dutch Book Arguments. Perhaps one can be made to work. But they've been subject to many other effective criticisms,[49] and I don't believe anyone has ever satisfactorily addressed the aforementioned worry. Given that these arguments are in such poor standing, I'll look elsewhere for a defense of EOV maximization.

We might also argue for EOV maximization via intuitions about cases, just as we did when we observed that non-comparativist theories give the wrong result in cases like the "magistrate" one we considered. But non-comparativist theories fell short in those cases because they were utterly insensitive to degrees of value. Comparativist theories, by contrast, are sensitive to degrees of value. Of course, there will be some comparativist theories that give extremely counterintuitive results. Consider an agent characterized by an F1(*) function such that she is an extreme probability-weighter. She may regard an action guaranteed to have a value of V as more rational than an action with a very slight

[47] The best discussion of sequential choice of which I'm aware is McClennen (1990).

[48] Thanks to Brian Weatherson for helping me to see this.

[49] See also Levi (2000) and Hajek (2005).

chance of having a value of V - N, and a very great chance of having a value of V + M, where the difference between V + M and V is, let's just say, 500 times the difference between V and V - N. EOV maximization is to be preferred to this level of probability-weighting.

But what about more reasonable types of weighting and sensitivity? When it comes to comparing EOV maximization with theories of this sort, my guess is that the cases will be a wash. For every case that, intuitively, EOV maximization gets right and a plausible alternative theory gets wrong, there'll be another case that seems to go the other way. And of course, there'll be many cases that all of the theories in question can easily capture. So if we're going to show that EOV maximization is the right theory, it won't be through a cavalcade of examples.

Instead, this is how I'll press my case for EOV maximization. First, I'm going to put the burden of persuasion on the defenders of alternative comparativist theories in two ways. The first way is through a "parity of reasoning" argument. The basic idea is that the proponent of risk-sensitivity or weighting under normative uncertainty must explain why we should behave differently thereunder than under non-normative uncertainty. The second way is as follows: The action with the highest EOV is, obviously, the action with the most expected value. All sides agree that, ceteris paribus, more EOV is a good thing. So the burden is on the defender of an alternative to show why we should sometimes "sacrifice" EOV to achieve his preferred distribution of EOV over epistemically possible worlds.

After I present these two “burden-laying” arguments, I’m going to mount a more direct defense of EOV maximization. First, I'll provide a series of arguments built around the well-known result that, subject to certain assumptions, a consistent policy of expected value maximization is certain to yield the maximum actual (i.e. not merely expected) value over the long term. Philosophers have too hastily dismissed this result as irrelevant to real-life decisions, and have failed to appreciate that it belongs to a family of results, each of which may figure as a premise in a plausible argument for EOV maximization. The second argument is, to my knowledge, a novel argument for expected value maximization – the Winning Percentage Argument. I’ll show that the action with the highest EOV will also have the highest probability of being better than a randomly selected action. Or, to telegraph my forthcoming neologism, it will have the highest “Winning Percentage”. What exactly this means and why it is normatively significant will be discussed later on.

Parity of Reasoning

Many people have suggested to me that EOV maximization is the right view of rational action under non-normative uncertainty – for example, when we don't know about the economic consequences of a Federal bailout – but not under normative uncertainty – when we don't know whether egalitarianism or prioritarianism is right. I want to urge that, if you firmly believe the first part of that view, you should reject the second part, and instead come to believe that EOV maximization is rational under normative uncertainty as well. If you don't believe the first part of that view, however, then you can move on safely to the next argument.

There is no formal incoherence in one's F1(*) function being different from one's F2(*) function – or in English, in having one view about rationality under normative uncertainty, and a different view about rationality under non-normative uncertainty. All the same, this sort of divergence demands an explanation. For it's typically just assumed that one's behavior under uncertainty about one domain ought to match one's behavior under uncertainty about other domains. Consider once again the general form common to EOV maximization, probability-weighting, and value-weighting:

$$\Sigma_{i,j}\ F1(p(N_i|NN_j)) \cdot F2(p(NN_j)) \cdot F3(v(A \text{ given } (N_i \wedge NN_j)))$$

Separating out the probabilities of normative propositions from probabilities of non-normative propositions allows us to make room for the possibility that F1(*) and F2(*) are not the same function.

But of course, we can "separate out" even more. Just as we can divide up the world into the normative and the non-normative, we can further divide the non-normative into facts about the United States of America, and all other non-normative facts. It's possible for an agent to have one probability-weighting function with respect to normative facts, another with respect to non-normative facts about the United States, and the identity function with respect other non-normative facts. We could represent this way of dividing things up as follows:

$\Sigma_{i,j,k}$ $F1(p(N_i|\text{NN-USA}_j)) \cdot F2(p(\text{NN-USA}_j|\text{NN-OTHER}_k)) \cdot F3(p(\text{NN-OTHER}_k)) \cdot F4(v(A \text{ given } (N_i \hat{} \text{ NN-USA}_j \hat{} \text{ NN-OTHER}_k)))$; where N_i, NN-USA_j, and NN-OTHER_k are just what you'd expect, and where F1(*), F2(*), and F3(*) are not the same function.

That's kind of a goofy example, but the same kind of separating out could be done along other, more philosophically interesting lines. We might divide the world into facts that can be known *a priori*, and facts that cannot; into necessary and contingent facts; into mental and non-mental facts. This makes room for the view that one’s behavior under, say, uncertainty about the mental – for example, uncertainty about whether comatose people are conscious, or about whether artificial systems can have intentional states – need have no relation whatsoever to one’s behavior under uncertainty about the non-mental.

And yet, this sort of heterogenous approach to uncertainty is rarely treated as a serious option. To consider the USA/non-USA case for a moment, nobody thinks that risk-neutrality is uniquely appropriate in US casinos, but that risk-aversion is uniquely appropriate in South African casinos. And let’s think about the mental/non-mental case. Consider two occasions of uncertainty: 1) You’re sure about some organism’s neurological structure, but unsure whether that organism can feel pain if kicked. In other words, you’re uncertain about the psychophysical “bridge” laws. 2) You’re sure about the psychophysical bridge laws, but unsure whether the organism has a neurological structure such that it will feel pain if kicked. It seems to me that we should respond in the same way to both of these kinds of uncertainty – that our attitudes towards risk in one case should mirror our attitudes towards risk in the other. If the proper response to uncertainty is invariant across all of these other ways of carving up the world, what's so special about the normative/non-normative way?

The More Expected Value, the Better

All remotely plausible views about rationality under uncertainty are Paretian with respect to objective value. That is to say, each of these views implies that the addition of objective value to one or more outcomes associated with an action performed under uncertainty makes this action more rational. Suppose, for example, that an action A has a .3 chance of having a value of 10, and a .7 chance of having a value of 5. On all theories that are Paretian with respect to objective value, another action, B, with a .3 chance of having a value of 11, and a .7 chance of having a value of 5 is better than A; so is another action, C, with a .3 chance of having a value of 10, and a .7 chance of having a value of 6. So all agree that EOV is a rationality-increasing feature of an action.[50]

But this is where the theories part company. The EOV maximization view says that total EOV is the only fundamentally rationality-affecting feature of an action

[50] This is one possible disanalogy between aggregation of value over epistemically possible worlds, and aggregation of value over persons. On some theories of the latter, like those that attach great significance to equality, the social utility function is not Paretian with respect to individual utility in the way that I've indicated. That is, the addition of more value "at" an individual will sometimes lower the overall quality of a state of affairs. This will happen when someone who is much better off than everyone else is made even better off, and everyone else stays where they are. This, of course, is a controversial feature of egalitarianism, and gives rise to the well-known "leveling down" objection. See Parfit (1997).

It's not hard to find an explanation for this possible lack of parallelism. If I'm badly off, and you're very well off, then it may be unfair in some way for more well-being to accrue to you, but not to me; it gives me grounds for complaint, if only against the indifferent cosmos. By contrast, we can make no sense of the accumulation of value at one epistemically possible world being unfair to another epistemically possible world, or of an epistemically possible world having ground for complaint.

performed under uncertainty. It does not matter how this value is distributed over possible outcomes. On all other views, however, distribution does matter, and not typically simply as a "tie-breaker" between actions with equal EOV's, either. These theories say that some actions with lower EOV's are more rational than others with higher EOV's, if the former have the preferred distribution of objective value over possibilities, and the latter do not. To put in another way, all other views demand that we be prepared to "sacrifice" some EOV to get either a riskier or less risky distribution of objective value over the epistemic possibility space. This means that proponents of these views are saddled with an extra burden of proof. They must show why distributive features matter in such a way that they're worth sacrificing EOV for. The proponent of EOV maximization faces no such demand; her theory sees nothing else for which we ought to sacrifice the acknowledged rationality-increasing feature of expected value.

<u>The Long Run, The Wide Run, the Medium-Long Run, and the Medium-Wide Run</u>.[51]

A policy of expected value maximization, implemented consistently over infinite occasions, will with a probability of 1 yield more total value than any other policy implemented consistently on those same occasions. This follows from the Strong Law of Large Numbers (SLLN), provided we make two assumptions. The first assumption is that the trials are independent – that the product of the probabilities of any outcome on any trial and any outcome on any other trial is equal to the probability of both of those outcomes obtaining. The second assumption is that the trials are quantitatively similarly enough so as to consist of actions with the same <u>mean</u> value and the same <u>variance</u>.

[51] Thanks for Lara Buchak and Jonathan Weisberg for helpful discussions about the material in this section.

Now, more actual value is obviously better than less actual value, and this shows that expected value maximization, repeated over and over again, is better than any of the other approaches, repeated over and over again. This supports the conclusion that expected value maximization is more rational in any particular instance than any other strategy, or so I claim. This is the basic argument; I'm sure it has the ring of familiarity.[52]

I expect two objections to this as an argument for EOV maximization under normative uncertainty. The first is that the required assumptions will almost certainly not hold in cases of normative uncertainty. Some decisions, like deciding whether to launch a nuclear weapon, are more momentous than others, like deciding between paper and plastic bags, and so the "same mean" and "same variance" assumptions won't hold. What's more, independence is highly implausible in the context of normative uncertainty. Only the most insanely particularistic of particularists is going to be such that normative "outcomes" – for example, it's being okay to turn the trolley in such-and-such a case, it's being better to eat organic buckwheat than non-organic buckwheat – are independent for him.[53]

There are two responses to this objection. The first is that the traditional SLLN is only one member of a class of so-called "Convergence Principles" that all guarantee actual value maximization over infinite trials of EOV maximization. Other principles allow for various weakenings of the independence and trial-similarity assumptions.[54]

[52] See Buchak (ms #2), Allais (1953), and Putnam (1986), for philosophical discussions, and Sen and Singer (1993), just to take one text at random, for a mathematical treatment.

[53] Thanks to Ruth Chang and Tim Maudlin for pressing me on this issue.

[54] Some of these Convergence Principles are explained in Sen and Singer (1993), Chapter 2, among other places. Whether there are convergence principles weak enough to guarantee actual value maximization in sequences of decisions under normative

The other response is that the result guaranteed by the SLLN may be indirectly relevant to the rationality of EOV maximization even in those cases where its assumptions fail to hold. Consider someone whose decisions are independent – "the most particularistic of particularists" – and who faces nothing but quantitatively similar decisions. Suppose we conclude, on the basis of the standard argument, that it's most rational of him to maximize EOV in some situation, S. Now imagine that an ordinary person like you or me confronts situation S. Is there any difference between you or me and this imaginary character that justifies a different verdict about what it's rational to do? I don't see that there is. Why should the particularist maximize EOV, while the generalist should not? Why should the person who sees similarly momentous cases over and over again maximize EOV in S,[55] while a person leading a more typical life should not? We cannot say, "This is because the imaginary character is guaranteed to maximize actual value by maximizing EOV over and over again, while ordinary folks are not." For it is no more plausible to suppose that the imaginary character will make infinite decisions than it is to suppose that we will. In each case, we're evaluating only the behavior in a single situation of an agent who will face only a finite number of situations. Given that, the standard argument seems to lend some support to the rationality of EOV maximization even for those whose circumstances don't match its assumptions.

At this point, I expect to hear the second objection: "Putting aside worries about whether the aforementioned assumptions apply to me, why is it relevant to the rationality

uncertainty like the ones we typically face is an extremely difficult question to which I have no answer at the moment.

[55] I should mention: It's well-established in statistics that there are SLLN-like Convergence Principles for so-called "Martingale" and "Reverse Martingale" sequences, which involve decisions that are – to put it loosely – progressively more and more, and less and less momentous, respectively. See Sen and Singer (1993).

of my following a rule in a single case what would happen if I followed that rule in an infinite number of cases?"

It's possible to read this objection as expressing a general skepticism about the relevance of normative judgments about some targets to the propriety of normative judgments about other targets – a flatfooted insistence that merely because infinite sequences of actions are one thing, and single actions or finite sequences another, what we say about the former ought to have no bearing on what we say about the latter. Such a skepticism is clearly misguided. Normative philosophy would grind to a halt if we couldn't take, for example, the propriety of blaming John for deceiving his friends as evidence that John oughtn't to have deceived his friends, or the rationality of some choice of a political system behind a Veil of Ignorance as evidence that that political system is the just one. There's nothing more irritating than the person who dismisses Rawls based on the mere fact that we're not behind a Veil of Ignorance but are rather "concrete individuals" with "particular attachments and commitments", etc., etc., etc. His counterpart in the present context is the skeptic who claims that the mere fact of our finitude is reason enough to dismiss the relevance of infinite sequences of actions.

But this is not what's lurking behind the objection as it'd be posed by most people. They don't think that, as a general matter, normative judgments about certain targets are irrelevant to the judgments we ought to make about other targets. Rather, they just think judgments about infinite sequences of actions, in particular, are irrelevant to assessing the behavior of finite individuals.

We might try to get the objector to give up his worry with prompts like: "Well, suppose you did follow a rule other than EOV maximization in a single case. Wouldn't it be strange if, in order to avoid a surely suboptimal total objective value over the long term, you had to give up this principle eventually? In other words, wouldn't it be weird if a principle of rationality had a “shelf life”, after which it had to be tossed in favor of EOV maximization?” Now, my intuition that EOV maximization is right actually is strengthened by little prompts like this, but this response is not universal.

Given that we seem to be at an impasse, it's worth exploring two variations on the standard argument that can be understood without much technical apparatus. The first differs from the standard argument in virtue of the “dimension” along which the trials are distributed. Rather than imagining the trials as acts that I do over an infinitely long life, we might imagine the trials as acts that the members of an infinitely large population do. That we're now distributing acts over different people rather than different time slices is mathematically irrelevant. The total value of the actions performed by all of the members of the population is guaranteed to be higher if they all maximize EOV than if they all act in accordance with any other rule.[56] And this should lend plausibility to the claim that it's rational for any single member of the population to maximize EOV. If the first argument was the “long run value” argument, then we might call this the “wide run value” argument.

This variation allows us to appeal to a new, and potentially powerful, intuition. This is the “What if everybody did that?” intuition: If everyone's acting in accordance with one norm compares favorably with everyone's acting in accordance with another norm, this is a reason to believe that the first norm is preferable to the second. In the present case, everyone's acting in accordance with EOV maximization compares

[56] This assumes that the EOV they're maximizing is relative to the assessor's credences.

favorably with everyone's acting in accordance with any other norm, in the specific sense that there will be more practical value in the world if we're all EOV maximizers.

Of course, it's possible to deflate “What if everybody did that?” intuitions in cases where the value of my action is thought to depend on whether other people actually do “that”, and where, as a matter of fact, they don't, or won't, do “that”. If everyone resorted to vigilantism, civil society would collapse. But it's reasonable to think that the disvalue of my resorting to vigilantism in a single case depends substantially on whether this will precipitate the collapse of civil society. Since most people won't resort to vigilantism, my doing so may not be so terrible after all.

You might resist such a deflationary effort, and say in cases like the one above that it doesn't matter if everyone actually does the act. You'd be in good company – specifically, Kant's company.[57] But whatever the merits of the effort in some cases, it clearly won't work in all cases. For EOV maximization is most obviously certain to yield the greatest objective value specifically in cases where the trials are independent – where the values of my actions do not depend on the values of your actions. Of course, the total value of all of our actions depends on what each of us does, but our actions may be mutually independent, and we still get the result I'm talking about.

So that's one variation, and a reason to think that variation adds additional strength to the case for EOV maximization. There's another variation worth discussing. As I mentioned before, many people are reluctant to draw conclusions from examples involving infinities. I may perform a lot of actions in my lifetime, but not an infinity of actions; there may be a lot of agents in the world, but not an infinity of agents; so why

[57] Kant (1785), Part II.

should those of us in the real world take our cues about what to do from these examples? We can appease doubters of this sort, though, for there are important conclusions to be drawn from cases involving lots of situations, even if they don't involve infinite situations. For the same reason that it's guaranteed that maximizing EOV over infinite trials will yield the highest total objective value, it's almost guaranteed that maximizing EOV over lots of trials will yield the highest total objective value. So in addition to the “long run value” and “wide run value”, arguments, we now have the resources for what we might call the “medium-long run value” and “medium-wide run value” arguments. (See Table 2.1)

Table 2.1

	Actions distributed over time	Actions distributed over agents
Infinite actions	“Long run”	“Wide run”
Finite actions	“Medium-long run”	“Medium-wide run”

“Interesting result,” you might say, “But as any salesman will tell you, there's a difference between a guarantee and an almost guarantee. If you could guarantee that EOV maximization would yield the highest value over the medium-long run or medium-wide run, then this would be a point in its favor. But an almost guarantee? No – there are some people who are very, very, very probability-weighting, value-weighting, and/or risk-sensitive with respect to medium-long- or medium-wide-run value, and they would opt for another strategy instead of EOV maximization if all you could give them was an almost guarantee.”

My response to the objector is simply that these people and their theories are crazy. Recall the very beginning of our discussion about the different comparativist theories of rationality. I said that certain of them were so extreme in their probability- or value-weighting, or their risk-sensitivity, that they should be dismissed out of hand. I didn't pursue this point at length, though, because there are plenty of non-extreme comparativist theories of rationality that deserved serious responses, and better to train our attention on the strongest opponents. But now we have occasion to address these extreme theories in new, mutated forms – not as theories about how to tote up the possible values of single actions (i.e. as theories of rationality) but as theories about how to tote up the values of the actions that make up a temporally extended life, or that compose the collective behavior of a population. As we add more and more actions, and their total value inches closer and closer to the number of actions multiplied by the actions' average EOV's, more and more theories of how to tote up the values of multiple actions will join the chorus of those that support EOV maximization in each instance over any other strategy in each instance. Only the most extreme theories will hold out – those that say, for example, that the slightest chance of an unfortunate statistical aberration (e.g. I get the worst possible normative "outcome" every time I do the EOV-maximizing act) should compel us to consistently choose an extremely risk-averse strategy instead.

So the medium-long and medium-wide run arguments are dialectically very effective after all. For all of the non-crazy theories of rationality other than EOV (mild risk-aversion in single cases, say) their multiple-action analogues (mild risk-aversion over lots and lots of cases) favor a consistent strategy of EOV maximization rather than a consistent strategy of themselves. In other words, when it comes to medium-long or medium-wide run value, a near-guarantee of optimal aggregate value is going to garner the support of every reasonable theory thereof. Single-case theories of rationality other than EOV-maximization are not self-defeating, exactly, but we might say that they're defeated by their analogues. Not so for the crazy theories of rationality; their analogues don't support EOV maximization.[58] But so as not to attack straw men, we eliminated those theories at the outset.

In summary, the "long-run value" argument can get off the ground with Convergence Principles much weaker than the SLLN, and so it doesn't rely on assumptions about the structure of act-sequences that can't possibly hold true. Even if it did rely on such assumptions, there's no good reason why someone whose situation matched those assumptions should maximize EOV, while someone whose situation didn't match them should do something else. Putting all of that aside, someone might be skeptical about the relevance of judgments about infinite act-sequences to the evaluation of finite agents. But by considering cases where the acts are spread out over members of a population rather than over time-slices of a single Methuselan agent, we can exploit the strong intuition that I ought to do something only if it would be desirable if everyone did it. And by considering cases involving very many, but not infinite acts, we can

[58] In seeking to undercut the medium-long run argument for expected value maximization, Lara Buchak makes the point that, just as the view that one ought to maximize EOV over the medium-long run supports EOV maximization in single cases, the view that one ought to Maximin over the medium-long run supports Maximining in single cases. See Buchak (ms #2), p. 7. Quite true. But again, this property of being supported by its multiple-case analogue is not one that any reasonable theory of single-case rationality other than EOV maximization shares with Maximin, so Buchak's argument doesn't generalize. And it doesn't matter so much that Maximin is endorsed by its analogue, because, again, it's a crazy theory of rationality (except in cases where we can't say anything about probabilities other than that they all fall in the zero-to-1 interval, and maybe not even then).

circumvent worries from those who are concerned particularly about passing over the boundary between the infinite and the finite.

Highest Winning Percentage

In this final section, I aim to show that the action with the highest EOV thereby has another heretofore unnoticed but very attractive feature: it has the highest Winning Percentage. This fact, I shall argue, constitutes another reason to regard higher-EOV actions as more rational than lower-EOV ones. An action's Winning Percentage is its probability of being objectively better than an action selected at random. In what follows, I will prove the connection between EOV and Winning Percentage, and explain why we should prefer actions with higher Winning Percentages.

Before we get into anything formal, I want to give you an intuitive feel for what Winning Percentage is. You may recognize the term "Winning Percentage" from the world of sports. In that context, it refers to the number we get by dividing an individual or team's number of wins by that individual or team's number of total games or matches played over a period. If a football team wins 7 games and loses 9 over a 16-game season, then that team's Winning Percentage is approximately .44. If a tennis player wins 60 matches and loses 7 during a calendar year, then that player's Winning Percentage is slightly less than .9.

This conception of Winning Percentage is not quite the one I have in mind. For this sort of Winning Percentage measures one's success against actual opponents, not all possible opponents. That's one reason why it's a crude indicator of an individual or team's ability. If, for example, a college football team plays a very easy schedule, full of opponents like Temple University and the University of Buffalo, its actual Winning Percentage may well be higher than that of a team with a difficult schedule, made up of games against the University of Southern California and Louisiana State University. And if someone were to seriously claim that the first team is better than the second on the grounds that its Winning Percentage is higher, it would be natural for the second team to reply, "Look, if they played our schedule, they'd win fewer games than we did." (Or, "If we played their schedule, we'd win more games than they did.") In other words, one would respond by adverting to what one suspects the teams' Winning Percentages would be against merely possible opponents.

Of course, there are good reasons in sports for not using Winning Percentage against merely possible opponents. Some of these reasons are even philosophically interesting (e.g. that success against actual opponents is partly constitutive of, rather than merely evidence of, one's athletic talent). But in assessing actions, I shall be concerned with their Winning Percentages against all possible actions – although "all possible actions" will need to be understood as having the particular sense I'll soon explain.

Now for the argument proper. I'll begin by showing that the objective value – not the expected objective value for now, but just the objective value – of an action is a positive linear function of that action's probability of being better than a randomly selected action.

We'll need to start by clarifying some key notions. In saying that an action is selected at random, one prompts the question, "From what set of actions?". The most tempting answer is, "From all possible actions – not just the ones that are possible in this situation, but all of the actions that could occur." But that answer is not quite what I'm looking for. For suppose that the values of the possible actions are not spread over the real numbers with uniform density. It may turn out that there are lots and lots of possible actions with values between, say, zero and 1000, and very few possible actions with values between 1000 and 2000. (See Diagram 2.1)

Diagram 2.1: Non-Uniform Action-Density of Values

VALUES

Zero 1000 2000

ACTIONS

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

In that case, an action with a value of 2000 would be twice as good as an action with a value of 1000, but its probability of being better than a randomly selected action would not, it seems, be twice as high.

The problem in the previous paragraph wasn't with the answer to the question. It was with the phrase that prompted the question: "randomly selected action". For what I really have in mind is not an action selected at random from any set of possible actions, but rather a value selected at random, whether or not the randomly selected value corresponds to an action that's truly possible. By a "randomly selected action", then, I shall mean "an action with a value randomly selected from among the real numbers". The real numbers are, of course, spread over the real numbers with uniform density. I will continue, however, to use the snappier "randomly selected action" rather than the more cumbersome “action with a value randomly selected from among the real numbers”.

But now there is another problem with the formulation. The interval containing the real numbers has neither an upper nor a lower bound. That is, the size of this interval is infinite. So the probability of an action with a value of V being better than a randomly selected action will turn out to be (V - -∞) / (∞ - -∞), which is undefined in standard analysis. This is obviously unsatisfactory. We want to somehow express the idea that the higher an action’s value, the more likely it is to be better than a randomly selected action. So we’ll have to use bounded intervals. An action A’s Winning Percentage, then, will be a ratio whose denominator is the size of some bounded interval and whose numerator is the size of the interval whose upper bound is A’s value, and whose lower bound is the lower bound of the interval in the denominator.

But where shall we set the bounds? Suppose that both the upper and lower bounds of the interval are less than A’s value. Then the probability of A being better than an action with a value randomly selected from that interval will be 1.0. But there will also be actions with lower values than A whose probabilities of being better than an action with a value randomly selected from that same interval will likewise be 1.0. So the upper bound of the interval must be at least as great as the value of A – or for that matter, any other action under consideration.

Mutatis mutandis for the “other direction”, as it were. Suppose that both the upper and lower bounds of the interval are greater than A’s value. Then the probability of A being better than an action with a value randomly selected from that interval will be zero. But there will also be actions with higher values than A whose probabilities of being better than an action with a value randomly selected from that interval will also be zero.

So the lower bound of the interval must be at least as great as the value of A – or for that matter, any other action under consideration.

With all of this in mind, we can sharpen our claim as follows:

> The objective value, V, of an action, A, is a positive linear function of A’s probability of being better than another action with a value randomly selected from the real numbers within any interval of non-zero size, the upper bound of which is at least as great as A's value, and the lower bound of which is no greater than A's value.

It’s not hard to show that this is true. Pick any bounded interval that satisfies these requirements; call the lower bound L and the upper bound U. With values of possible actions distributed uniformly over this interval, the probability of an action with a value of V being better than a randomly selected action from this interval will be:

$(V-L) / (U-L)$

$= V/ (U-L) - L/ (U-L)$

$= V(1/(U-L)) - L/(U-L)$

Substituting “a” for 1/(U-L) and “b” for -L/(U-L), we get the standard form of a linear function:

Winning Percentage = **a**V + **b**

Since the size of the interval is non-zero, U-L is positive, so 1/(U-L) is positive, so **a** is positive, which makes f(V) =aV + b a positive linear function. Of course, the values of **a** and **b** will depend on the values of U and L, so the linear function will be different for different intervals. But given some particular interval, these values will be constant. Winning Percentage is a probability, which means it should conform to the axioms, and in particular, to the Normalization axiom that probabilities can be no less than zero, and no greater than 1. Luckily, this axiom is obeyed here. Since V can be no greater than U, (V-L)/(U-L) can be no greater than 1. Since V can be no less than L, (V-L) can be no less than zero, so (V-L)/(U-L) can be no less than zero.

Now for the next stage of the argument. I'm going use the result above in showing that the action with the highest EOV has the highest (what I'll call) Expected Winning Percentage. Expected Winning Percentage is the probability-weighted sum of an action's possible Winning Percentages. It is calculated as follows:

$\Sigma_i\ p(S_i) \cdot (\text{Winning Percentage given } S_i)$

This should remind you a bit of the formula for expected value. This is no coincidence. Expected Winning Percentage is to actual Winning Percentage as EOV is to Objective Value.

Since we know that Winning Percentage = **a**V + **b**, we can substitute accordingly:

$\Sigma_i\ p(S_i) \cdot (a(V \text{ given } S_i) + b)$

Then we distribute p(Si):

$$\Sigma_i \, (p(S_i) \cdot (a(V \text{ given } S_i) + (p(S_i) \text{ x } b)$$

$$= \Sigma_i \, p(S_i) \cdot (a(v(A \text{ given } S_i))) + \Sigma_i \, p(S_i) \cdot b$$

$$= a \, (\Sigma_i \, p(S_i) \cdot (V \text{ given } S_i)) + \Sigma_i \, p(S_i) \cdot b$$

$\Sigma_i \, p(S_i) \cdot (V \text{ given } S_i)$ is just EOV, and $\Sigma_i \, p(S_i) \cdot b = b$, so we can substitute to get:

a(EOV) + **b**

Since **a**, once again, is positive, we get the result that Expected Winning Percentage is a positive linear function of EOV. By consequence, the higher the EOV, the higher the Expected Winning Percentage.

Now for the final step. Expected Winning Percentage is a type of Winning Percentage. Earlier, we showed that the action with the highest objective value has the highest Winning Percentage. But in doing so, we underspecified things a bit. For what we really demonstrated is that the action with the highest objective value has the highest Winning Percentage, relative to its actual value. But when we act under normative uncertainty, we're not privy to actions' actual values; we must make our decisions based on our credence distributions over a set of possible values. Just as there are notions of "better than" and "ought" that are belief-relative, there is a kind of Winning Percentage that is belief-relative – or, more specifically, relative to the agent's credence distribution over possible values of an action. And this kind of Winning Percentage is Expected Winning Percentage relative to actual value.

What's going on here is a sort of "nested gamble" – a gamble involving two "levels" of probability. At the first level, there's the probability that the action's actual value will exceed the value of a randomly-selected action. And at the second level, there are the agent's credences in the action's having various actual values. We get the total probability of some outcome of a nested gamble by computing the expected first-level probability, just as we did above.

This brings us to the question, "Why introduce this notion of Winning Percentage at all?"

The first-pass answer is metaphysical. We might think of the debate between proponents of EOV maximization and proponents of other comparativist theories as a debate about how degrees of belief and degrees of value should be "combined", if you will, to yield the rational values of actions. The proponent of EOV maximization thinks they should be combined via one function; his opponents think they should be combined via another. It is difficult to gain any traction in this debate, because degrees of belief and the degrees of value that are their objects seem like such totally different entities. Many debates are like this. Consider the almost exactly parallel debate between, say, utilitarians and prioritarians in population ethics/welfare economics. This is a debate about how to combine two factors – levels of well-being, and numbers of people, to yield an optimum distribution of welfare. And it, too, is a debate in which it's often difficult to come up with decisive arguments.

The purpose of introducing Winning Percentage is to circumvent the need to discuss one of these factors – in this case, degrees of value. Since an action's EOV is linearly correlated with Winning Percentage (relative to one's credences distribution over objective value rankings), we can stop talking about the two determinants of EOV, and just talk about probabilities of favorable events – in this case, probabilities of actions under consideration being better than randomly selected actions. If it is a bit of a mystery how probabilities and value should go together in determining which action to perform under uncertainty, it is less of a mystery how the probability of some favorable result should determine which action to perform. It seems to me an attractive feature of the action with the highest EOV that it is most likely to be better than a randomly selected action – that it'll "win" in more conceivable situations than an action with the a lower EOV.[59]

Let me close by saying a bit more directly why having the highest Winning

[59] Is there an analogous maneuver we can make in population theory? I think this may be an example: Suppose we are deciding which distribution of length-of-life over people is best – 1) John living a life of some quality for 100 years and Mary living a life of the same quality for 60 years, or 2) John living a life of that quality for 75 years and Mary living a life of that quality for 75 years. Again, it can seem difficult to prosecute this debate, in large part because we're combining two different factors – years of life, and people whose years they are. But we might, loosely following Parfit (1984), think of things this way: John and Mary are collections of person-slices, each of whom persists for, let us say, 1 year. So while there may have been a concern about equality or priority between John and Mary, there is surely none between the person-slices, each of whom enjoys the same length of life. Since the first scenario involves 160 person slices getting a year each, and the second involves only 150 person slices getting a year each, the first distribution is better. Now, of course that kind of strategy is controversial. First, it's plausible that there's a normative difference between a person living for X years, and a set of X disconnected person-slices, laid out temporally end-to-end, living one year each. And second, it's not totally obvious that it's better for there to be 160 person slices living a year each rather than 150 living a year each, just as it's not obvious that possible world in which there are more people thereby has more value than an otherwise similar possible world in which there are fewer people.

Percentage is important. Winning Percentage marks a sort of normative modal robustness. The higher an action's Winning Percentage, the better it fares not just in the actual world, but across all conceivable worlds, against actions with all conceivable values, in all conceivable situations. To select the action with the highest Winning Percentage, then, is to select that action that compares favorably to more of the actions that you've done in other situations, that others have done in other situations, that you could imagine yourself going in other situations, and so forth. The difference between an action with a higher Winning Percentage and an action with a lower Winning Percentage that ends up being objectively better in the actual world is similar to the difference between an object with some dispositional property, and an object that lacks that dispositional property but behaves in the actual world as though it manifested such a property. To see whether an object is actually fragile, for example, we cannot simply look at its behavior in the actual world; we must see how it behaves across a range of possible worlds. Otherwise, it would be end up being a matter of luck which objects were fragile. Faberge eggs that didn't break would not get counted as fragile, while solid steel crowbars that did break would get counted as fragile. But fragility should not depend on luck in this way. Similarly, to see whether an action is actually rational we should see how it fares against actions across possible worlds. Otherwise, it would end up being a matter of luck which actions were rational. Actions that fared worse against the entire field of conceivable actions would get counted as more choiceworthy in the belief-relative sense than actions that fared better against this field. But if fragility should not depend on luck, rationality should certainly not depend on luck.

Parity, Incomparability, and Indeterminate EOV

I've been arguing that it's most rational to do the action with the highest EOV. But sometimes there won't be such an action. This will happen in at least some of the cases where EOV is indeterminate. (In other cases of indeterminate EOV, the indeterminacy will concern sub-optimal actions.) When EOV is indeterminate, actions are either rationally on a par or else rationally incomparable, depending on whether the cause of the indeterminacy is the agent's credence that actions are objectively on a par, or her credence that actions are objectively incomparable. In this section, we'll see how credence in objective parity and incomparability can give rise to rational parity and incomparability, and I'll say something about action in the face of each of the latter two.

One explanation for the rational parity of two actions is an agent's high-enough credence that those two actions are objectively on a par. For example, if John's credence is .1 that A is better than B, .1 that B is better than A, .1 that the two are equal, and .7 that they are on a par, then it may turn out that they are rationally on a par as well. The same goes for incomparability: If John's credence is .7 that A and B are incomparable, then they may turn out to be rationally incomparable as well.

But this is not the only route to rational parity and incomparability. If we have only ordinal rankings of actions, then parity and incomparability show up only as relations that A and B might bear to one another. But once we introduce cardinal rankings, parity and incomparability may rear their heads again with regard to A and B, as features that generate imprecise cardinal rankings. For imagine that A is better than B. A might, in that case, be on a par or incomparable with several other actions, C, D, E, each of which is better than the last, and B might be on a par or incomparable with several other lower-ranked actions, F, G, H, each of which is better than the last. (See Diagram 2.2)

Diagram 2.2: Parity- or Incomparability-Induced Cardinal Indeterminacy

C – parity or incomparability with A, better than B
D – parity or incomparability with A, better than B
E – parity or incomparability with A, better than B
⋮
F – parity or incomparability with B, worse than A
G – parity or incomparability with B, worse than A
H – parity or incomparability with B, worse than A

If that's the case, then this value difference will lack a determinate size. It will be no smaller than the difference between E and F, and no larger than the difference between C and H, but that's all we can say. The same sort of indeterminacy may happen, of course, on the hypothesis that B is better than A. With indeterminacy on all sides, it will end up being indeterminate how the difference between A and B if A is better compares to the difference between B and A if B is better. It may be, e.g., that the first difference isn't exactly 4 times the second, but is instead between 3.5 and 4.5 times the second.

This doesn't mean, of course, that any chance of parity or incomparability between A and B, or parity- or incomparability-induced cardinal indeterminacy, leads to A and B's being rationally on a par or incomparable. This will only happen some of the time. The probabilities and associated value differences may work out so that A's expected value is, say, higher than B's. But it will be higher by an indeterminate amount if there's any chance of parity or incomparability.

How do rational parity and rational incomparability differ from rational equality, in practical terms? This depends upon one's theory of practical reason. You may have the view that parity and incomparability represent “reasons running out”, and so one must enter a second stage of deliberation to resolve things.[60] By contrast, one simply acts with indifference when faced with equality. A more obvious difference, though, involves the value of gathering new information that is relevant to the values of actions. As we'll see in the next chapter, the gross EOV of normative deliberation is almost always positive. I show this with the help of a proof given by I.J. Good, and clarified and expanded upon by Brian Skyrms. So if two actions are equal, then normative deliberation will always make sense provided that the EOV cost is lower than the EOV gain from deliberation. But if two actions are on a par or incomparable, then even if the EOV gain exceeds the EOV cost, the net EOV may not be enough to break the parity or incomparability between A and B. It's helpful here to think of cases like the Mozart/Michelangelo one: If the two are equal, then a minor improvement to one makes that one better. But if the two are on a par, say, then a minor improvement may preserve parity. So if two actions are rationally on a par, it may not, for the purposes of decision-making, be worth engaging even in an EOV-positive bit of normative deliberation. The same, again, goes for incomparability.

We needn't forgo EOV maximization as the correct theory of rationality simply because there's a possibility of rational parity and incomparability. That there won't always be a highest-EOV action in no way disparages the view that, when there is one, we should do it. But we do need to be cognizant of the two ways in which rational parity and incomparability may come about. The two perhaps-not-as-obvious points in this section, then, are: 1) The possibility that A and B are on a par or incomparable should not be treated, for purposes of EOV calculation, just like the possibility that A and B are

[60] This view is defended in Chang (2009).

equal. But it's equally a mistake to assume that the possibility of parity or incomparability automatically makes A and B rationally on a par or incomparable, and 2) There is a route to rational parity and incomparability that doesn't depend on the possibility of A and B being objectively on a par or incomparable; instead, it depends on parity- and incomparability-induced cardinal fuzziness.

CHAPTER THREE: CHALLENGES TO EXPECTED VALUE MAXIMIZATION

Introduction

In this chapter, I'll consider some challenges to my view of rationality under normative uncertainty. The previous chapter consisted of positive arguments in favor of the view; in the present chapter, I try to show how some arguments against the view are mistaken.

A few preliminary remarks about the arguments to come: First, you will notice that some of the objections have targets other than just EOV maximization. Some are objections to comparativist theories of rationality in general and some are worries about the very project of developing a theory of rationality under normative uncertainty. But since I'm defending both this specific theory and one of this family of theories, and, of course, engaging in this project, these are all objections I must address.

Second, I should say that the two most important objections are not represented here. The first of these is what I'd called the Problem of Value Difference Comparisons – that it may be impossible to compare differences in value between actions across different normative hypotheses. If such comparisons are impossible, then EOV maximization is impossible. I'll address this problem in Chapters 4 and 5.

The second of these is Problem of Uncertainty about Rationality: One might be uncertain not only about first-order normative hypotheses, but also about theories of what it's rational to do under such uncertainty. One might develop a theory of what it's rational to do under uncertainty of the latter sort, but of course, it's possible to be uncertain about that, too, and so on. I'll identify the problems to which this possibility gives rise, and attempt to solve them, in Chapter 7.

Why treat these problems separately from the rest? For one thing, I'll have to devote a lot of space to solving them, and it seems better to have several 40-or-so page chapters than one 180-page chapter. For another, these problems are fundamental in a way that the others aren't – fundamental in a sense that grappling with them will teach us important lessons about the structures of normative theories and other hypotheses (in the case of the Problem of Value Difference Comparisons), and about the nature of rule-following and rationality (in the case of the Problem of Uncertainty about Rationality). Better, then, not to lump them in with objections like the ones in this chapter, which are more easily resolved.

Demandingness

The first objection is that EOV maximization is too demanding a theory of rationality under normative uncertainty. Here's how the charge might be put:

> "Look, I believe murder is really, really bad. If I have any credence at all that abortion is as bad as murder, that not giving to UNICEF is as bad as murder, that killing a fly is as bad as murder, that non-procreative sex is as bad as murder, and so on, then all of these things may have less-than-maximal EOV's, and will therefore be rationally forbidden by your theory. But it's clearly permissible for me to do all of these things, so your theory is too demanding."[61]

[61] This objection has been pressed by Weatherson (2002), and in conversation by

Now, at least as stated, this objection seems to get the expected values wrong for most agents. My credence is low enough in non-procreative sex being as bad as murder that the EOV of non-procreative sex is probably quite often higher than any alternative. But that's a very superficial flaw in the objection. The real problem with this objection is that it fails to find a target in my theory as presented so far. For I've not said that one is rationally required to do the action with the highest EOV, or that one is rationally forbidden to do any action with a less-than-optimal EOV. I've argued only that it's more rational to perform actions with higher EOV's than it is to perform actions with lower EOV's. In other words, the demandingness objection is about the statuses of actions, while my theory so far says only the action with the highest EOV enjoys the highest ranking.

The focus on statuses rather than rankings is a feature of demandingness arguments more generally. Consider: Utilitarianism is often cited as an overly demanding moral theory for, to give one example, forbidding a person from giving to a very effective charity on the grounds that this promotes less utility than giving to a slightly more effective charity. But the objection here has nothing to do with the rankings of the two

Ruth Chang, Richard Chappell, and Govind Persad. It's also taken up by Lockhart (2000), p. 109, whose solution is to impute to agents a substantial degree of belief in an egoistic normative hypothesis according to which sacrificing one's own interests is much, much worse than doing or allowing harm to other beings' interests. The upshot of this, of course, is that giving to UNICEF, killing a fly, and so on, will frequently have lower EOV's than their alternatives, even if there's some chance that they're as bad as murder. But Lockhart's response strikes me as unsatisfactory, since it is unresponsive to what I regard as the stronger form of the demandingness objection. According to this form of the objection, EOV maximization places excessive demands even on agents whose credences in hypotheses like egoism are insufficient to ever render self-sacrificial actions the ones with the highest EOV's. This problem with Lockhart's response illustrates a more general point: One cannot buttress an account of what it's rational for agents to do, given their credences, by imagining agents to have a particular set of credences.

actions. Surely it is better to give to a more effective charity than to a less effective one. (Why else would it make sense to do research into the effectiveness of charities?) Rather, the objection is that giving to the less effective charity is, contra classical utilitarianism, not forbidden. Utilitarianism's alleged mistake lies in its purported mis-assignment of status, not its mis-ranking.[62]

This is not to say that I'm unconcerned with rational statuses. Indeed, I'll discuss them at some length in Chapter 6. But nothing I say on this score should trouble the exponent of the demandingness objection, either. For all I say about statuses, it may be rationally permitted to do an action with a much higher EOV rather than an action with a much lower EOV. In summary, the demandingness objection is not well-put against the main part of my theory, which concerns rational rankings only, and nothing I do say about statuses renders that part of my theory overly demanding, either.

Staying True to Yourself

On the theory of rationality that I'm proposing, there's no ruling out cases in which I have an extraordinarily high credence in normative Hypothesis 1 and an extraordinarily low credence in Hypothesis 2, but am nonetheless more rational in doing what's better according to Hypothesis 2 than what's better according to Hypothesis 1. There's an argument that EOV maximization's tendency to deliver results like this is, at the limits at least, a disadvantage.

[62] And indeed, there are non-standard versions of utilitarianism that rank actions just as classical utilitarianism does, but do not assign the status "forbidden" to all sub-optimal actions. See Howard-Snyder (1994) and Norcross (1997).

The argument is that there’s a kind of rational value that inheres in “staying true to yourself”.[63] On a not-implausible construal of the notion, staying true to oneself in the context of normative uncertainty requires you to attach disproportionate weight to higher credences. So, for example, a credence of X in a hypothesis according to which the value of my action is Y contributes more rational value than a credence of X/4 in a hypothesis according to which the value of my action is 4Y. The thought, perhaps, is that fidelity to one's credences is a more important part of fidelity to oneself than is fidelity to the values of actions if those credences are true. EOV maximization fails to account for this, since it treats credences and values equally by leaving them both “unweighted”, in the terminology of the last chapter.

This argument may strike you as more convincing the more extreme the disparity in credence is between hypotheses. If my credence is .55 that killing 1 person is better than letting 5 die, and .45 that letting 5 die is better than killing 1, it sounds silly to say that my identity is tied to the former hypothesis – that I’m somehow betraying myself by acting in accordance with a hypotheses that I believe to degree .45. By contrast, if the credences are .99 and .01, respectively, then it may seem, perhaps, that I should essentially treat the first hypothesis as true for the purposes of decision-making.

With that said, I have trouble understanding the appeal of this argument. I'm not suggesting that people who are close to certain regarding some normative matters should second-guess themselves, or not act with conviction, or sell out their deepest values for money, or just stay home and watch television until the uncertainty vanishes completely. That's not what the sort of “normative hedging” counseled by EOV maximization is

[63] This objection is due to Geoff Anders.

about. If anything, the theory of rationality I'm proposing explains why irresolution, sloth, and selling-out are irrational. If, according to the hypotheses in which you have credence, these behaviors rank poorly in the objective sense, then EOV maximization will ensure that they rank poorly in the rational sense, too.

I also have the suspicion that this objection is based on a subtle mistake about how EOV maximization under normative uncertainty works. That some people make this mistake is suggested by the way they phrase their questions about my approach, at least as it's applied to uncertainty about well-known moral theories. There is a tendency for people to begin, “So suppose you're uncertain between Kant's moral theory and Bentham's. Now, Kant says...and Bentham says...”, which suggests that EOV maximization would take as inputs my credences in the two philosophers' theories, and the values of actions according to those theories *according to those philosophers*. It's as though by having some credence in Kant's theory, I've signed a blank check, and Kant gets to fill in the amount. If that's how EOV maximization works, then it's not crazy to think that it is inconsistent with staying true to oneself. For on this understanding, my credences in hypotheses are part of me, and the values according to those hypotheses are not; they're whatever The Great Philosophers say they are. So of course credences should be weighted more than values in determining what it's rational for me to do.

But again, this is all a misunderstanding. When I have some credence in utilitarianism, the values of actions if that credence is true depend on utilitarianism *as it's represented by me*. The works of Jeremy Bentham may have some causal influence on these values, insofar as I've read and admired Bentham, but they play no constitutive role whatsoever. Degrees of value represented in hypotheses, then, are just as much a part of me as my credences in those hypotheses. Given that, the motivation for credence-weighting vanishes, and we see that EOV maximization is not at all at odds with staying true to oneself. Indeed, maximizing EOV just is staying true to oneself, on the correct understanding of how the respective inputs to EOV relate to one's self.

The Value of a Life versus the Values of Actions

There's also room to object to EOV maximization not on the grounds that any single EOV-maximizing action, taken in isolation, is a mistake, but rather because there may be something deficient about a life consisting of a string of such actions. Before we get to the argument for that, here are some examples that may motivate us to sever our evaluations of actions from our evaluations of lives consisting of those actions: 1) The thought of breaking out a six-pack of Corona beer, laying out on a beach chair, and listening to Eddie Money's Greatest Hits strikes me as very appealing. The thought of doing nothing but that for my entire life strikes me as pathetic. 2) If a stranger on the train told me that he was training to become a monk, I'd think this was a legitimate life choice. If another stranger told me that he was training to become a hip-hop dancer, I'd think the same. If the very same stranger told me on one day that he'd started monastic training, and then a week later told me that he'd given that up and started training to be a hip-hop dancer, I'd think he was a lost soul. In the first example, we find the overall contour of the person's life to be deficient on the grounds that it contains too much of the same thing. In the second example, we find the overall contour of the person's life to be deficient on the grounds that he leaps from one serious pursuit to another serious pursuit of an entirely different type. And these assessments do not depend on our assessments of the actions taken in isolation. There's nothing wrong with listening to Eddie Money, training to be monk, or training to be a hip-hop dancer.

Could following a strategy of EOV maximization result in one's leading a life whose contour is similarly skewed? Ruth Chang has suggested to me that it might. It will help to see this to contrast EOV maximization with Credence Pluralitarianism (CPlur) from the previous chapter. Suppose that your credence in Kantianism is .7, and your credence in utilitarianism is .3. If you live your life in accordance with CPlur, you will always do what Kantianism recommends, assuming your credences stay the same. If you act in accordance with EOV maximization, it may turn out that you sometimes act in accordance with Kantianism and sometimes with utilitarianism, for the "stakes" of the situation may be different for different theories on different occasions. With respect to these normative theories at least, your life will take on the air of inconsistency – of doing the "Kantian thing" sometimes and the "utilitarian thing" at other times. As Chang put it in conversation, your life will appear not to "make sense" or not to be characterizable by a "coherent narrative". And this failure of sense and/or narrativity may count as a strike against the overall merit of the EOV maximizer's life, and by consequence, a strike against EOV maximization.[64]

It may help to diffuse Chang's objection to consider how we might explain away the tension that arises in the "Eddie Money" and "monk/dancer" cases – of behavior being laudable in isolation but less so in constituting a life. In thinking about both cases,

[64] This type of objection is reminiscent of some views expressed by Alastair MacIntyre (1981) and Charles Taylor (1989), but neither considered the implications of his theory for action under uncertainty. See Strawson (2004) for objections to MacIntyre and Taylor.

we shall want to note that the evaluative properties of an action at a time can depend on facts about other times. We are not, for instance, stuck saying that it's best to relax on a beach chair every single day, but that a life consisting of these "best-at-a-time" actions is somehow bad. Rather, we can simply say that the *first* day of relaxing is a day well spent, but perhaps that the *tenth consecutive* day of doing so is not a day well spent. The same sort of response works for the "monk/dancer" cases: it's good to become a dancer after years of dancing school or as a first job after college; it's not quite as good to become a dancer when only a day before you were serious about becoming a monk. To put the point generally, we can hold fast to the stance that the best life is the one consisting of a series of actions, each of which is the best that could be done at its time, because there's no bar to the value of an action performed at one time depending upon features about other times, including actions performed at those other times.

Here's how this might be applied to Chang's worry about EOV maximization. Being an EOV maximizer is perfectly consistent with having a very high credence that the value of an action performed at one time depends, in part, on the actions one has performed at other times. So suppose my credence is divided between normative hypotheses H1, H2, and H3, and that that at time T1, I act in accordance with H2. I can attach greater value to future actions that accord with H2 *precisely because* I believe they exhibit a kind of narrative consistency with my action at T1. So I can be an EOV maximizer and follow one normative hypothesis over and over, so long as I have a high enough degree of belief that there's objective value to following one normative hypothesis over and over.

But that may be an inadequate response. While it’s possible for an EOV maximizer to display the kind of narratively consistent life we’ve been sketching, it’s also possible for her not to. Such a person will live this sort of life if she attaches enough objective value to narrative consistency; she may live a more fragmented life if she does not. Chang and others may want to say that someone who acts in accordance with the correct theory of rationality under normative uncertainty will exhibit narrative consistency whether she herself attaches objective value to narrative consistency or not.[65] EOV maximization quite clearly does not have this upshot, and so we’re confronted yet again with the possibility that it may be mistaken. To reiterate, we cannot escape this objection by imagining an agent with just the right beliefs about the objective value of narrative coherence, for the point of the objection now is that it’s rationally better for agents to live narratively coherent lives, whatever normative beliefs they have.

Now, one response to the renewed objection is that there’s a very obvious way in which the EOV maximizer’s life “makes sense” and falls under a “coherent narrative”: She acts in accordance with the same theory of practical rationality at every opportunity. Every one of her actions may be subsumed under a quick and easy explanation: given her credence distribution over normative hypotheses, she did the action with the highest EOV. If that’s not consistency, we might ask to the objector, then what is?

The objector is not without the resources for a reply. She will grant, I’m sure, that someone who maximizes EOV over and over again exhibits narrative consistency at one level. But someone who acts in accordance with CPlur exhibits narrative consistency at

[65] That is to say, they may want to levy the same sort of criticism against my initial response to Chang that I levied against Lockhart's response to the demandingness objection.

two levels. Like the EOV maximizer, his actions conform to the same theory of rationality under normative uncertainty on each occasion. But while the EOV maximizer's actions may accord with different hypotheses about objective value on different occasions (recall the Kantianism/utilitarianism case above), the CPlur-adherent's actions conform to whichever such hypotheses enjoy the plurality of his credence on each occasion. So until his credences in normative propositions change, he will be consistent at this second level, too.

With that said, it's unclear why consistency at every level is required for one's life to make sense or form a coherent narrative. Consider first the title character of Hermann Hesse's Siddhartha, who started his life as a sheltered Brahmin, then became a wandering ascetic, then studied with the Buddha, then had his "sex, drugs, and rock 'n' roll" phase, and finally settled peacefully at the banks of a river.[66] It would be absurd to suggest that the book and the life it depicted were nonsensical or incoherent, even though Siddhartha cycled rather drastically through different ways of life. All of these changes made perfect sense; we can explain them all in terms of Siddhartha's pursuit of Enlightenment. I think we can make a parallel claim in support of EOV maximization. Although the EOV maximizer exhibits a kind of inconsistency at one level, this very inconsistency is explicable in terms of consistency at another level, and so the EOV maximizer's life makes just as much sense as anyone else's. The degree of narrative coherence one's life enjoys is not a positive function of the number of levels at which one's behavior is consistent.

[66] See Hesse (1922).

Furthermore, it's far from obvious that narrative coherence requires the particular sort of consistency associated with acting in accordance with the very same first-order normative theory over and over and over again. Someone who maximizes EOV on each occasion repeatedly behaves in a way that is fully sensitive to his values. Being fully sensitive to your values involves letting some values take precedence at some times, and others take precedence at other times, depending on the relative degrees of each value at stake on these occasions. So long as all of the values are exerting influence on each occasion, it doesn't seem objectionable that some win out some times, and others win out other times. Just as someone who is certain of a commonsensical ethical theory may behave perfectly coherently by killing (in self-defense) in one instance, and refraining from killing (for kicks) in another instance, someone who is uncertain among ethical theories may behave perfectly coherently by acting in accordance with one theory in one case, and in accordance with another theory in another case. As always in life, stakes matter.

Nor should we lose sight of the unattractive aspects of guaranteed consistency at the level of objective value. It requires that one act in accordance with a theory of rationality like CPlur, the flaws of which we've already made manifest.

Absolutism and Infinite Value

There is an oft-raised objection that concerns agents with credence in so-called “absolutist” theories of value. Before we address the objection, though, let’s get a bit clearer about the nature of such theories. I’ll first explain them schematically, and then describe two theories that fall under the schema.

Absolutist theories may be characterized quite generally as follows:

a) Some factor F positively affects the value of actions that involve it,

b) Some factor G positively affects the value of actions that involve it, and

c) It is better to do an action that involves any degree of F than an action that involves any degree of G but a lower degree of F.

That's horrifically abstract, but I think necessarily so. Let me flesh it out with some examples. The theories most commonly associated with absolutism are a subset of the non-consequentialist theories. On some of these, it's wrong to kill no matter how many lives one might save by doing so. Here is how to state theories like this in terms of the schema above:

a) Not killing is a factor that positively affects the value of actions. (Alternatively, killing is a factor that negatively affects the value of actions),

b) Saving lives is a factor that positively affects the value of actions, and

c) It is better to do an action that involves any degree of not killing than an action that involves any degree of saving lives (i.e. that involves saving any number of lives) but a lower degree of not killing (i.e. that involves more killing).

That may seem like a tortuous way of re-stating this theory, but it's important to see that absolutist theories can all be accommodated under the same rubric.

Now, although it's sometimes overlooked, there are absolutist consequentialist theories as well. Some such theories are absolutist about the axiological values of outcomes, and are for this reason absolutist about the practical values of actions. Larry Temkin is fond of discussing cases involving, on one hand, an outcome in which one person is very badly off, and on the other, an outcome in which lots and lots of people are only slightly badly off. Someone's suffering excruciating torture for five years is an outcome of the first type; millions of people each suffering a mild headache is an outcome of the second type. Temkin tends to conclude about such cases that the first outcome is worse than the second, no matter how many people are implicated in the latter.[67]

It's not hard to see how this kind of absolutism about outcomes might translate into a consequentialist absolutism about the values of actions. A normative view that fits comfortably with Temkin's axiology might say that it's better to prevent the one person from being tortured than it is to prevent any number of people from suffering headaches. Here's how this theory falls under our general schema:

> a) Preventing people from being horribly tortured is a factor that positively affects the values of actions,
>
> b) Preventing people from suffering headaches is a factor that positively affects the values of actions, and

[67] See Temkin (2005) and (forthcoming), also Scanlon (1998).

> c) It is better to do an action that involves any degree of preventing people from being tortured than an action that involves any degree of preventing people from suffering headaches, but a lower degree of preventing people from being tortured.

The problem generated by absolutist theories is this: If both F and G positively affect the values of actions, but any degree of F is better in that regard than any degree of G with a lower degree of F, this must mean that the value of an action that involves that degree of F is infinite. Consider the non-consequentialist theory above. If it's better and better to save more and more lives, but not killing is always more important than saving lives, this can only be explained by the value of not killing's being infinite (or, put equivalently, the disvalue of killing's being infinite). If my credence is greater than zero in a theory according to which the value of some action is infinite, then the EOV of that action is also infinite. An action with infinite value will always be better than an action with only finite value, and so EOV maximization will always prescribe the action that the absolutist theory says has infinite value. Consequently, whenever I have even the slightest credence in an absolutist theory, I must do exactly what that theory says in situations about which it is absolutist. But it's highly counterintuitive that even the slightest grain of belief in an absolutist theory grants it so much "say" over my behavior. So EOV maximization must be false.[68]

[68] This objection is discussed briefly in Ross (2006) and (ms). Ross's response is essentially the one that I give in the next paragraph. Tim Campbell, Ruth Chang, Johnny Cottrell, and Josh Orozco have also presented me with this objection.

Let me start off with some gentle reminders before I launch into a more substantial response. First, while it does seem rather unappealing to be led around by theories in which you have so little credence, it's not entirely clear that the fault lies with EOV maximization, rather than with you for having any credence in such theories at all. Theories of rationality, after all, are handicapped in their ability to serve as normative cure-alls by the "Garbage In, Garbage Out" rule. If someone has the wrong beliefs or degrees of belief, even the right theory of rationality may recommend that he perform actions that are in some sense bad. It may be rational, in the internal, local sense with which I'm concerned, for the die-hard racist to behave in racist ways. So maybe the lesson here is not "EOV maximization is mistaken", but rather "Don't have any credence in absolutist theories". And for that matter, it may be that one of the reasons you shouldn't have any credence in such theories is that, if you do, you'll be led astray if you act in accordance with the correct theory of rationality under uncertainty: EOV maximization (or, really, if you act in accordance with any comparativist theory of rationality).

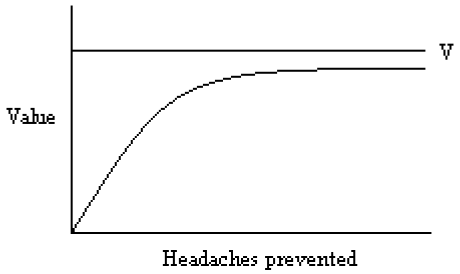
I should also remind you, once again, that EOV maximization is not a theory of what one must do, or of what one is required to do. It is a theory about what's more rational to do – that is, a theory about rankings, not statuses. So it's perfectly consistent with the foregoing objection's soundness that one is not required to act in accordance with the absolutist theory or theories in which one has credence. It may be, rationally the best thing to do, but perhaps one is permitted to do less than the best.

Now for the meatier responses. I wish to take issue with the characterization of absolutist theories in terms of infinite value. While it's tempting to think that a theory may meet conditions a) through c) only if it assigns infinite value, this temptation is to be avoided. There are other ways to assign numerical values to actions that are perfectly consistent with a) through c), and indeed, with our intuitive understanding of how absolutism works.

First, we might say that an action that involves some non-zero degree of factor F has a finite value, and that the value of an action that involves a lesser degree of F, and some degree of factor G, is an increasing function of the degree of G, bounded by a value no greater than the value of the first action.[69] That way, no matter what degree of G is involved in the second action, it can never match the value of the first action.

Here's how this method of representation would work for the Temkin-inspired theory: We say that preventing someone's excruciating five-year torture has a value of V, and that preventing headaches has a value that increases with the number of headaches prevented, but that can never equal or exceed V. (See Diagram 3.1)

Diagram 3.1: The Bounded Function Approach



There's also a sneakier way of assigning values consistent with a) through c). Let me introduce it by focusing on condition b) in the schema. If we remove this condition, then there is obviously no need to resort to infinite value, bounded functions, or any other

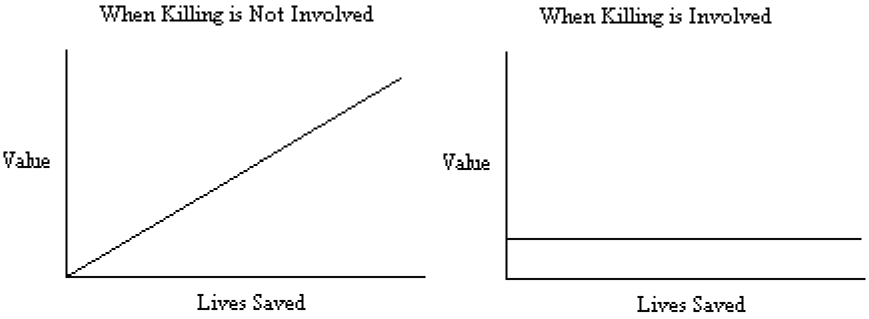
[69] Jackson and Smith (2006) also note this possibility.

clever machination. Consider, for example, views according to which saving lives has no value, or has a value that is a constant function of the number of lives saved. Then it's no mystery how not killing could be better than killing to save any number of lives. But few absolutists will want to go this route, for the fact that an action will save lives clearly does contribute positively to that action's value, and this contribution is surely greater the more lives that are saved.

But I think there's a way for the absolutist to have his cake, and eat it too – to maintain the spirit of b) without being pushed into infinite value or bounded functions. For we might suppose that the contribution of factor G to an action's value is context-dependent in what, for the absolutist, is a helpful way. Here are two ways this could work, focusing on the killing/saving case:

1) We could say that the value of an action is a positive function of the number of lives saved when the action does not also involve killing, but that the value of an action is only a constant function of the number of lives saved when the action also involves a killing. In other words, if I don't have to kill to save lives, then the more lives I can save, the better. But if I do have to kill to save lives, then the value of my action doesn't depend on the number of lives saved. So long as the value of not killing is greater than the value of that constant function, it will turn out that, when I'm faced with the choice between killing and letting any number of people die, it's always better to choose the latter. This is completely consistent with the value of saving, say, 30 lives without killing in one situation being greater than the disvalue of killing in another situation. (See Diagram 3.2)

Diagram 3.2: Context-Dependence Approach, #1



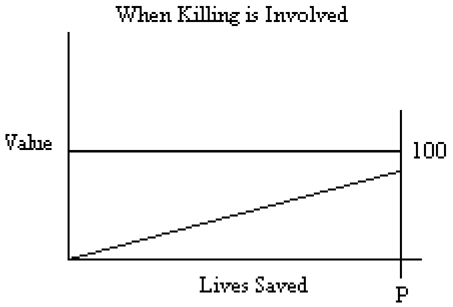
Now, that can seem unpalatable, since it implies that the number of lives saved makes no difference when there's killing involved.[70] No matter; the context-dependence approach can work in another way:

2) Perhaps the value of an action is always a positive function of the number of lives saved, but the function varies depending on a) whether one must kill to save that number of lives, and b) how many lives it's possible to save. It will help to illustrate this with an example. Suppose the disvalue of killing is 100. This means that the value of saving any number of lives cannot exceed 100 when it's possible to do so only through killing. Here's a set of functions consistent with that: When an action doesn't involve killing, then the dependence of the action's value, V, on the number of lives saved, N, is given by the function V = 10(N). This means that the action's value might be greater than 100, but since there's no possibility of killing, this would never license killing to save any number of lives. When an action does involve killing, then the dependence of V on N is given by the function V = (99/P)(N), where P is the maximum number of lives its possible to save in the situation. This function is always positive, but the multiplier (99/P)

[70] Structurally, however, this position is no different from the one adumbrated by John Taurek in his "Do the Numbers Count?" (1977), or from the one that undergirds Frances Kamm's "Principle of Irrelevant Utilities". See Kamm (1993) and (1996).

keeps getting smaller the more lives it's possible to save. The highest value the function can have is 99, since one will never be able to save more lives than it’s possible to save. Therefore, there’s no danger of licensing killing. (See Diagram 3.3)

Diagram 3.3: Context-Dependence Approach, #2



This may strike you as just too convenient, in the pejorative sense of the word, but formally speaking, it’s a perfectly adequate way of assigning numerical values to actions such that conditions a) through c) hold.

What we now have on the table are three approaches to representing absolutist theories, where, again, such theories are characterized in terms of conditions a), b), and c) above. Given that only one of these approaches involves infinite value, it’s a mistake to assume that, simply because someone has some credence in an absolutist theory, there is some action that has infinite EOV for her.

That doesn't diffuse the problem just yet; infinite value is still one way to represent conditions a) through c). However, I also think it's a particularly bad way of representing them, given the typical agent's other commitments. These other commitments are incompatible with the propriety of representing such an agent's absolutist hypotheses using infinite value. One of the other numerical representations must be correct instead. To see why this is so, consider how strange an agent's other normative beliefs would have to be if her absolutist hypotheses were representable using infinite value:

First, she would have to think it worse, in the non-normative belief-relative sense, to run any risk of doing the infinitely disvalued act than to run no risk of doing it, no matter how much might be gained through the former. This does not accord, I'd imagine, with most absolutists' beliefs. For example, most people who say you cannot kill one person to save any number of other people will not also say that you can't risk even the slightest chance of killing one. Furthermore – and even more decisively – since every act has <u>some</u> chance of being an infinitely disvalued act, it would more often be the case that <u>all</u> acts have a infinite expected disvalues; as such, they are all on the same level, in the non-normative belief-relative sense. Most absolutists will reject this. They will not say, for example, that there's no difference in non-normative belief-relative value between a act that has a .99 probability of being an instance of killing and one that has a .0000001 probability of the same. Finally, one would have to think that the difference between the infinitely disvalued act and any finitely valued act, on the absolutist hypothesis, bore an infinite ratio to any value difference on a non-absolutist hypothesis. Most of us will think it very implausible that killing, on absolutist deontology, stands to killing on normal deontology or consequentialism the way the latter stands to placing a small obstacle in an ant's pathway on moderate deontology or consequentialism. It accords much more with our intuitions to think that, from the absolutist deontological perspective, the moderate deontologist or consequentialist overestimates the value of saving lives than that she underestimates the disvalue of killing to this completely absurd degree.

That the results detailed above strike us as not at all what we had in mind suggests that the correct formal representations of the absolutist views in which we have credence do not involve infinity. It's often said that they do, but I suspect this is only because, first, people don't see the horrors wrought by these representations, and second, because people don't see that there are alternatives. But there are alternatives – at least two of them, as far as I can see: we say that the value of saving lives approaches the value of not killing asymptotically as the number of lives saved increases; or we say that the value of saving lives is contextually variant, such that it can never exceed the disvalue of killing when saving involves killing, but that it can increase unboundedly otherwise.

How things will look on these other representations is less important for now. It's sufficient to diffuse the problem that they exist, and that insofar as we disown the commitments of infinite value representations, our absolutist hypotheses are represented by one or both of these others.

Near-Absolutism and Extreme Values

The possibility of infinite value presented a challenge that was grave if genuine, but unlikely to be genuine. There is similar possibility that presents a challenge that is less grave if genuine, but more likely to be genuine. This is the possibility that values are not infinite, but simply extreme. Extremist hypotheses assign very great, but not infinite, values or disvalues to some actions. It's not the case that if one has any credence whatsoever that an extremist hypothesis is true, one is bound by its verdicts. If one's credence is low enough, the highest-OV action on this hypothesis will not have the highest EOV. The problem is that, if the hypothesis is extreme enough, then the threshold credence will be extraordinarily small. Any higher credence, even if it is still very small, and the extremist hypothesis wins out over its competitors. And this seems, at least on face, to be an undesirable result.[71]

How might we respond to this challenge? One solution is to restrict the scope of EOV maximization as a principle of rationality. On such a hybrid approach, we first exclude from consideration those hypotheses in which the agent has credence below a certain threshold. Stated in terms of my framework, we treat as locally rational relative to an agent's credences over all value rankings the action that is locally rational relative <ins>only to those credences in value rankings that are at or above the threshold</ins>. We then apply EOV maximization to just that subset of credences.

The idea is that the typical agent's credences in extremist hypotheses will be below some plausible-sounding threshold, and so they won't affect (some might say "skew") the EOV calculation. And if an agent does have credences above the threshold in extremist hypotheses, then perhaps it's not a counterintuitive result that they should play such a major role in determining what it's rational to do.

This is not the sort of move I wish to make. For one thing, I've shown that EOV maximization has advantages over another views of rationality – advantages that it'd be wise not to simply jettison when it delivers counterintuitive results. Furthermore, it's particularly unwise to throw away EOV maximization in favor of a theory that seems <ins>cobbled together</ins> in this way – a mixture of a general theory of rationality, and a tail awkwardly pinned to it for reasons that have nothing to do with the reasons supporting the general theory. Finally, and perhaps most crucially, such an approach seems clearly

[71] Thanks to Ruth Chang for urging me to distinguish this objection from the previous one, and to take it seriously.

mistaken in the context of <u>non-normative</u> uncertainty, and this gives us reason to doubt its suitability for cases of normative uncertainty. It is unwise to treat a small non-normative risk of catastrophe as though it were no risk at all, and if certain other arguments in this chapter are correct, then there's no good reason for treating normative uncertainty differently from non-normative uncertainty.

Nor do I see that there's much ground to be gained by directly rebutting the intuition that EOV maximization gives too much weight to extremist theories. It's a reasonable intuition, and if we put aside the arguments in the second half of the last chapter, there are reasonable theories of rationality that can capture it. The only thing I'd want to offer in direct response is a reiteration of my commitment to treating normative and non-normative uncertainty similarly, and an intuition of my own that discounting the value accorded to actions on unlikely extremist hypotheses makes little sense in the non-normative case. Suppose that we could save people a good deal of money by relaxing the safety standards for nuclear power plants, while increasing only slightly the risk of nuclear fallout. It'd be irrational, I should think, to relax the standards on the grounds that nuclear fallout is unlikely and extreme. Or suppose that the United States could make public life significantly more pleasant by repealing the First Amendment to the Constitution, while only slightly increasing the risk of tyranny. The remote but frightening possibility of tyranny is a reasonable ground for keeping the First Amendment intact.

My main counter-argument, however, is that the sort of intuition upon which the “extremism” argument rests is plausibly explained away. One problem with intuitions of this sort is that thinking about extremes is in general very difficult. As John Broome has argued, our moral intuitions were honed in environments consisting mostly of less extreme cases, and so it's not surprising that these intuitions would go off the rails when they are trained on alien environments involving billions of people, unimaginable pain, wealth beyond our wildest dreams, and so forth.[72]

Another problem with these intuitions is that imagining extreme values – especially extremely *low* values – of *one's own actions* is difficult. There's a sense of vertigo we feel when we imagine ourselves doing very evil actions. It's the sense that, once we've passed a “point of no return”, any censure cast in our direction for greater evils falls on an agent so utterly transformed by having passed that point that he ceases to be the sort of creature against whom censure is appropriate. It's difficult to imagine doing great evil because it's difficult to imagine being such a creature. There's a reason that, despite the well-documented “banality of evil”[73] (and of evildoers), the evil characters that connect most strongly with the popular imagination are unpredictable sociopaths, stony automatons, or self-worshipping emperors who seem so unlike anything we could ever think of ourselves being.

It's also possible that many of us hold, without realizing it, a view about objective value that leads us not to take seriously the value assignments on unlikely extremist hypotheses. It's the view that, in virtue of the low probability of some hypothesis, the objective values on that hypothesis really can't be that extreme after all. If there's a small chance that, say, some moral theory is correct, morality “won't allow” the possibility that someone could “win the moral lottery” or “lose the moral lottery” by making her actions, if that theory is indeed correct, incredibly good or incredibly bad. This is analogous in

[72] See Broome (2004), p. 56-59.

[73] The phrase is due to Arendt (1965).

some ways to the thought, shared by many, that if the probability of God's existence is very low, then if God does exist, He wouldn't punish non-believers by sending them to Hell.

This view requires that an agent's credence in a hypothesis is among the features that may help determine the objective values of actions on that very hypothesis – not the rationality of actions if that hypothesis is true, mind you, but their end-of-the-day objective values. I'll explore this interesting possibility at the very end of this chapter. I myself think it's implausible, but that's not what I'm concerned to show now. The present claim is simply that, if you hold such a view, then you're not taking at face value the possibility that you could have a very low credence in a hypothesis, and that that hypothesis could at the same time be extreme. You claim to be imagining extreme hypotheses, but your intuitions are really the result of imagining more moderate ones.

Each of these biases will lead us to underestimate the normative significance of very extreme hypotheses. If Broome's right, then I simply have trouble imagining extremes in general. If the "point of no return" suggestion is right, then I have particular trouble imagining myself doing something that is extremely bad. If the "no lottery" suggestion is right, then I'll resist the imputation of great objective values or disvalues according to very improbable hypotheses. I don't know how to get rid of these thoughts. But we should be aware of them, and insofar as they cause us not to take seriously the values assigned by extreme hypotheses, we should compensate for their effects.

Normative Fetishism and Grabby Theories

There are two otherwise insightful objections that trade on the same subtle misunderstanding of my project, and so I'll address them in the same section.

The first objection is that maximizing EOV under normative uncertainty requires one to be a "normative fetishist". The mark of the normative fetishist is his manner of motivation to do what he has reason to do. While the non-fetishist is motivated to do what he has sufficient reason to do, understood *de re*, the fetishist is motivated to do what he has sufficient reason to do, understood *de dicto*.[74] The non-fetishist's motivational states have content like *that I visit my elderly grandmother*, *that I pick my children up from school*, and *that I help those in need*. By contrast, the fetishist's motivational states have the content *that I do what I have sufficient reason to do* or something very similar. Michael Smith has argued that fetishism, particularly moral fetishism, is objectionable.[75] He has further argued that the motivational externalist can only explain the connection between moral judgment and motivation by imputing to people a motivation to do the morally right thing as such – in other words, by imputing a kind of fetishism. This, he says, puts the externalist in an unenviable position.

The fetishism-based argument against EOV maximization goes like this: Someone following CPlur, for example, can be motivated to do whatever the most probable normative hypothesis says is best, where this is understood *de re*. If my credence is .7 that it's better to give than to receive, and .3 that it's better to receive than to give, I can act in accordance with CPlur by gathering up the motivation to give. That is to say, I can safely be a non-fetishist. But someone following EOV maximization must be motivated to do whatever has the highest EOV, where this is understood *de dicto*, rather than to do any

[74] See Smith (1994), p. 127-128.

[75] Ibid.

particular action. So the EOV maximizer must be a normative fetishist. And since fetishism is objectionable, EOV maximization is also objectionable.[76,77]

The second objection also starts from the premise that EOV maximization requires that the content of one's motivational state be *that I maximize EOV*. But this objection does not imprecate this motivation on the grounds that it's fetishistic, but rather on the grounds that it may be disvalued by one or more of the normative hypotheses in which the agent has credence. Some normative theories are what I'll call "grabby": the values they assign to actions depend not only on features of the actions that are external to the body, but also on the motivations with which the actions are performed. Consider a theory that I'll call "Bastardized Kantianism", or "BK". According to BK, an action done from the motive of duty has a much, much higher value than an otherwise similar action done from any other motive.[78] The motive of maximizing EOV is not the motive of duty (according to this theory), and so if you follow the theory of EOV maximization, you'll have done an action with a much lower value according to BK than the action you would

[76] This argument was suggested to me by Brian Weatherson. A slightly different argument involving fetishism and normative uncertainty may be found in Weatherson (2002).

[77] I should say, for whatever it's worth, that I reject an underlying premise of both Smith's original fetishism argument, and the modified fetishism argument now on offer. This is the premise that fetishism is objectionable in the first place. I find myself straining to see what's so terrible about being motivated to do what you have most reason to do. And in some cases, it seems as if there's something commendable about being so motivated. As Lillehammer (1997) argues, there are certain intuitively commendable ways of changing one's mind in ethics that can only be rationalized through the imputation of the motivation to do the right thing, understood *de dicto*. But for argument's sake, I put these sorts of worries aside, and simply assume that there is something objectionable about fetishism.

[78] It is "bastardized" mainly because it uses ranking terms like *higher value than* rather than the absolute status terms in which Kant's own theory is couched, and because it elides Kant's distinction between the *rightness* of an action, and an action's *moral worth*. See Kant (1785).

have done had you not followed the theory of EOV maximization. And if you have credence in BK, then this reduction of value according to BK also means a reduction in EOV, so following EOV maximization may have the effect of reducing the EOV of one's actions. This is an unsettling result.[79]

The problem with these objections is that they misconstrue the nature of the theory on offer. I'm defending a theory about which actions are most rational to perform under normative uncertainty. It is not a theory about what our motivations should be, nor is it what we might call a "decision procedure". Therefore, it does not recommend that we be fetishistically motivated, or that we be motivated in a way that is disfavored by any of the normative hypothesis in which we have credence. I can be motivated to perform the actions that have the highest EOV, where this is understood de re, rather than by doing whatever maximizes EOV, understood de dicto. So too can I be motivated in whatever ways are favored by the normative hypotheses in which I have credence. It's not worrisome that BK attaches additional value to actions done from the motive of duty, for there is nothing in my theory of rationality under normative uncertainty that requires one to act from a contrary motive.

This is, however, a slightly churlish stance on my part. For while my theory of rationality is officially silent regarding motivation, there are limitations on the ways I can be motivated while still doing performing the actions with the highest EOV's. For example, the motivation to avoid harming others is non-fetishistic, but it's also probably not the best motivation to have for the purposes of maximizing EOV. Sometimes, after all, the action with the highest EOV will be an action that harms others. In developing our

[79] Thanks to Tim Campbell for helping me to develop this argument.

motivations, then, we must strike a careful balance. On one hand, we must develop those motivations that tend to lead to actions whose other features tend to render them high-EOV actions. On the other hand, we must avoid developing motivations that are themselves objectionable in some way or other.

How exactly we strike this balance will depend on, among other things, the way in which some motivations are alleged to be objectionable. The "grabby theory" objection states that the objectionability of motives other than duty may be cashed out in terms of their effects on EOV. Taking into account this type of objectionability does not require any theoretical overhaul; requires merely that, in considering which actions have the highest EOV's, we take into account <u>all features</u> of those actions, including the motivations that give rise to them. The fetishism objection is different. It does not take issue with any motive on the grounds that that motive is disfavored by any of the normative hypotheses in which the agent has credence. Rather, it takes issue with the fetishistic motive on the grounds that such a motive is somehow bad, independently of the agent's credence distribution over normative hypotheses – in other words, whether or not the agent himself believes to whatever degree that this motivation is bad. So in calibrating our motives to avoid objections like the fetishism objection, we will need to weigh the importance of EOV-maximization against the importance of the type of normativity that the anti-fetishist claims attaches to motivations. The exponent of the fetishism objection will have to say more about what kind of normativity this is before such a weighing may be fruitful.

<u>Why Think More About Normative Matters?</u>

This next objection is not only an objection to EOV maximization, but also, I believe, to all theories of rationality that take only the agent's credences in normative hypotheses as inputs. It's easiest to illustrate if we consider someone who is uncertain among normative theories, but it should be obvious how the objection extends to uncertainty about other normative hypotheses. Suppose my credence is divided between Theory 1 and Theory 2. Theory 1 says that A is the best thing to do, and that B is the second best. Theory 2 says that B is the best thing to do, and that A is the second best.

It seems as though, in some such cases, the most rational thing to do is to think more about the normative aspects of the situation. You should reflect on the features of the actions, consider the theoretical virtues of Theories 1 and 2, debate the matter with a worthy adversary, maybe consider some hypothetical situations, and so forth. And don't be fooled into thinking that it's only *theoretically* rational to think more about normative questions. It's also sometimes *practically* rational to think more. More thought will lead you to what is, from your current perspective, a better decision. So in the same way that it's practically rational to gather more evidence about the non-normative features of one's actions, it also seems practically rational to gather more "normative evidence" as well.

However, it's not obvious why, on any of the theories of practical rationality canvassed so far, this should be the case. For the act of engaging in normative deliberation is an action, too, and as such, will fall somewhere in the rankings implied by each of Theory 1 and Theory 2. But what if it ranks very low according to each of those theories? Then not only will it not have the highest EOV; it may be disfavored by risk-sensitive and non-comparativist theories, too. Actions A and B are the true candidates for being rational, while the act of thinking looks from this perspective like a sure loser. Nor are Theories 1 and 2 aberrational in this respect. Most well-known normative theories will assign high values to actions like helping others and keeping your promises, but very low values to sitting in your office and thinking about the values of helping others and keeping your promises.

In saying this, I'm not denying that some normative theories may assign very high value to normative reflection. They may. And obviously, it's not hard to explain why it's rational for agents with high enough credences in such theories to occasionally stop and think about normative matters, rather than engage in some other action. But this can't be the complete explanation for why this is sometimes rational. For it doesn't explain why it's rational for agents with very low credences in "pro-thinking" theories to reflect on their own normative views. It leaves us with the dispiriting conclusion that people who don't believe that normative reflection is very good shouldn't bother with normative reflection.[80] We need an explanation that steers us away from this conclusion.

Luckily, we may find the materials for such an explanation in I.J. Good's important paper, "On the Principle of Total Evidence" (1967). Good's primary aim in that paper was to respond to one of A.J. Ayer's (1957) criticisms of the logical interpretation of probability. Ayer claimed that the logical interpretation lacked the resources to explain why non-personal probabilities that were relative to larger bodies of evidence were in any sense privileged over non-personal probabilities that were relative to smaller bodies of evidence. Good wanted to show that the logical interpretation could make sense of this fact. In doing so, he proved a closely related, and probably more well-known result – that the expected value of gathering further evidence is almost always positive.

[80] I thank Barry Loewer and Larry Temkin for their input regarding this objection.

I will give a proof, based on Good's, of the result that the expected value of normative deliberation is almost always positive. I say "*based* on Good's" for two reasons. First, my proof starts from the version of the expected value function I employed in the last chapter, which separates out the probabilities of sets of normative facts from the probabilities of sets of non-normative facts. Good's expected value function does not. Second, my exposition is based on Brian Skyrms' version of the proof in *The Dynamics of Rational Deliberation* (1990), rather than on Good's original; Skyrms' version struck me as much easier to follow. Here is the proof:

The rational value of choosing now, assuming it's most rational to maximize EOV, is just the EOV of the best action now available:

1. $\text{Max}_m\ \Sigma_{i,j}\ p(N_i|NN_j) \cdot p(NN_j) \cdot v(A \text{ given } (N_i \wedge NN_j))$[81]

This is equivalent to:

2. $\text{Max}_m\ \Sigma_{i,j,k}\ p(N_i|NN_j) \cdot p(NN_j) \cdot \mathbf{p(E_k|(N_i|NN_j))} \cdot v(A_m \text{ given } (N_i \wedge Nn_j))$

Where $p(E_k|(N_i|NN_j))$ is the probability, upon engaging in normative inquiry, of arriving at a set of normative conclusions, $\{E_k\}$, conditional on a set of normative facts, which are themselves conditional on a set of non-normative facts.

[81] Max(F(*)) simply yields the maximum value in F(*)'s range. The maximum value for the EOV function is just the EOV of the action with the highest EOV, whichever action that is.

Now for the next step. The rational value of choosing later, after further normative reflection, is the EOV of the best action that will be available then. This is given by the quantity:

$$3. \qquad \mathrm{Max}_m\ \Sigma_{i,j,k}\ p((N_i|NN_j)|\mathbf{E_k})\cdot p(NN_j)\cdot v(A_m \text{ given } (N_i \wedge NN_j))$$

Formula 3 is just the same as formula 1, except that it replaces the probability of N_i conditional on NN_j with this very probability conditional on E_k. If you think about it, this replacement accords with our intuitions. Before I engage in normative reflection and thereby reach some normative conclusion E_k, the EOV's of the actions available to me will depend partly on the subjective probabilities of normative hypotheses not conditioned on E_k. After all, how could my credences in utilitarianism, Kantianism, or whatever, depend on normative conclusions that I haven't yet reached? But after I arrive at E_k, then, assuming I conditionalize on this evidence, the EOV's of my actions will depend on the subjective probabilities of utilitarianism, Kantianism, and so forth, conditional on E_k.

Now that we know how to calculate the rational value of acting later after you've reached some normative conclusion, we can calculate the rational value now of this future act, given that you're presently uncertain which normative conclusion you will reach. It is the expectation of the rational value of acting later, which is the expectation of quantity 3. above, which the expected EOV (that is, expected expected objective value, which you may recall from our discussion of the Winning Percentage Argument) of the best action available later, given the conclusions that you might reach as a result of your normative thinking:

4. $\Sigma_k\, p(E_k) \cdot (\mathrm{Max}_m\, \Sigma_{i,j}\, p((N_i|NN_j)|E_k) \cdot p(NN_j) \cdot v(A_m \text{ given } (N_i \wedge NN_j))$

By Bayes' Theorem, this is equal to:

5. $\Sigma_k\, p(E_k) \cdot (\mathrm{Max}_m\, \Sigma_{i,j}\, \mathbf{p(E_k|(N_i|NN_j)) \cdot p(N_i|NN_j\,)/p(E_k)} \cdot p(NN_j) \cdot v(A_m \text{ given } (N_i \wedge NN_j))$

6. $\Sigma_k\, \mathrm{Max}_m\, \Sigma_{i,j}\, \mathbf{p(E_k|(N_i|NN_j)) \cdot p(N_i|NN_j\,)} \cdot p(NN_j) \cdot v(A_m \text{ given } (N_i \wedge NN_j))$

Now compare 6 to 2. It is true on general mathematical grounds that 6 is greater than 2 whenever the EOV-maximizing post-reflection act does not have the same EOV for all values of k – basically, whenever what you will do after normative reflection will depend on the conclusions you reach.[82] This means that the rational value of acting after normative reflection is greater than rational value of acting without normative reflection, so the rational value of normative reflection is positive.

This proof depends upon several assumptions that are worth making explicit. Skyrms points out some of these: First, it assumes that the same generic actions are available both before and after deliberation. An example should help to show why this is a key assumption. Suppose I must decide this very second whether to marry Miss Z. If I hesitate for a moment, this opportunity will be lost. In that case, it is obviously of no rational value to sit around and think. I might as well just say "no". Second, the proof assumes that agents are conditionalizers, and moreover, that they know they are

[82] See Skyrms (1990), p. 89.

conditionalizers. Again, an example will help. Suppose that I respond to evidence in some way other than conditionalization. To make the case especially stark, imagine that I just ignore evidence entirely; it makes no dent in my credal state. If we suppose further than I know this, then it makes no practical sense for me to sit around and think, since this will have no effect on my future credence distribution over normative hypotheses, and hence no effect on what I do. (Of course, I shouldn't ignore evidence in the first place, but that's another matter. Given that I do, it's irrational to sacrifice other opportunities so that I may gather more of it.)

There is another important assumption that is overlooked by Good, Skyrms, and every other discussion of this issue of which I'm aware.[83] All this proof shows is that, from the present perspective, post-deliberative action has a higher rational value than pre-deliberative action. We're then whisked to the conclusion that the deliberative act itself has some kind of rational value, which accrues to it on the grounds that it makes possible post-deliberative action. But we cannot draw this conclusion without a further assumption about how the values of future acts affect the instrumental values of present acts that affect the probabilities of those future acts being performed. For it is an odd, but not obviously incoherent, stance to say, "I realize it's better to act after thinking a bit than it is to act right now without thinking. But I don't think my present reasons for present actions are in any way affected by my future reasons for future actions. So I don't think I have any instrumental reason to think about what to do. This will leave my future self worse off, but that's his loss, I say."[84]

[83] See, for example, Lindley (1971), Graves (1989), and Loewer (1994).

[84] Frankly, I think it's difficult for many non-consequentialists to explain why this bizarre stance is a mistake. For they will typically say that I cannot kill one person now to

I should also say something quickly about conceptualizing the results of normative inquiry as evidence, upon which it's proper to conditionalize. This way of thinking may strike some as jarring, but it doesn't seem problematic to me. If there are normative facts, then those normative facts are evidence for normative propositions (just as physical facts are evidence for hypotheses in physics). So, for example, if it's a fact that it's wrong to kill one person to save five, then that might be evidence against consequentialism as a normative theory. So what about, say, normative intuitions? Well, to put it informally, evidence of evidence is itself evidence. That someone has the intuition that it's wrong to kill one to save five is evidence that it's wrong to kill one to save five – the more reliable the intuition, the stronger the evidence. And as we noted above, the fact that it's wrong to save five may be evidence against consequentialism. Normative intuitions play the same secondary role that observations of physical facts play in physics. If physical fact F is evidence for physical hypothesis H, then the observation that is evidence of F is itself evidence for H. Of course, you might think that normative intuitions are unreliable. But that just means they're akin to non-reliable observations (or perhaps to unreliable methods of detection like guesses or tarot card readings, etc.).

And what about normative arguments? I want to suggest that we should not think of arguments as evidence. Suppose there is a valid argument from premises P, Q, and R to conclusion S. I'd want to say that P, Q, and R, together, constitute evidence for S. But the fact that there's an argument is not further evidence. For there to be such an argument is just for P, Q, and R to be a certain kind of evidence for S. So counting the existence of the

prevent myself from killing two in the future. The very "time-relativity" that's required for this view also seems to rule out doing the presently sub-optimal act of normative reflection so that I may do a better act in the future than I could have done without such reflection.

argument as evidence is double-counting. Does this mean that arguments are irrelevant to what I should believe, given that they don't constitute further evidence over and above their premises? Certainly not. When I come to know an argument, I'm made aware of the evidential relationship between the premises and the conclusion. So even if the evidence was already "out there", in some sense, it's only part of *my* evidence after I've learned the argument. And I can only *rationally* update my beliefs based on my evidence; undiscovered or unrecognized evidence is, in respect of rationality, inert.

The Precarious Balance of Reasons

Joshua Gert argues that there are two sorts of strength of reasons – *requiring strength* and *justifying strength*. From this, he draws some conclusions that are highly uncongenial to my project:

> “...when one appreciates the nature of the two kinds of normative strength, it will become clear that maximizing is not really a coherent goal, that the general advice “Act on the stronger reasons” is typically quite useless and confused, and that phrases such as ‘the balance of reasons’ or ‘what there is most reason to do’, even taken as metaphorical, are so misleading that they ought never to be used.”[85]

If Gert is correct, then I'm wrong to focus my attention on objective rankings as inputs, even objective rankings that include some parity and incomparability here and there, for there are no such things. I'm also wrong to focus on rationality rankings as outputs, for it will make no sense to speak of what's “most rational” in my sense either. So it's incumbent upon me to show that Gert is mistaken.

[85] Gert (2007), p. 535.

Once we see how he defines “requiring” and “justifying” strength, respectively, this should not be difficult. He begins with the idea of a reason's playing a requiring and/or justifying role. A reason plays the requiring role insofar as it “explain[s] why actions that would otherwise be rationally permissible are in fact irrational”.[86] A reason plays the justifying role insofar as it “explain[s] why actions that would otherwise be irrational are in fact rationally permissible.”[87] He then defines requiring and justifying strength in terms of these roles, respectively. A reason, R, has more requiring strength than another reason, S, just in case, “in playing the requiring role in actual and counterfactual circumstances, R can overcome any reason or set of reasons that S can overcome, and there are some reasons or sets of reasons that R can overcome but S cannot.”[88] R has more justifying strength than S just in case, “in playing the justifying role in actual and counterfactual circumstances, R can overcome any reason or set of reasons that S can overcome, and there are some reasons or sets of reasons that R can overcome but S cannot.”[89]

A few comments about the way Gert sets things up: First, we should be clear that his notion of rationality is not the “internal” one that I've been employing. He's working with a conception according to which what's rationally required is what I'd call “objectively required”. Second, there's something askew about his adverting to “what would otherwise be” permissible or impermissible. Suppose you're deciding whether to decapitate someone who's lying on your mother's new white sofa. That this would ruin

[86] Ibid., p. 537.
[87] Ibid., p. 538.
[88] Ibid., p. 538.
[89] Ibid., p 539.

the sofa suffices to render the action impermissible. What then, do we say about the role played by the reason not to kill the person? On Gert's formulation it wouldn't play the requiring role, since the action is already one that you're required not to do. But of course it wouldn't play the justifying role, either. This is the sort of small problem that we should want cleared up before jettisoning the traditional picture of univocal reason strength in favor of Gert's revisionary picture.

Now let's move on to the important stuff. What does Gert mean by “irrational”, or “rational”? Is he using these as ranking terms, or as status terms? It's clear that, by “irrational”, say, he can't mean “not supported by the balance of reasons”, for he wants to claim that it's nonsense to speak of the balance of reasons. More generally, these can't be ranking terms, because he doesn't believe in all-reasons-considered rankings of the sort I've been working with. So they must be status terms – “irrational” means “forbidden”, “rational” means “permitted”, perhaps, and “rationally required” means “required”.

But there's the problem for him if this is so. It's not a conceptual truth that an action's status is determined by its ranking in such a way that you're required to do what you have most reason to do. There are other determinants of statuses – features other than reasons, for example, or the strength of particular types of reasons (e.g. moral reasons), or quantitative features of reasons other than their raw strength. (Again, this will be our focus in Chapter 6.) Given that statuses are determined by features other than overall reason strength, though, it's perfectly possible for reason R to be able to overcome more in playing the requiring role than reason S and for S to be able to overcome more in playing the justifying role than reason R, even if reasons have a univocal strength. This can be the case because R and S differ with respect to the determinant(s) of status other than overall reason strength.

Consider one particular view of statuses, whereon the determinants of statuses are strength of reasons and blameworthiness: a required action is one that I have most reason to do, and that I'd be blameworthy for not doing. Perhaps you disagree with this view; I'm using it here only to illustrate a more general point. On such a view, it may be that R is a reason of some strength, and one tends to be blameworthy for not doing the action(s) that R supports; however, S may be a reason of much greater strength, but one tends not to be blameworthy for not doing the actions(s) that S supports. R, then, can overcome more in playing the requiring role than S. However, R may be very poor at defeating blameworthiness, since its strength is less, while S may be very good at defeating blameworthiness, since its strength is greater. In that case, S can overcome more in playing the justifying role than R.[90]

All of this means that we can explain exactly what Gert wants to explain, while holding onto the notion of a univocal strength of reasons. I take it that the picture I'm offering is, in fact, the standard picture: where an action "ranks on the scale" is one thing; whether it's required or permitted or forbidden is another. This seems to be the way of carving up the terrain that's employed in debates about whether we're required to do what's best – the two opposing sides of that debate can agree about (univocal) ranking, but they disagree about statuses. I can see nothing in what Gert says to imperil this standard picture.

It's been suggested to me in conversation that my way of setting up the debate is a

[90] That statuses are determined by features other than overall reason strength seems to me to play a crucial role in Kamm's (1985) arguments for the intransitivity of obligation and permission. So the point I'm making against Gert isn't entirely new.

mere notational variant of Gert's way.[91] If this is true, though, it's a point against Gert and for me. He's the one trying to argue from the premise that it's possible for R to be better than S at making things required but S to be better than R at making things permitted, to the conclusion that talk of strength of reasons is nonsense. My claim is that his premise can be retained in a way that's consistent with such talk's making perfectly good sense. All that's required is a conceptual gap between ranking and status, such that the determinants of the latter are features other than the former. And it seems to me that such a gap must exist in order for many of the standard debates in practical philosophy to make any sense at all.

Pressing further on the same theme, there are better and worse notational variants, and it seems to me that my way of setting things up is more helpful than Gert's. Rankings and statuses play different roles in our normative practices. The guidance of action depends most fundamentally on my judgments about rankings. If I judge that there's more reason to do A than B, then in the absence of akrasia, I'll do A. Statuses play a less fundamental role in the guidance of action. For example, my judgment that A and B are both permitted will impel me neither one way nor the other. And suppose that I'm choosing between A, B, and C, and I judge that both A and B are forbidden vis a vis C. It still may be more practically rational for me to do A than to do B, if A ranks higher than B. But all of this is consistent with statuses playing other important roles – in explaining the propriety of post facto attitudes like regret and blame, in explaining the presence or absence of duties of repair following from sub-optimal actions, and so forth. My way of dividing up the terrain separates out notions that seem to play different roles. Gert's way

[91] This was by Sergio Tenenbaum, in conversation.

doesn't; it focuses only on the role of reasons in determining actions' statuses, and denies that there is any way of ranking actions that is independent of this role. But there's more to normativity than deontic status, and in particular, there's more to the action-guiding aspect of normativity than deontic status.

With that said, the claims of this last paragraph are optional for the purposes of my project. All I need to show is that Gert's claims about requiring and justifying roles of reasons may be accommodated within the standard ranking/status picture. Once we see that the determinants of statuses are not conceptually limited to actions' rankings, this is not difficult to show.[92]

The Limits of Form

Often, when we propound formal constraints on action, we have in the backs of our minds images of what would count as acting within those constraints, and of what would count as violating them. We're then surprised – sometimes even dismayed – when what would seem like transgression can be shown, through some theoretical fancywork,

[92] Jonathan Dancy (2004) employs a distinction that's similar to Gert's – between what he calls "preemptive reasons" and "enticing reasons". They differ, Dancy claims, in that only the former are relevant to what I ought to do. What we make of Dancy's view will depend on how we understand "ought". If we take "I ought to do A" to mean "I'm required to do A", then Dancy's view is essentially the same as Gert's (except that Dancy assigns the different roles to reasons themselves, rather than to kinds of strength, each of which might inhere in the same reason; Gert is right in assigning the roles to kinds of strength rather than to reasons themselves.) Otherwise, however, I can make no sense of Dancy's suggestion that there are reasons that play no role in determining what one ought to do. It is possible, perhaps, for features to be relevant to what, following Thomson and Wiggins, I'd called the "evaluative" assessment of action but not the "directive" assessment of action. Nothing in my way of setting things up in this dissertation precludes there being features of this sort, since, again, I focus entirely on the "directive" side of the directive/evaluative line. But I'd be disinclined to call such features reasons, on the same grounds that I'd be disinclined to call the pleasure that exists in a state of affairs a normative reason for that state of affairs.

to be compliance all along. Here are two examples: Consequentialism is thought to constrain action in certain paradigmatic ways. For instance, it's thought to require us to push the fat man in front of the trolley in order to save the five people in the trolley's path. After all, five deaths is surely a worse consequence than one, no matter what accounts for the badness of death. But there are ways to render <u>not</u> pushing the man compatible with consequentialism. We can say that <u>killings</u> contribute greater disvalue to outcomes than do "natural" deaths – so much so that the outcome in which I kill the man is actually worse than the outcome in which the five die by runaway trolley.[93] Or, we can be even more sophisticated, and countenance such theoretical devices as agent- and time-relative value of outcomes, which allow us to represent every normative theory as a form of consequentialism.[94]

The second example: The rule that one's preferences must be transitive is also thought to place serious constraints on action. For instance, it's thought to rule out my trading an ordinary banana and a nominal sum for an ordinary apple, then trading the apple and a nominal sum for an ordinary peach one minute later, then trading the peach and a nominal sum for the banana one minute after that. But there are ways to render such a sequence of transactions consistent with transitivity of preferences. We can say that the first banana and the second banana count as different options because the second is two minutes riper than the first. Or, we can include in the description of each the options the choice situation in which it is embedded. That is, we can call the first banana <u>a banana</u>

[93] See the "utilitarianism of rights" raised and rejected by Nozick (1977)

[94] See, e.g., Dreier (1993) and Portmore (2009). This sort of approach is criticized in Schroeder (2006) and (2007).

when the other option is an apple, and the second banana *a banana when the other option is a peach*. Intransitivity averted.

This normative slipperiness also afflicts the view that it's most rational to maximize EOV. For this is a theory of what it's rational to do, *given* your degrees of belief in normative propositions. It has no "say" over what your degrees of belief should be in the first place. And as we shall see, people with substantial enough degrees of belief in certain normative views will maximize EOV by acting in ways that we might have thought were inconsistent with EOV maximization. In this section, I want to have a look at how and why this is, and then consider some ways of responding to this problem, which I call the Limits of Form Problem.

At the heart of the problem lies the fact that there are certain features of the world which obtain only when you are normatively uncertain, or when you act under normative uncertainty. For example, anyone can boil a lobster alive. But only someone who's uncertain about the lobster's moral status can boil a lobster alive *while being uncertain about its moral status*. An agent is not conceptually barred from having a non-zero degree of belief that her very own normative uncertainty is among the factors that affects the objective values of her actions. She could have some credence that an action performed under normative uncertainty has a different objective value than the same action performed under normative certainty. More generally, she could have credence that an action performed with a certain array of credences over normative propositions has a different objective value than the same action performed with a different array of credences over normative propositions.

Let's be a bit more concrete, and return to the example of abortion. We may imagine an agent whose credences are .8 that having an abortion is better than not having one, and .2 that not having an abortion is better than having one. And suppose that her difference in value between the two actions on the former hypothesis is 1/20 of her difference between the two actions on the latter hypothesis. Not having an abortion has the higher EOV, so on my view, it's the more rational action.

But now imagine a different agent – one who says the following: "I've got high credence that this is true: If you're really sure that abortion is worse, and you have an abortion anyway, that's very bad. That is to say, it has low objective value. But I've also got high credence that this completely consistent view is also true: If you think that abortion is probably not worse, and you have an abortion anyway, then that's not so bad at all. It's objective value isn't very low. And I happen to have a very low credence that abortion is worse."

This person might have a credence of .8 that having an abortion with her current credal state is better than not having an abortion given this credal state, and .2 that not having an abortion with this credal state is better than having an abortion with this credal state. And the difference between her two actions on the former hypothesis might be equal to the difference on the latter hypothesis, in which case having the abortion will have the higher EOV. This very same agent might have a credence of .4 that having abortion with a different credal state is better than not having an abortion with that other credal state, and .6 not having an abortion with this different credal state is better than having an abortion with that state. And the difference between the two actions on the former hypothesis might be, let's say, 1/20 the difference between the two actions on the latter hypothesis. After all, there's a non-normative difference between my credal state being .8/.2, and its being .4/.6, and so its possible for her to think this makes a normative difference, too, at the level of objective value.

This erodes the force of the injunction to maximize EOV. What I consider the gist of EOV maximization, and indeed of most comparativist theories of rationality, is that risk matters. Even if my credence is only .01 that doing A is better than doing B, it will sometimes be rational for me to do A, if, on that hypothesis, the difference between the two actions is sufficiently great. But on the views we've been sketching, the differences in value on different normative hypotheses will themselves depend on my credences – either in those very hypotheses, or in other normative propositions. If such views are right, then risk still matters when acting under normative uncertainty; it's just that the very condition of one's being normatively uncertain may, for example, dissolve the risk by reducing the differences in value between actions on less probable hypotheses.

A few comments on this possibility before we address it head-on. First, it should be easy to see that this phenomenon is not unique to action under normative uncertainty. An agent could also have substantial credence in a view according to which the relative sizes of value differences depended on that agent's own credence distribution over a set of non-normative hypotheses. Second, I should warn you away from tempting response that is nonetheless mistaken. One might want to say that the attitudes of the agents we've been discussing are incoherent. I've been imputing to these characters credence in views on which their own degrees of belief are relevant to the objective values of actions. But, you might ask, isn't objective value, by definition, independent of the agent's beliefs? Isn't only belief-relative value, well, relative to the agent's beliefs? As we saw in Chapter 1, though, this is much too sloppy a way of drawing this distinction. An agent's beliefs are as much a part of the world as anything else, and it's possible for someone to hold views on which these beliefs form part of the supervenience base for facts about the objective values of actions.[95]

Third, it might now be objected that, even if the imputed beliefs aren't incoherent, they're at the very least silly, and not worth spilling so much ink over. Nor are they silly in the sense of "obviously wrong"; rather they're silly in the sense that nobody would ever actually hold them. This objection strikes me as underestimating the currency of these views. I think a decent number of people, even very intelligent people, have the idea that there's little-to-no potential for normative catastrophe when we act under uncertainty – that the mere fact that an action is done under uncertainty about its normative or non-normative features is enough to divest that action of the substantial objective disvalue it might have had were it performed under conditions of certainty. This was the "anti-lottery" intuition I registered in discussing extreme theories above. This explains, I should think, why a troublingly large proportion of people consider the proper response to uncertainty to be "flip a coin", or "go with your gut". It's not that they hold bizarre views of rationality on which risks don't matter; it's that they hold bizarre views about objective value according to which the situations I've been calling "normatively risky" are not really so.

[95] An unpublished note of Preston Greene's helped me to see this point.

My own response to the Limits of Form problem is, frankly, to acknowledge it and suggest that this dissertation is not the place to address it. My aim was to defend comparativism, and in particular, EOV maximization, under normative uncertainty. All along, we've acknowledged that some agents may have such strange views about what their reasons are that they will wind up behaving in rather awful ways under the banner of EOV maximization. As we're now able to see, some of these are ways that we'd intuitively think are inconsistent with EOV maximization. This is not a strike against EOV maximization as a theory of rational action. Rather, it is a reminder that this sort of theory is not a comprehensive theory of normativity. The task of showing that the agents discussed in this section are making some sort of error falls to someone else – someone who will argue that they're irrational, or at the very least wrong, in having any credence in these nettlesome normative views in the first place. Once again: "Garbage In; Garbage Out". But that doesn't mean that reducing the "Garbage In" is not a valuable task. It's just not *my* task.

CHAPTER FOUR: THE PROBLEM OF VALUE DIFFERENCE COMPARISONS

If there's anything well-established in the literature on normative uncertainty, it's that comparativist approaches like EOV maximization suffer from a potentially debilitating problem that I call the Problem of Value Difference Comparisons (PVDC).[96] Imagine that my credence is divided between two rankings – A is better than B, and B is better than A. To determine whether A or B has the higher EOV, I must know how the degree to which A is better than B if the first is true compares to the degree to which B is better than A if the second is true. For example, if my credence is .1 that A is better than B and .9 that B is better than A, the difference between A and B on the former hypothesis must be greater than 9 times the difference between B and A on the latter hypothesis if A is to have the higher EOV.

I can't determine that from the hypotheses themselves. Each of them tells me which actions are better than which other actions, but neither tells me how the differences in value between actions, if it is true, compare to the differences in value between actions if some other hypothesis is true. It may help to think of the matter in terms of comprehensive theories. Some consequentialist theory may say that it's better to kill 1 person to save 5 people than it is to spare that person and allow the 5 people to die. A deontological theory may say the opposite. But it is not as though the consequentialist theory has, somehow encoded within it, information about how its own difference in

[96] See Hudson (1989), Lockhart (2000), and Ross (2006), Sepielli (2009). This problem is also noted in unpublished work by John Broome, Andy Egan and Alan Hájek, and Nick Bostrom and Toby Ord. Something like it seems to have been recognized in the Catholic tradition's treatment of normative uncertainty, specifically in connection with the Compensationist position. See The Catholic Encyclopedia (1913) and Prümmer (1957).

value between these two actions compares to the difference in value between them according to deontology. The problem, then, is that although it seems as though we ought to be sensitive to value difference comparisons across normative hypotheses, we lack a way of making such comparisons.[97]

It's important to emphasize that this is a metaphysical and not solely an epistemological problem. The claim is not that there may be some fact of the matter about how the A-B difference above compares with the B-A difference, but we just can't know it because of insufficient access to our own minds, or deficiency at recognizing what normative hypotheses logically entail. Rather, the claim is that there's just no fact of the matter about how value differences on one hypothesis, as it's represented by an agent, compare to value differences on another hypothesis, as it's represented by the same agent. If all that's in our heads is a credence that A is better than B, and a credence that B is better than A, there's just not enough structure there to make it the case that the size of the former difference compares to the size of the latter in any way whatsoever. As we'll see later, my solution depends on that's <u>not</u> being all that's in our heads. But first let's look at what Ted Lockhart and Jacob Ross have said about the Problem of Value Difference Comparisons.

<u>Lockhart's Proposed Solution</u>

Ted Lockhart focuses on uncertainty among comprehensive moral theories, but

[97] This problem is in some ways analogous to the welfare economists' problem of interpersonal comparisons of utility. See Robbins (1938), Harsanyi (1955), Hammond (1976), Elster and Roemer (1991), and Broome (1995) and (2004) for just a sample of the work on that topic. As we'll see later, what's required for a solution to the PVDC is structurally analogous to what's required for a solution to this other problem.

since he ignores features of theories other than the rankings they assign, what he says is perfectly applicable here. He suggests that we compare moral value across theories via a stipulation he calls the Principle of Equity among Moral Theories (PEMT):

> “The maximum degrees of moral rightness of all possible actions in a situation according to competing moral theories should be considered equal. The minimum degrees of moral rightness of possible actions in a situation according to competing theories should be considered equal unless all possible actions are equally right according to one of the theories (in which case all of the actions should be considered to be maximally right according to that theory).”[98]

The idea is that, if I have credence in theories T, U, and V, I set the value of the best action according to theory T equal to the value of the best action according to theory U equal to the best action according to theory V; same goes for the worst action according to each theory. (I'm using “value”/”strength of reasons” rather than “degrees of rightness”, but I can only imagine that Lockhart and I are expressing the same concept.)

So how does the PEMT fare as a solution to the Problem of Value Difference Comparisons? Lockhart claims that the PEMT makes these comparisons possible, and that it is attractive in its own right. I think he is wrong on both counts.

First, it is incompatible with the intuitive claim that moral theories disagree not only about what to do in different situations, but about which situations are “high stakes” situations and which are “low stakes” situations, morally speaking. A momentous decision from the perspective of traditional Christian ethics may be a relatively unimportant decision from the utilitarian perspective. But according to PEMT, all moral

[98] Lockhart (2000), p. 84.

theories have the same amount “at stake” in every situation.[99]

Second, the PEMT is arbitrary. Consider: It’s not difficult to find a method of comparing values of actions across theories. I could, for example, declare by fiat that the difference in moral value between lying and not lying, on a Kantian deontological theory, is equal to the moral value of 23 utils, on a utilitarian theory. But if there’s no principled reason for that “rate of exchange”, I haven’t solved anything. And, similarly, if there’s no principled reason to use the PEMT, rather than some other possible method, Lockhart hasn’t solved it either.

Lockhart recognizes that the PEMT may appear ad hoc, and tries to provide a reason why it, rather than some other principle, is the correct method of comparing values of actions across theories. He says:

> “The PEMT might be thought of as a principle of fair competition among moral theories, analogous to democratic principles that support the equal counting of the votes…in an election regardless of any actual differences in preference intensity among the voters.”[100]

Lockhart is right that the PEMT is analogous to this voting principle. But while the latter makes good sense, the former does not. One cannot be unfair to a moral theory as one can be unfair to a voter. And yet, presumably, fairness is why we count votes equally, regardless of preference intensity. Insofar as we care only about maximizing voters’ preference satisfaction, equal counting of votes seems like quite a bad policy. Rather, we would want to weight peoples’ votes according to the intensity of their preferences regarding the issue or candidates under consideration. Similarly, insofar as

[99] A version of this objection also appears in Ross (2006), p. 762, n. 10.

[100] Lockhart (2000), p. 86.

we care about maximizing EOV, it seems quite bizarre to treat moral theories as though they had equal value at stake in every case. If some act would be nightmarish according to one theory, and merely okay according to another, it seems right to give the first theory more "say" in my decision.

The gist of the analogy, though, is that we should somehow treat moral theories equally. But even granting that some "equalization" of moral theories is appropriate, Lockhart's proposal seems arbitrary. Why equalize the maximum and minimum value, rather than, say, the mean value and the two-standard-deviations-above-the-mean value? And especially, why equalize the maximum and minimum value with regard to particular situations, rather than the maximum and minimum conceivable or possible value? This is all to make a more general point: It seems as though we could find other ways to treat theories equally, while still acknowledging that the moral significance of a situation can be different for different theories. Thus, even if we accept Lockhart's "voting" analogy, there is no particularly good reason for us to use PEMT rather than any of the other available methods.

Third, the PEMT is nearly useless. It requires that all theories have highest-possible-valued acts of equal value, and lowest-possible-valued acts of equal value. But it tells us nothing about how to assign values to the acts that are intermediately ranked according to the various theories. That is, PEMT is a way of doing what I'll later call "normalizing" the different normative hypotheses, but not at all a way of "cardinalizing" them. Lockhart recognizes this, and rather halfheartedly suggests that we employ the "Borda count" method as a solution. On this method, we assign values to options equal to their numerical ranking on an ordinal worst-to-best scale. So, suppose I am deciding which Sonic Youth album to listen to. My worst option is Washing Machine; my second worst option is Murray Street; my second best option is Sister; and my best option is Daydream Nation. The Borda count method would assign values of 1, 2, 3, and 4, respectively, to these options.

Lockhart recognizes the flaws of this method, perhaps the most interesting of which is that it has the consequence that the value of any of the options depends on how many other options there are. Anyhow, it seems clear that if the defender of PEMT needs to use Borda counting to make use of his principle, then his principle isn't particularly useful.

Fourth, PEMT, even without the Borda count method, has the consequence that the EOV's of actions will depend on which other actions are possible in a situation. This is a violation of something akin to the Independence of Irrelevant Alternatives, which we discussed in Chapter 2 as a problem for non-comparativist theories.[101] Suppose that your credence is divided between Theory 1, according to which A is better than B, and Theory 2, according to which B is better than A. Now, if A and B are the only two options in a situation, then by PEMT, the value of A on Theory 1 must be equal to the value of B on Theory 2; mutatis mutandis for B on Theory 1 and A on Theory 2. For illustration's sake, let's just assign absolute values to these action-theory pairs:

Theory 1: Value of A = 10; Value of B = 0

Theory 2: Value of A = 0; Value of B = 10

[101] Thanks to Brian Weatherson for suggesting this argument.

But now suppose an additional action becomes available that is better than A according to Theory 1, and, say, ranked between A and B according to Theory 2. This action – call it "C" – will then be the best action on Theory 1, and neither the best nor the worst action according to Theory 2. So the value assignments will have to change slightly:

Theory 1: Value of A = ?; Value of B = 0; Value of C = 10

Theory 2: Value of A = 0; Value of B = 10; Value of C = ?

As we observed earlier, Lockhart gives us no good way of assigning values to sub-optimal but super-minimal actions, so the value of A according to Theory 1 and the value of C according to Theory 2 will have to stay "up in the air" for now. But one thing's for sure: The value of A according to Theory 1 must be less than 10, since the value of C according to that theory is 10, and the theory ranks C over A. However, A's value on Theory 1, before C got added to the mix, was 10. So the addition of C had the consequence that A's value on one theory, and therefore A's expected value, were reduced, vis a vis B's. This is an unwanted result.

Now, as we said in Chapter 2, it's possible to deny the IIA (or rather, it's possible to say that the so-called "irrelevant" alternatives are relevant after all). But the arguments for doing so all involve concrete features of particular acts. PEMT, like non-comparativism about rationality, leads to violations of this principle in virtue of merely formal features of the choice situation – the number of actions, and the rankings of the actions according to the various theories.

Fifth, the PEMT leads to inconsistent results when applied. Suppose my credence is divided between two moral theories. According to Theory T, the value of an action is a positive linear function of the number of instantiations of some property P that the action causes. According to Theory U, the value of an action is a positive linear function of the number of instantiations of some property Q that it causes. Now imagine two situations. In one situation, I can either cause 100 instantiations of P and no instantiations of Q, or 10 instantiations of Q and no instantiations of P. In the other situation, I can either cause 100 instantiations of Q and none of P, or 10 instantiations of P and none of Q.

Situation 1

100 P's + 0 Q's OR 10 Q's + 0 P's

Situation 2

100 Q's + 0 P's OR 10 P's + 0 Q's

If PEMT is true, then in both situations, the maximum possible moral value according to Theory T must be equal to the possible moral value according to Theory U.

But this is impossible. Since the moral values assigned by T and U are positive linear functions of the number of instantiations of P and Q, respectively, the theories' respective value functions are:

V_T = (# of instantiations of P) · (W) + X

V_U = (# of instantiations of Q) · (Y) + Z

In the first situation, the highest possible value according to T is 100W + X, since the best possible action according to T is the one that causes 100 instantiations of P. By PEMT, this must be equal to the highest possible value according to U. This value is 10Y + Z, since the best possible action according to U is the one that causes 10 instantiations of Q. In the second situation, the highest possible value according to T is 10W + X. If PEMT is correct, this value must be equal to the highest possible value according to U, or 100Y + Z.

But PEMT cannot hold in the second situation if it held in the first situation. Why not? Well, the highest possible value according to T in the second situation (10W + X) is lower than the highest possible value according to T in the first situation (100W + X), because W is a positive number. On the other hand, the highest possible value according to U in the second situation (100Y + Z) is *higher* than the highest possible value according to U in the first situation (10Y + Z), because Y is a positive number. So if the highest possible values according to the two theories were equal in the first situation, they cannot also be equal in the second. PEMT fails.

Now, that example was simplistic in two respects. First, the theories were monistic; each assigned moral relevance to only a single factor – the promotion of some property. Many moral theories, however, are pluralistic; they assign moral relevance to several factors. Second, one of the acts in each situation promoted only what was valuable according to one theory, while the other act in each situation promoted only what was valuable according to the other theory. This is, of course, rarely the case in real life. Lockhart might respond, then, that even if I have shown PEMT to be inapplicable to monistic theories in rather stark choice situations, I have not shown it to be inapplicable to the more complex scenarios in which we typically find ourselves.

Consider, then, a modified version of that example. Suppose my credence is divided between two moral theories. According to Theory T, the value of an action is a positive function of the number of instantiations of some property P that the action causes, and the number of instantiations of some property Q that it causes. Another theory, Theory U, also assigns value to instantiations of P and Q, but weights P less heavily, and Q more heavily, than T did.

We can imagine the two theories' value functions as:

$$V_T = (\#\text{ of instantiations of P}) \cdot (W1) + (\#\text{ of instantiations of Q}) \cdot (W2) + X$$

$$V_U = (\#\text{ of instantiations of P}) \cdot (W1 - A) + (\#\text{ of instantiations of Q}) \cdot (W2 + B) + Z$$

Now imagine two situations. In one situation, I can either cause 100 instantiations of P and 10 instantiations of Q, or 50 instantiations of Q and 5 instantiations of P. In the other situation, I can either cause 100 instantiations of Q and 10 of P, or 50 instantiations of P and 5 of Q.

<u>Situation 1</u>

100 P's + 10 Q's OR 50 Q's + 5 P's

<u>Situation 2</u>

100 Q's + 10 P's OR 50 P's + 5 Q's

It should not be difficult to see how, if PEMT holds in the first situation, it cannot also hold in the second situation. If we take it as given that PEMT holds in the first situation, then Theory U must have more available value in the second situation than Theory T. This is because Theory U weights the production of Q more heavily, and the production of P less heavily, than Theory T does, and there is more Q and less P available in the second situation than in the first.

The lessons of these examples could be applied to still richer cases. All one needs to generate an impossibility result for the PEMT are at least two theories, and at least two scenarios each allowing at least two possible acts. PEMT must hold in the first scenario and in the second. This is impossible unless the difference between the highest possible value in the first situation and the highest possible value in the second situation, according to one theory, is precisely the same as the difference between the highest possible value in the first situation and the highest possible value in the second situation, according to the other theory. But this will almost never be the case.

Still, Lockhart has a response available. I asked you to imagine moral theories as represented by single value functions. But perhaps this is not the only way, or even the best way, to understand moral theories. Instead, moral theories might specify different value functions for different situations. For example, the value a hedonistic utilitarian theory assigns to an action that produces some number of hedons might depend on the situation; it might be situation-relative.

If situation-relativity is possible, then Lockhart can reply to my impossibility arguments as follows (here I imagine his reply to the first, more simplistic, impossibility argument): It is false that the maximum value in Situation 2 according to Theory T must be lower than the maximum value in Situation 1 according to Theory T, simply because fewer instantiations of P may be produced. Similarly, it is false that the maximum value in Situation 2 according to Theory U must be higher than the maximum value in Situation 1 according to Theory U, simply because more instantiations of may be produced. Both theories could, after all, have different value functions corresponding to the different situations. If that's so, then PEMT could hold in both the first situation and the second.

An interesting approach, but still, I think, a multiply flawed one.

First, the mere possibility of situation-relative value is not enough to rescue PEMT. It is not sufficient simply for theories' value functions to vary depending on the situation. They must vary in precisely the way that ensures that PEMT will hold in every situation. But why wouldn't theories' value functions vary in one of countless other ways that are *not* amenable to PEMT? Absent some kind of answer to this question, the defender of PEMT can find help from situation-relativity only by employing it in a suspiciously ad hoc way.

But let's put this worry aside, and assume that theories' value functions vary across situations such that PEMT is preserved. This has some counterintuitive implications. Suppose my credence is divided between Theories T and U. According to T, the rightness of an action in some situation is some positive function of the number of instantiations of P it produces. According to U, the rightness of an action in that situation is a positive function of the number of instantiations of Q it produces. Now, imagine that in that situation, I can either create some instantiations of P and slightly fewer instantiations of Q, or else some instantiations of Q and slightly fewer instantiations of P.

By PEMT, the value according to Theory T of taking the first option must be equal to the value according to Theory U of taking the second option. So far, so good.

Now, suppose I start by believing that I can create 10 instantiations of P, but later come to believe that I can create 100 instantiations of P. That is, I start believing that I'm in one situation, and then come to believe that I'm in another, P-richer situation. All of my other beliefs about the number of instantiations of P and Q that I can produce remain constant, and the rest of my relevant beliefs correspond to the facts as laid out in the previous paragraph.

It's natural to think, "Okay, I thought the situation was such that only a few instantiations of P were possible. Now I think the situation is such that many more instantiations of P are possible. So, according to the P-favoring theory (Theory T), more value should be possible than I'd previously thought." But this sort of thinking is disallowed by the type of situation-relativity that necessarily preserves PEMT. For whatever the situation turns out to be, T's and U's value functions for that situation will be such that the value of the best action according to T must be equal to the value of the best action according to U. So if the situation turns out to be particularly P-poor, T's value function will "expand" so that the best action according to T has as much value as the best action according to U. If the situation turns out to be particularly P-rich, T's value function will "contract" to preserve the same.

We can more easily see how odd this kind of situation relativity is by imagining an agent who's deciding whether to find out how many instantiations of P are possible on option one. If she knows for sure that some act will produce more instantiations of P than any other act, it seems as though she should take no effort whatsoever to find out exactly how many P instantiations this is, since it will have no effect on the maximum value of this act as compared with Theory U's favored act. But this just seems incredible. Suppose she's got some credence in utilitarianism and some credence in deontology, and is deciding whether to kill one person to save X people. Let's stipulate that utility is maximized if X is 2 or greater, and that she knows this. Is it really a complete waste of time for her to find out whether X is 2 or 2 million? Again, tough to believe.

Let me consider one final response available to Lockhart. His formulation of the PEMT states that "the maximum degrees of moral rightness of all possible actions in a situation according to competing moral theories should be considered equal." It's the *in a situation* part that has generated controversy so far. But perhaps PEMT can be modified. Why not instead say that the maximum *conceivable* degrees of moral rightness according to competing moral theories should be considered equal? Call this the "Conceivability PEMT".

This PEMT doesn't suffer from all of the problems of the first, but it suffers from some of them. First, it does seem a bit counterintuitive, in the following respect: There may be some theories, like utilitarianism, according to which there just isn't a maximum conceivable value. If infinite utility is possible, and value is an unbounded function of utility, then an infinite amount of value is possible, too. We might just stipulate that utilitarianism's value function must be bounded, but this gives rise to two problems. First, what could possibly be the argument for setting the bound at one place rather than another? In the absence of such an argument, requiring a bound introduces a significant element of arbitrariness into the proceedings. Secondly, as Frank Jackson and Michael

Smith demonstrate in a recent paper, interpreting theories as bounded value functions leads us to bizarre conclusions.[102] At the very, very least, unbounded utilitarianism is a live option, and one for which the Conceivability PEMT does not allow.

The Conceivability PEMT is just as arbitrary as Lockhart's version – why equalize the maximum and minimum, rather than the mean, two and a half standard deviations from the mean, and so on? If anything, the very possibility of this new PEMT ought to make the original PEMT seem even more arbitrary. Why go with the old version rather than this new one? And insofar as the old PEMT is a viable option, it ought to make <u>this</u> PEMT seem more arbitrary.

This PEMT is, if anything, even more noticeably useless than the original version. Unless one of my possible actions in some situation is <u>the best or worst conceivable action</u> according to one of my theories – and let's face it, when's <u>that</u> ever going to be the case? – the Conceivability PEMT will say nothing about it. Lockhart's PEMT at least had something to say about two actions in every situation for every moral theory.

For what it's worth, the Conceivability PEMT doesn't generate violations of the Independence of Irrelevant Alternatives, and isn't vulnerable to the kind of "impossibility arguments" that I made against the original PEMT.

<u>Ross's Two Proposed Solutions</u>

Jacob Ross also considers the Problem of Value Difference Comparisons, and suggests two solutions.

Let me begin with the latter. Ross begins with a suggestion for how we can

[102] See Jackson and Smith (2006).

compare values within moral theories, and then argues that we can simply extend this proposal and thereby compare value differences between moral theories. Here is his suggestion for intratheoretic comparisons:

> "To say, for example, that according to [Peter] Singer's moral theory, a given amount of human suffering is equally bad as the same amount of animal suffering is to say, among other things, that according to Singer's theory, we should be indifferent between producing a given amount of human suffering and producing the same amount of animal suffering, other things being equal. Likewise, to say that according to the traditional moral theory, human suffering is a thousand times as bad as animal suffering is to say, among other things, that we should be indifferent between a probability of P that a given quantity of animal suffering is produced and a probability of P/1000 that the same quantity of human suffering is produced, other things being equal. In other words, intratheoretic value comparisons can be explicated in terms of claims about what choices would be rational on the assumption that the theory in question is true."[103]

And his parallel suggestion for intertheoretic comparisons:

> "Similarly, we can explicate intertheoretic value comparisons in terms of claims about what choices would be rational assuming that the evaluative theories in question had certain subjective probabilities. Thus, to say that the difference in value between ordering the veal cutlet and ordering the veggie wrap is one hundred times as great according to Singer's theory as it is according to the traditional moral theory is to say, among other things, that if one's credence were divided between these two theories, then it would be more rational to order the veggie wrap than the veal cutlet if and only if one's degree of credence in Singer's theory exceeded .01."[104]

I think this proposal is a non-starter, but we won't be able to see why until we rectify some slipperiness in it. To start with, suppose I have a credence of X that the

[103] Ross (2006), p. 763.

[104] Ibid.

ranking of actions is A, B, then C, and a credence of Y that it's C, B, then A. And suppose further, as Ross does without saying as much, that EOV maximization is the correct theory of rationality. What's it rational for me to do? That depends on the value differences on each side. What are the value differences on each side? That depends, on Ross's proposal, on what it's rational for me to do. This is to say: my credences over the two rankings, along with the correct theory of rationality, don't determine either which action is rational, or what the value differences are. They determine only <rational action, set of value differences> pairs. On some ways of assigning value differences to the rankings just mentioned, A will be most rational; on others, B; on others, C. Going the other direction, if A is the rational action, then the value differences must turn out of these ways; if B is the rational action, one of these other ways; if C is the rational action, one of still other ways.

This means that Ross has a chicken-and-egg problem on his hands.[105] None of A, B, or C is most rational unless we fix the value differences, and no assignment of value differences is correct unless we fix which action is rational. If all we've fixed are credences in the ordinal rankings, and the correct theory of rationality, then the choice between <A, set of value differences #1> and <B, set of value differences #2> is arbitrary. So Ross will have to either use the agent's credences, a theory of rationality, and the agent's beliefs about value differences to fix an action as being rational, or the first two items plus the agent's belief about which action is rational to fix the value differences.

His proposal looks bad either way. He can't go with the first option, because the

[105] See Matthen (2009) for a discussion of this problem's literal analogue.

agent doesn't <u>have</u> beliefs about how value differences compare across hypotheses. That's the source of the PVDC in the first place!

Suppose he goes with the second option: We start with a theory of rationality, an agent's credences regarding rankings, and her beliefs about which actions are rational, and use these to assign value differences that rationalize that action (given that theory and those credences). This will fix value differences, all right, but at a significant cost. If we're simply assigning whatever value differences are required to render rational the action that the agent believes is rational, then it's going to turn out that the agent's beliefs regarding what's rational will always be correct. But if this is the game – I just form beliefs about what's rational, and Ross's method guarantees that I'm right! – then I can only be irrational by failing to do what I think is rational. This is, for one thing, highly counterintuitive. I can make mistakes about morality, about math, about my own name, about almost every rule of rationality, but somehow, not about rationality under uncertainty? Incredible. For another thing, those of us who defend EOV maximization should be a bit exasperated upon discovering that all agents necessarily regard as rational the actions that we say are rational. It makes what we're doing the most purely intellectual exercise possible. And what about our opponents? If we feel bad that everyone necessarily agrees with our recommendations, imagine how they feel, given that everyone necessarily disagrees with their recommendations, insofar as they come apart from ours. The point of all this, of course, is that our method of assigning value differences should leave room for a gap between what it's actually rational to do in some situation, and what an agent in that situation believes is rational. Ross's first method doesn't.

It's worth saying that Ross isn't the only one who faces this problem. The standard way of assigning credences and utilities in decision theory assigns them in such a way that the agent's preferences will necessarily come out as maximizing expected utility. Since the going assumption in decision theory is that maximizing expected utility is necessarily rational, this means that agents will necessarily have fully rational preferences. In support of my approach to the Problem of Value Difference Comparisons, I'm going to be doing in the next chapter something akin to what Ross and the decision theorists are doing here, but as you'll see, my approach will not yield such obnoxious results.

But that's just one of Ross's proposed solutions to the Problem of Value Difference Comparisons. The other solution is somewhat similar to the one I'll develop, but also different in a few ways that tell in favor my mine. Here's Ross:

> "Suppose, for example, that I am uncertain about what is the correct theory of rights. My credence is divided between two such theories, T1 and T2. Suppose, however, that I have a background theory, TB, that evaluates my options in relation to all considerations, other than those deriving from rights. And suppose I am fully confident that this background theory is true. Thus, my credence is divided among two complete ethical theories, the first, which we may call TB+1, consisting in the conjunction of TB and T1, and the second, which we may call TB+2, consisting in the conjunction of TB and T2. Now suppose there is a pair of options, I and J, such that, according to both T1 and T2, no one's rights are at stake in the choice between I and J.... Since no rights are at issue, TB along will suffice to evaluate these options, and so TB+1 and TB+2 will agree concerning their values. Therefore, these alternative ethical theories will agree concerning the difference between the values of these options. We may now define "one unit of value" as the magnitude of this difference. And having thus defined a common unit of value for the two theories, it will follow that so long as we can compare the value intervals within each of these theories, there will be no difficulty

comparing value intervals between the two theories."[106]

This proposal suffers from several limitations and defects. First, it only applies to uncertainty about theories that include information not only about the rankings of actions, but also about the considerations – rights, for example – that give rise to those rankings. But it's entirely possible for one to be uncertain about rankings without having any thoughts about considerations or factors. It's also possible for one to be uncertain about rankings even if one is totally certain about which factors are normatively relevant, and totally certain about the natures of those factors. It is, after all, a further question how considerations should be weighed against each other. Different weighings may give rise to different rankings. So agents can be uncertain about what's better than what without being uncertain about the factors that explain why. These agents seem to fly under the radar of Ross's system, at least as it's stated.

Second, Ross seems to want to require certainty in a background theory, TB, that evaluates actions with respect to all considerations other than rights. I doubt whether very many people will have such a theory. More plentiful, I suspect, are those people who are uncertain about the correct theory of rights, and about whether equality as such is normatively relevant, whether the intention with which one acts affects its permissibility, whether my desiring something is a reason for me to pursue it, and so forth.

Furthermore, there are many features of this system that are left unexplained. This is not a criticism of Ross, necessarily. He's offering only a quick sketch of how to solve the Problem of Value Difference Comparisons. It's just that there are some kinks in his system that would need to be worked out before we could consider it viable.

[106] Ross (2006), p. 764-765.

Here's an example: Consider that there's not just one theory that combines the factors relevant according to TB with the factors relevant according to T1. There are many of them. Some of them regard rights as more important vis a vis other factors; others regard rights as less important. There will be TB+1 (#1), TB+1 (#2), TB+1 (#3), and so on. How do we get the value differences according to T1 from the value differences according to various members of this family of complete theories? My guess is that Ross will want to say something like this: If my credence in TB+1 (#n) is X, then my credence in the value differences according to T1 being the value differences according to TB+1 (#n) value is also X.

This seems like a natural move, but I'm not sure it's defensible. My credence in TB + 1 (#n) may be affected by the way that theory weighs rights against all of the other normatively relevant factors. For example, my credence may be very low in theories that weigh rights considerations very lightly, maybe because I think it's part of the very idea of a right that it outweighs, or trumps other considerations in most cases. But if this feature of a complete theory is what's responsible for my low credence in it, it's not clear why this should impact my credence only in the value differences according to T1 that are implied by those of the complete theory. The point, put very generally, is that there's no reason why my finding a theory implausible because of one of its aspects should mean that I find one of its other, seemingly independent aspects implausible.[107]

Finally, and most importantly, it's not clear that Ross's second approach solves the PVDC at all. Ross claims that the differences between I and J according to TB+1 and

[107] Another example: Classical Deontology includes both the act/omission distinction and the Doctrine of Double Effect. One might find Classical Deontology as a whole very implausible because one finds the act/omission distinction implausible, but I see no reason why this should bear on one's opinion of the Doctrine of Double Effect.

TB+2, respectively, will be equal to one another, by virtue of each being equal to the difference between the two actions on TB. But he gives us no reason to think that the former two differences will each be equal to the latter. Why mightn't the difference between I and J, say, shrink, once we amend the theory so it is concerned with rights as well? Not only is this a possibility; it makes good sense. If existing value differences were preserved as we added more and more morally relevant factors to theories, we'd be left with the result that theories that "cared" about more factors would, in general, have larger value differences. More to the point, why mightn't the difference between I and J be altered in some way once we add T1 to TB, and altered in a different way once we add T2 to TB, such that the I-J difference on TB+1 is different than the I-J difference on TB+2? It might seem odd that this should happen. Perhaps. Later, I'll explain why it might strike us as odd. What matters at the moment, though, is whether Ross's theory can rule this possibility out, and I see no reason to think that it can. And unless it can, it offers no solution to the PVDC whatsoever.

My Own Solution

My view is that Lockhart and Ross each have a part of the truth. Here's the whole truth, in rough outline: Suppose you have some credence in the ranking A, B, then C, and some in the ranking, C, B, then A. We will need to do two things to these rankings before the differences according to the first can be compared to the differences according to the second. First, we'll need to cardinalize each one of them – to impute to agents, on principled grounds, credences not only in the ordinal rankings just mentioned, but in cardinal rankings of the same actions. This is where Ross's first approach has things right, and where Lockhart's approach, insofar as it specifies only maximum and minimum values in situations, falls short.

But cardinalization can't be the whole story. To say that the difference between A and B on the first ranking is 5 times the difference between B and C on the first ranking, and that the difference between A and B is 3 times the difference between B and C on the second ranking is to say *nothing* about how either difference on the first ranking compares rationally to either difference on the second ranking.

So we'll also need to *normalize* the rankings. Once we give the rankings cardinal structure, we'll also need to put them on a “common scale”, you might say, so that we can compare differences between them. This is where Lockhart, whatever the merits of his execution, has got things right. As I noted at the end of my criticism of Lockhart, we could normalize the rankings by setting the best conceivable action according to each of them equal to the best conceivable action according to the others; *mutatis mutandis* for the worst action. This strategy is flawed and, as it stands, incomplete, but it's at least the *sort* of thing that'll have to be done.

In the sections to follow, I'm going to defend a *cardinalizing* proposal, and then a *normalizing* proposal. For reasons having to do with the nature of normative concepts, the latter will of necessity be less precisely developed than the former, which will unfold slowly throughout the remainder of this chapter and the entirety of the next.

Cardinalization

Some of the specifics of the cardinalizing proposal will have to wait until the next chapter, but its essence is simply Frank Ramsey's method of assigning value differences through probabilities.[108] Suppose acts A, B and C are ranked ordinally in that order. And suppose I believe that an act that has a .2 chance of being an instance of A and .8 chance of being an instance of C has the same value as an act that has a 1.0 chance of being an instance of B. Then we can impute to me a belief that the difference between A and B is 4 times the difference between B and C. That's because the former difference would have to be 4 times the latter in order for the first act, given its probabilities of being an instance of A and C, respectively, to have the same expected value as the latter.

Here are two examples of an action's having a chance of being an instance of other actions. Suppose I believe that P, but I tell you that ~P. Then my act has some chance of being an instance of telling a lie – if P is indeed true – and some chance of being an instance of something other than a lie – if ~P turns out to be true. Or suppose I throw a stick of dynamite down a mineshaft. My act has some chance of being an instance of killing – if there are miners in the shaft, perhaps – and some chance of not being an instance of killing. To say that an act has a chance of being an instance of another act is not, however, simply to say that the act has a chance of producing some outcome. For it can be less-than-certain whether an act possesses some feature, even if that feature is one that we'd naturally think of as part of "the act itself", rather than as part of its consequences.

Again, this will all require much more elaboration, but that will take place in the next chapter. Hopefully you get the basic idea for now.

How can we use this method to cardinalize rankings among which we're

[108] See Ramsey (1926).

uncertain? There are, broadly speaking, two types of ways. First, there are ways that involve counterfactuals; second, there are ways that don't involve counterfactuals. Of the ways that don't involve counterfactuals, one way relies on shared cardinal rankings among hypotheses in which the agent has credence, and another way relies on some hypotheses being "indifferent" among actions regarding which other hypotheses are not indifferent.

Let me describe each of these in turn, starting with the Counterfactual Method. Suppose I am uncertain between two rankings of actions A through Z; call them Ranking 1 and Ranking 2. We can say that my difference in value between A and D on Ranking 1 is 3 times my difference between D and M on Ranking 1 if, were I certain of Ranking 1, I would believe that an action that had a .25 chance of being an instance of A and a .75 chance of being an instance of M was equal in value to an action that was certain to be an instance of D. On this approach, I just put to one side all of the rankings except the one I'm trying to cardinalize, and determine what the agent's beliefs would be regarding "probabilistic acts" of the sort described, if all of his credence were allotted to the ranking in question.

There are two potential difficulties with this approach. The first is that counterfactuals may, as a metaphysical matter, be indeterminate, in which case the ratios of value differences according to the cardinalized ranking will also be indeterminate. The second is that, indeterminacy aside, it might be difficult for anyone, even the agent, to evaluate the relevant counterfactuals. Answering the question, "If you were certain of some normative hypothesis of which you're in fact uncertain, what would be your beliefs about the comparisons of probabilistic acts on that hypothesis?" may only be possible when the hypothesis is a comprehensive theory like utilitarianism, for which a particular cardinalization readily suggests itself. (This cardinalization is one on which the ratios of value differences between actions are simply the ratios of utility differences between the outcomes produced by those actions.) It is of particular importance that the agent herself have epistemic access to her cardinalized ranking, or else she will be unable to deliberate in accordance with the theory of rationality I've been suggesting.

Another proposal, the Shared Ranking Method, depends on shared cardinal rankings, and does not involve counterfactuals. Here a more concrete case may help. Suppose a woman is uncertain about whether it's better to have an abortion or not to have one. So she has some credence in a ranking that ranks abortion above non-abortion, and some credence in a ranking that ranks non-abortion above abortion. It is perfectly consistent with this, and perhaps expected even, that she is certain regarding the ordinal rankings of other actions. (Along the same lines, it's to be expected that the pro-choicer and pro-lifer will agree about how lots of actions rank vis a vis one another, even though they disagree about such a salient issue.) The former ranking might look like this: A, B, C, D and abortion tied, E, F, G, H and non-abortion tied, I, J, K. The latter ranking might look like this: A, B, C, D, E, F, G and non-abortion tied, H, I, J and abortion tied, K. In such a case, uncertainty about abortion stands amidst large swaths of certainty. (See Diagram 4.1)

Diagram 4.1: The Shared Ranking Method, #1

Ranking 1	Ranking 2
A	A
B	B

C	C
D, abortion	D
E	E
F	F
G	G, non-abortion
H, non-abortion	H
I	I
J	J, abortion
K	K

In such a case, the agent might also have the beliefs regarding probabilistic acts required to impute cardinal rankings to her. She might, for example, believe that an action that has a .1 chance of being an instance of B and a .9 chance of being an instance of J is equal in value to an action with a 1.0 chance of being an instance of H, in which case she can be represented as believing that the difference between the first two acts is 9 times the difference between the latter two.

If acts about which the hypotheses disagree are equal on those hypotheses to acts about which they agree, then it will be possible to cardinalize even with respect to the controversial acts. For example, since abortion is tied with J on the second hypothesis, if we've managed to cardinalize with respect to J using the shared ranking method, we've thereby managed to cardinalize with respect to abortion on the second hypothesis. So perhaps contrary to initial appearances, the shared ranking method can be used to rank cardinally, on each hypothesis, even those actions of whose relative position I'm uncertain. (See Diagram 4.2)

Diagram 4.2: The Shared Ranking Method, #2

	Ranking 1	Ranking 2
___	B	B
9x		
___	H, non-abortion	H
1x		
___	J	J, abortion

There's no need, as there was on the Counterfactual Method, to ask what my cardinal rankings of actions would be were I certain of this ordinal ranking or that ordinal ranking. That's because the Shared Ranking Method applies to cases in which the cardinal ranking I would have were I certain of *any* ordinal ranking is just the same cardinal ranking I have in the actual world. So we can avoid considering difficult-to-evaluate counterfactuals, and simply impute to me a cardinal ranking in the actual world.

If the actions on the shared cardinal ranking divide the range of values very finely, then chances will be good that, for any controversial act on any hypothesis, it will be equal to some or other act on that shared cardinal ranking. It helps matters that every action falls under multiple descriptions, so there are plenty more action types to place on a shared cardinal ranking than there are action tokens. This method, then, has the potential to be a very powerful tool.

There is one more proposal that does not require inquiring into counterfactual scenarios – the *Lone Ranking Method*. Suppose I am uncertain between several rankings of the actions A, B, and C. According to *all of the rankings but one*, A, B, and C are equally ranked. Perhaps A, B, and C are, respectively, going to see a Mike Kelley exhibit, going to see a Tim Hawkinson exhibit, and going to see a Donald Judd exhibit, and I enjoy all three artists' work equally. But let's suppose I have some credence in a theory that valorizes cool impersonal simplicity and abhors its opposite. That theory will say that C is better than B, which is better than A. If that is the only theory according to which the actions are not equal, then my judgments about rankings of actions in general will also serve as judgments about rankings according to that theory. If, for instance, I regard as equally valuable an action with a .5 chance of being an instance of A and a .5 chance of being an instance of C, and an action with a 1.0 chance of being an instance of B[109], then I can be represented as believing that the difference between A and B is equal to the difference between B and C – not only in general, but also on the condition that the "cool impersonal" theory is true, since that theory is the only one that "cares" about this situation. (See Diagram 4.3)

Diagram 4.3: The Lone Ranking Method

	Ranking 1	Rankings 2, 3, 4...
___	Judd	
1x		
___	Hawkinson	Judd = Hawkinson = Kelley
1x		
___	Kelley	

The degree to which cardinalizing will prove successful will depend,

[109] Perhaps the former act involves going to an exhibit the content of which is determined by a coin flip.

unsurprisingly, on the agent's psychology. If the counterfactuals to which the first method adverts are indeterminate, if the rankings among which an agent is uncertain are not amenable to a shared cardinal ranking, and if there are no cases of actions being differently-valued on only a lone ranking, then cardinalization will be difficult. The hope, however is that at least one of these methods will apply. There may also be, independently of doubts about whether the agent's mind is "structured" in the optimal way, doubts about whether the probabilistic method of cardinalizing that I briefly sketched is appropriate. These will be dealt with at length in the next chapter.

Normalizing

That there are principled ways to normalize competing rankings is much less obvious than that there are ways to cardinalize the rankings, for it can seem that there are no mental features as clearly helpful for the former task as beliefs about probabilistic acts were for the latter. Beliefs about probabilistic acts on the condition that, say, Kantianism is true, will help to cardinalize that theory – and again, I'll show precisely how in the next chapter – but where in one's mind might we find the resources to compare the differences on that cardinal ranking with those on a cardinalized version of Bentham's theory? In this section, however, I'll try to show that this pessimistic judgment is premature, and that in fact we can find features of agents' minds that will permit normalization.

Let me start by noting a procedure that won't work, but that I once believed would.[110] I cannot normalize rankings simply by applying the Shared Ranking Method

[110] See Sepielli (2009).

discussed above.[111] Suppose, e.g., that my credence is divided between two comprehensive rankings that disagree about the relative positions of various actions. However, they agree that G is better than H, which is better than I. Furthermore, suppose I'm certain that an act with a .2 chance of being an instance of G and a .8 chance of being an instance of I is equal in value to an act with a 1.0 chance of being an instance of H. I've shown, then, that according to each ranking, the difference between the first two acts is 4 times the difference between the second two. What I have <u>not yet</u> shown, however, is anything about how the value differences on one ranking compare to the value differences on the other. I have not shown, for instance, that the difference between G and H on the <u>first</u> ranking is 4 times the difference between H and I on the <u>second</u> ranking. It is possible for two rankings to "agree" that one value difference is X times another, even if the first difference and second difference on one ranking are much, much larger or smaller than their counterparts on the other ranking.

I think that I assumed this method would work because I was focusing my attention on cases in which competing rankings could be normalized through one of the methods I'll survey in a moment, and indeed, I was applying one or more of these methods without realizing it. This blinded me to the fact that shared cardinal rankings guaranteed just that – shared cardinal rankings – and took us no closer to normalization. Let me now turn to some methods through which normalization is possible.

The key to normalizing is getting competing rankings on a "common scale", and the key to getting rankings on a common scale is exploiting their common features – the

[111] Thanks to Ruth Chang and Toby Ord for helping me to see this.

features of normative concepts as such. Normative concepts have rich conceptual roles, and the our beliefs involving them have complex functional profiles.[112] After all, my beliefs about rankings of actions are not simply related to my beliefs about probabilistically-described actions. They are also constitutively related to attitudes like blame, praise, regret, and relief, to my motivation to act, and even – although this is more controversial – to my beliefs involving non-normative concepts.[113] For example, it may be that part of what it is for me to believe that dancing is wrong is for me to be disposed to blame for dancing, for me to be motivated not to dance, for me to regret dancing if I end up doing it, and perhaps for me to think that dancing has the same general normative status as certain paradigmatically wrong acts like hurting someone's feelings without cause.

All of this means that the "sparsely populated" mind to which I alluded at the beginning of the chapter – the one that has only credences in the different competing rankings of actions – is not only unlikely, but is necessarily chimerical. We can never *just* have normative beliefs and nothing else, for it's a condition on our having normative beliefs that we have the other mental states to which these beliefs are constitutively connected.

To see how these connections might help the cause of normalization, let's explore some possible methods of normalizing that involve the relationship between normative beliefs and the attitude of blame. Much of what follows might also be applied, *mutatis mutandis*, by substituting other connections in place of the normative belief-blame

[112] Thanks to Gustaf Arrhenius for helpful discussions about the material here.

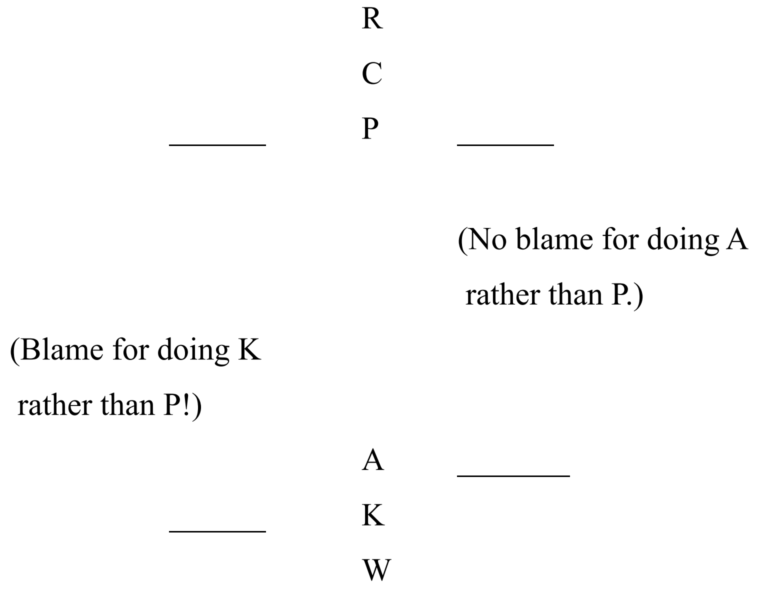
[113] This last relation is postulated, most notably, by "analytical descriptivists" like Jackson (1998), and their close cousins, "thickened" conceptual role semanticists about normative terms like Peacocke (2004).

connection, although I won't explore all of these here.

Here's one idea, based on the Counterfactual Method above: Suppose I am uncertain between two rankings of the actions A through Z – Ranking 1 and Ranking 2. Suppose it is the case that, were I certain of Ranking 1, I would blame someone for doing K rather than P, if those were the only two actions available, but that I would not blame someone for doing the less valuable of two actions that were closer together according to that ranking. The K-P difference marks the Blame Interval, you might say, for Ranking 1. (See Diagram 4.4) Suppose now that, were I certain of Ranking 2, I would blame someone for doing Q rather than S, if those were the only two actions available, but that I would not blame someone for doing the less valuable of two actions that were closer together according to that ranking. The Q-S difference marks the Blame Interval for Ranking 2.

Diagram 4.4: Blame Intervals

If I were certain of Ranking 1...



R

C

P

(No blame for doing A rather than P.)

(Blame for doing K rather than P!)

A

K

W

G

Mutatis mutandis for Ranking 2...

The relation between normative judgment and blame is not something that it makes sense to say varies from ranking to ranking. It is a feature that depends on the role in thought of normative concepts as such. Insofar as we say that my tendency to blame someone for doing an act is conceptually tied to the degree by which I believe that act falls short of the best act available, then two "blame intervals", as I'd called them above, must be of the same size. So the K-P difference on Ranking 1 must be equal to the Q-S difference on Ranking 2. Once we've fixed these differences as equal and cardinalized the rankings through one of the methods suggested above, we can compare other value differences across rankings. If, for example, a value difference on Ranking 1 is 6 times the K-P difference on that ranking, it must 6 times the Q-S difference on Ranking 2.

There is also a normalizing method that's somewhat similar to the Shared Ranking Method above. We might call it the Invariance Method. Suppose I have some credence that A is better than B, and that B is better than C; and some credence that A is better than C, and that C is better than B. Suppose further that I have a disposition of strength S to blame for doing B rather than A, and that this strength would remain constant regardless of my credence distribution between the two rankings. If I became certain of the first, I'd have a disposition of strength S to blame someone for doing B rather than A. The same goes if I became certain of the second, or if my credence distribution shifted to 75/25, 25/75, or 30/70, and so forth. In that case, the strength of my disposition to blame for doing B rather than A is invariant with respect to my credences in the rankings, and so the size of the value difference to which this strength of disposition is conceptually tied is the same on both rankings, which means they've been normalized. The A-B difference on the first ranking is equal to the A-B difference on the second ranking. (See Diagram 4.5)

Diagram 4.5: The Invariance Method

Credences:	0	1	…..	.43	.57	…..	1	0
______	A	A		A	A		A	A
(Disposition of strength S to blame for doing B!)	C			C			C	
______	B	B		B	B		B	B
		C			C			C

The procedures just outlined depend, of course, on a particular view about how blaming is tied to beliefs about value differences. They rely on a characterization of a value difference's being a certain size on which to believe this is to be disposed to blame someone for doing the act on the bottom end of that difference. This is a way of giving a meaning to the concept of a value difference's being of that size – not by giving a descriptive definition thereof, but by giving an operational definition of the role of the concept in thought. On this definition, the concept is the one such that you count as believing that a value difference falls under it just in case your dispositions to blame work in the way specified.

This is, to be sure, a very simplistic picture of the semantics of normative concepts. For one thing, the relationship between blame and value differences might be different, and for that matter, more complicated, than what I've sketched here. Rather than there being a smallest blame interval, as I suggested in bringing out the first approach, it may be that both the strength of my disposition to blame for doing Y rather than X, and the degree to which I blame for doing Y rather than X, are related complexly to the degree to which I believe X is better than Y. Or it may be that what matters is not how X and Y compare to each other, per se, but how each compares to a salient set of alternatives, whether possible in the current situation or not. Or perhaps the constitutive relationship is not between blame and value differences in general, but between blame and my beliefs about how actions compare in terms of a specific kind of value – moral value, say, or value that depends on our behavior's effects on others.

Or, as I said earlier, it's probably appropriate to see normative concepts as tied not only to blame, but to other things as well. Perhaps we should think of my degree of motivation to do X rather than Y as constitutively related by some positive function to the degree to which I believe X is better than Y. Or here is an approach that I find attractive but that other conceptual role semanticists like Ralph Wedgwood, for example, might object to:[114] There might be some acts that I think of as "paradigmatically okay" – in the strict sense that they serve as paradigms for the concept OKAY: going to the store on an ordinary day, listening to music, and so forth. There might also be some acts that I think of as "paradigmatically heinous": killing for fun, for example. Because these non-normatively described acts are paradigms for the associated normative concepts, the okay-ness of going to the store and the heinousness of killing for fun will be invariant across rankings in which I have credence. After all, I can't have any credence in a hypothesis according to which a concept's paradigm fails to fall under it; that wouldn't

[114] Wedgwood (2001) favors a "thin" CRS for normative terms on which the only concept-constitutive connection is between OUGHT and motivation.

count as a belief involving that concept. Now that we've given the meanings of OKAY and HEINOUS through their relationships to ranking-independent features of my psychology – specifically, to beliefs involving non-normative concepts, we can say that the "okay/heinous" difference on any ranking is the same size (or at least, roughly the same size) as the "okay/heinous" difference on any other ranking.

Anyhow, these are suggestions about how we might use the conceptual roles of normative concepts to normalize rankings. Chances are that, once we get the psychology and semantics all sorted out, the relationships between normative concepts and other attitudes, behaviors, feelings, etc. will turn out to be enormously complex. But because it would be difficult to state my approach in terms of such complex relationships, and because our understanding of normative concepts is still at such an early stage, it's best to present my suggestions in terms of simpler constitutive relationships.

Normalizing rankings by appealing to non-ranking features of the agent's mind is analogous to something that might be, but rarely is, done in welfare economics – normalizing differences in well-being across individuals by appealing to features other than those individuals' preference rankings. If, for example, we took it as constitutive of well-being how well one fared relative to an index of objective goods, or how positive one's subjective hedonic state was, the problem of interpersonal comparisons of well-being would disappear. To use a cartoonish example, we might say that the difference in well-being between living George Plimpton's lifestyle and living Philip Larkin's lifestyle counts as the same for everyone. We could then use this fixed-size difference in well-being to normalize different people's rankings. It's my contention that we can make the

Problem of Value Difference Comparisons disappear if we take as constitutive of my having certain beliefs about value differences that I have functionally-associated beliefs about non-normative features of actions, or that I am disposed to blame in such-and-such situations, or that my motivations are structured in such-and-such a way. As long as we simply stick with beliefs about rankings in trying to solve the Problem, though, we'll be out of luck, in the same way welfare economists will not be able to solve their problem without dropping the simple preference-based view of well-being.

My normalizing proposal differs from Lockhart's essentially in that he establishes value difference comparisons across hypotheses by fiat, while I do not. We might wonder why he does this, given that his proposal gives rise to so many problems. (Recall that when PEMT is applied in the way he suggests, it leads to contradictions and grievous violations of the Independence of Irrelevant Alternatives; when it is applied not at the level of particular situations, but to all conceivable acts, it becomes all the more arbitrary, and also runs into problems with normative theories that seem to admit of no upper bound of value.)

Why, then, does he resort to normalization-by-fiat? My suspicion is because, like Ross, like my past self, and like most others whose work on this topic is influenced by decision theory,[115] he sees no way of normalizing in an antecedently meaningful way. Since there is no such way of giving content to the notion of a value difference on utilitarianism bearing some ratio to a value difference on Kantianism, we must simply offer stipulations about how value differences compare across hypotheses. If normative

[115] See, e.g., unpublished work by Bostrom and Ord.

concepts' roles involved only other normative concepts, and if normative beliefs were only constitutively connected to other normative beliefs, Lockhart might be right. If we can't break out of what Allan Gibbard (2003) calls “the normative web”, there is no way to normalize rankings. But we can break out of the normative web – if not so far as to explicitly define normative terms in non-normative terms, then at least to give the connections among mental states that are constitutive of normative concepts and normative beliefs. It's not utterly meaningless to say that abortion is as bad on the “pro-life” hypothesis as murder is on either the pro-life or pro-choice hypotheses, or that letting your poor relative drown in the bathtub is as “sick”, “cold-blooded”, or “messed up” on the consequentialist hypothesis as pushing him into the bathtub is on either that hypothesis or the deontological one.

To be sure, my exploitation of connections like these makes my normalization procedures much messier than Lockhart's procedure. Just as it's simpler to stipulate a definition for a term than to find out the antecedent meaning of a term, it's simpler to stipulate the relative sizes of value differences across normative hypotheses than to determine them using the actual constitutive roles of normative concepts. As a result, I don't have a quick-and-easy theory of what it is for a difference in value on one normative hypothesis to be, say, equal to a difference in value on another one. But in a sense, I shouldn't have a theory like this – nobody should – for such a thing could only be justified by a satisfactory account of how beliefs involving all the different normative concepts are functionally related to blame, motivation, and all the other mental features I suggested. Such an account is (at least what I would call – others may demur) a psychologically realistic conceptual role semantics for all the different normative concepts, which is something that hasn't yet been developed. In the meantime, though, I hope what I've provided here has convinced you that the principled normalization of competing normative hypotheses makes good sense, in much the same way that the principled normalization of different peoples' scales of well-being makes good sense.

Conclusion

In order to solve the PVDC, we need to do two things – first, we need to impute a cardinal structure to each of the rankings about which an agent is uncertain. Second, we need to normalize those rankings so that it's possible to compare differences across them. I've suggested that, in order to do the second of these, we need to appeal to features other than normative beliefs. We need to appeal to features like the dispositions to praise and blame that are, as a matter of conceptual constitution, linked to normative beliefs, or perhaps more controversially, to beliefs about non-normative features of actions that are similarly conceptually linked to normative beliefs. It's impossible to give a precise recipe here without settling on a view about which links between other mental states and normative beliefs are concept-constitutive and which aren't, and about the precise ways in which the former are constitutive of particular normative concepts. I can only say at this point that it makes good sense to us to say that, for example, the relative importance of not murdering is the same across the pro-choice and pro-life hypotheses, and that we can explain this sense in roughly the manner I've been suggesting. We ought to start with the assumption that certain commonsensical thoughts are meaningful, and then search for more rigorous ways of giving their meanings.

We can cardinalize using agents' beliefs regarding probabilistic acts, in much the way Ross suggests in his first proposal. His proposal was problematic, though, in that it delivers the result that agents' beliefs about what's rational to do under uncertainty are necessarily correct. My proposals don't have that upshot. I'm able to cardinalize agents' normative rankings without relying on their beliefs about what's rational under normative uncertainty. However, the proposal as stated does have the implication that agents' beliefs about what's right in the non-normative belief-relative sense are always correct. That is, I assume for interpretive purposes that expected value maximization under non-normative uncertainty is correct, and that the value differences on each of the normative hypotheses are whatever they have to be in order for agent's beliefs about the equality of certain probabilistically described acts to come out true. This improves on Ross in one way, but is still not entirely satisfactory. The next chapter, which is all about cardinalization, will more carefully lay out my views on cardinalization, and will enable us to do away with this interpretive assumption.

CHAPTER FIVE: A METHOD OF CARDINALIZATION

Introduction

Some normative hypotheses seem readily amenable to cardinalization. For example, it's natural to think that, according to classical utilitarianism, the difference between actions A and B is 4 times the difference between actions B and C just in case the difference between the utility produced by A and the utility produced by B is 4 times the difference between the utility produced by B and the utility produced by C. Some normative hypotheses do not obviously lend themselves to cardinalization. Nothing in Kant's corpus, for example, suggests a way of representing Kant's moral theory by a cardinal ranking. The aim of this chapter is to show that such cardinalization is possible over a wider range of normative hypotheses than you might have expected. The first few sections provide my positive cardinalizing proposal. The later sections modify the proposal in response to objections and highlight some of its less obvious merits. What we'll be left with is a way of giving content to the idea of a cardinal scale of reason strength that is richer and more plausible than any other method developed so far.[116]

Ramsey as a Template

My account finds its roots in Frank Ramsey's paper "Truth and Probability" – specifically, in Ramsey's method of deriving differences in value between states of affairs, or "worlds". I'll begin by introducing you to Ramsey's method; I'll then modify it in a piecemeal way until we're left with my own method of defining differences in value

[116] This chapter has undergone so many revisions that I can't remember whose suggestions led to which changes. I do remember comments from Lara Buchak, Ruth Chang, Holly Smith, Brian Weatherson being especially helpful.

between actions. Here's a succinct statement of the core of Ramsey's Method:

> "...we define [the difference in value between worlds A and B being equal to that between worlds C and D] to mean that, if p is an ethically neutral proposition believed to degree ½, the subject has no preference between options 1) A if p is true, D if p is false, and 2) B if p is true, C if p is false."[117]

An "ethically neutral" proposition, says Ramsey, is one such that two possible worlds differing only in regard to the truth of that proposition are necessarily of equal value.[118] In other words, neither the truth nor the falsity of p in some world contributes to or detracts from the value of that world. So we needn't worry about, say, the difference between A and D with respect to the truth value of p "skewing the results" of the stated method. A fair coin will come up heads the next time it's flipped is an example of an ethically neutral proposition.

Now, the option A if some ethically neutral proposition with probability .5 of being true is true, D if that proposition is false is equivalent to the option a .5 probability of A, and a .5 probability of D, assuming A and D are mutually exclusive. So for ease of exposition, let's take the following restatement of Ramsey's method as our jumping-off point:

> Ramsey's Method Restated**:** The difference in value between worlds A and B is equal to the difference in value between worlds C and D =df. The subject has no preference between options 1) a .5 probability of A, and a .5 probability of D, and 2) a .5 probability of B, and a .5 probability of C.

[117] Ramsey (1926), p. 33.

[118] Ibid.

To see the appeal of a method like Ramsey's, consider an example. I'm given the choice between two lotteries. If I choose Lottery 1, a fair coin is flipped. If the coin lands heads, I win a healthy Saint Bernard puppy. If the coin lands tails, I win a basket of strawberries. If I choose Lottery 2, a fair coin is flipped. If it lands heads, I win a healthy Irish Wolfhound puppy. If it lands tails, I win a basket of blueberries. Ramsey says that, if I'm indifferent between these two lotteries, then the difference in value, for me, between the world in which I get the Saint Bernard and the world in which I get the Irish Wolfhound must be equal to the difference in value between the world in which I get the blueberries, and the world in which I get the strawberries. And this makes sense. Suppose I like Saint Bernards better than Irish Wolfhounds; then I have to like blueberries better than strawberries by exactly the same margin in order to balance this out, and render the second lottery just as good as the first.

There are two important things to note about Ramsey's method. First, the value differences here are not absolute values of value differences; they take into account "direction" as well as "magnitude". So, for example, if the difference between A and B is 20, the difference between B and A is not also 20. Rather, it is -20. Without this stipulation, Ramsey's method and my subsequent revisions of it will seem not to make sense. To see this, have another look at Ramsey's Method Restated. Suppose A is better than B, and D is better than C. The magnitude of the difference between the first two might be equal to the magnitude of the difference between the second two, but clearly the agent will prefer 1) to 2). If the measure of a difference were just its magnitude, this case would be a counterexample to Ramsey's method. But direction also matters. Therefore, the difference between D and C is not the same difference as the difference between C and D, and so there is no counterexample.

Second, Ramsey is not merely giving us a way of *measuring* differences in value between worlds; rather, he is giving a *stipulative definition* of when two differences in value are equal. We are of course free to reject his definition in favor of another one, but if we accept it, there is no room to say that an agent is indifferent in the way specified by the method, but does not have value differences between worlds. If you’ve got these preferences, you’ve got the corresponding value differences, whether or not you’re fond of the locution “value differences”. Later in the chapter, though, I will discuss the possibility of rejecting a Ramseyan definition in favor of some alternative.

From Worlds to Actions

I’m concerned with differences in value between *actions*, not *worlds*, so I’ll need to modify Ramsey’s method a bit to fit my needs. Let’s start by dividing actions into two types – *prospect actions* and o*utcome actions*. Prospect actions play roughly the role in my scheme that options played in Ramsey’s, and outcome actions play roughly the role that worlds played. A prospect action is an action under a probabilistic description; it has some subjective probability of being an instance of one or more outcome actions. For example, the prospect action that has an X probability of being an instance of killing 1 person, a Y probability of being an instance of killing 10 people, and a Z probability of being an instance of killing 100 people, has X, Y, and Z probabilities, respectively, of being an instance of the outcome action *killing 1 person*, the outcome action *killing 10 people*, and the outcome action *killing 100 people*. *Detonating a bomb in a city plaza* is an action like this; it is potentially deadly, but one is unsure how many people, if any, will be killed. We may modify Ramsey's method accordingly:

> *Ramsey's Method (Action Version)*: The difference in value between outcome actions A and B is equal to the difference in value between outcome actions C and D =*df*. The subject has no preference between the prospect actions 1) an action with a .5 probability of being an instance of A, and a .5 probability of being an instance of D, and 2) an action with a .5 probability of being an instance of B, and a .5 probability of being an instance of C.

Let's introduce some variations on this approach. Sometimes, we'll have just three outcome actions – A, B, and C – and we'll want to know how the difference between A and B compares to the difference between B and C. The abortion example from the previous chapter was a "three act" case; the three actions were *killing an adult human being*, *bearing a child*, and *using an innocuous form of birth control*.

The approach I'm developing can handle "three act" cases, for we can simply treat two of "A", "B", "C", and "D" as naming the same action, and combine the probabilities that go along with the two names. So, in the abortion example, "A" might name *killing an adult human being*, "B" and "C" might both name *bearing a child*, and "D" might name *using an innocuous form of birth control*. We can add the probabilities of B and C together to get the probability attached to bearing a child. The point can be generalized as follows:

> Ramsey's Method (Three-Action Version): The difference in value between outcome actions A and B is equal to the difference in value between outcome actions B and C =df. The subject has no preference between the prospect actions 1) an action with a .5 probability of being an instance of A, and a .5 probability of being an instance of C, and 2) an action with a 1.0 probability of being an instance of B.

We'll also want to compare value differences when they're unequal, as of course they were in the cases from Chapter 4. There are two ways of deriving unequal value differences.

The first is similar to the method Richard Jeffrey credits to Ramsey in Jeffrey's The Logic of Decision.[119] It involves repeated applications of the Three-Action Version. After one application, we will have three actions placed on cardinal scale – A, B, and C – with the difference between A and B equal to the difference between B and C. In other words, B will lie at the "midpoint" of A and C. Next, we will find some other outcome action, D, such that the subject has no preference between prospect actions 1) an action with a .5 probability of being an instance of A, and a .5 probability of being an instance of B, and 2) an action with a 1.0 probability of being an instance of D. The difference between A and D will then be equal to the difference between D and B; D will lie at the midpoint between A and B. Then, we can find some other outcome action, E, such that the subject has no preference between prospect actions 1) an action with a .5 probability of being an instance of A, and a .5 probability of being an instance of D, and 2) an action

[119] See Jeffrey (1990), Chapter 3.

with a 1.0 probability of being an instance of E. The difference between A and E will equal the difference between E and D; E will lie at the midpoint of A and D. (See Diagram 5.1.)

Diagram 5.1: Illustration of Jeffrey's Method

←Better Worse→

A E D B C

What we're doing here is determining differences in value between actions, chopping those differences in half, chopping those differences in half, and so on. As we do so, we place more and more actions on the cardinal scale, with differences of varying sizes between them.

The second way involves an extension to, rather than the repeated application of, the Three-Action Version. It requires the assumption that the agent is an expected value maximizer. On this assumption, the values assigned to the outcome actions will be whatever they have to be in order for her preferences among prospect actions to line up with their expected values – for her to prefer one prospect action to another just in case the first has a higher expected value than the second. This means that, if the ratio of the probabilities attached to A and C in the first prospect action is X:Y,[120] then the ratio of the difference in value between A and B to the difference in value between B and C must be Y:X. This is what's required in order for the two prospect actions to have equal expected values, and hence, to be equally preferred. (Or rather, it must be Y:X if A is better than B,

[120] To take a specific case, this ratio is 3:1 when the first prospect action is an action with a .75 probability of being an instance of A and a .25 probability of being an instance of C, since .75 divided by .25 is 3, which is expressed in rational form as 3:1.

and B is better than C. For the subject may also have no preference between two prospect actions like those mentioned in the Three-Action Version if A, B, and C are all of equal value, in which case the differences between A and B and between B and C will both be zero rather than Y:X.) Thus:

> Asymmetric Three-Action Version: The difference in value between outcome actions A and B is Y:X times the difference between outcome actions B and C =df. The subject has no preference between prospect actions 1) an action with an X probability of being an instance of A, and a Y probability of being an instance of C, and 2) an action with a 1.0 probability of being an instance of B, and the agent prefers A to B and B to C.

A few words about the assumption that agents are expected value maximizers: In at least one strand of contemporary decision theory, it's intended to be an interpretive tool rather than an empirical thesis.[121] There is not, on this view, some other way of assigning value, the expectation of which we're blithely assuming that everyone will maximize. The assumption is our basis for assigning values in the first place; values, as they're understood in this context, just are whatever they have to be in order for expected value to be maximized. So this assumption has not been as controversial as we might otherwise expect.

All the same, it's plausible there may be other ways of assigning values that don't require the expected value maximization assumption or anything like it. This possibility

[121] For further explication of this point, see Hurley (1989), Broome (1995) and Maher (1993).

will be explored later in the chapter. For now, though, we will proceed on the assumption that agents are expected value maximizers under non-normative uncertainty, and that values are as assigned using this assumption.

Before concluding this section, I should alert you to a method of deriving value differences that will *not* work. Suppose an agent is indifferent between two prospect actions: 1) an action with an X probability of being an instance of A, and a Y probability of being an instance of D, and 2) an action with a W probability of being an instance of B, and a Z probability of being an instance of C, with the limitation that neither X = Y nor W=Z, and neither "A" and "D" nor "B" and "C" name the same act. For whatever values we assign to W, X, Y, and Z within this limitation, there are several different sets of values we could assign to A, B, C, and D such that the EOV of the first prospect action will be equal to the expected value of the second prospect action (i.e. such that (X)·(Value of A) + (Y)·(Value of D) = (W)·(Value of B) + (Z)·(Value of C)). And, what matters for our purposes, the ratio of the difference between A and B and the difference between C and D will depend on which of those sets of values we assign. Thus we can't pin down the relative sizes of the value differences just by filling in the probabilities. So while there's an *Asymmetric Three-Action Version,* there's no *Asymmetric Four-Action Version*.

From Preferences to Normative Judgments

Now for a more fundamental modification of the Ramseyan method. The backbone of Ramsey's approach is *preference*. We've been determining differences in value between outcome actions by looking at agents' preferences – or really, their lack of preference – between prospect actions. Decision theorists typically define the preference of A over B as the disposition to choose A over B.[122] A disposition to choose is not a cognitive attitude regarding the values of the objects of choice.[123] There's a difference between saying someone is disposed to choose vanilla over chocolate, and saying that someone thinks vanilla is better than chocolate. More pointedly, there's a difference between saying that someone is disposed to shoot heroin, and saying that someone thinks shooting heroin is good.

My aim, on the other hand, is to impute beliefs about cardinal rankings, and it's hard to see how these can be derived from mere preferences. Instead, I'll impute these beliefs via the following two-step process. First, I'm going to give an account of what cardinal evaluative rankings are – of what it is for the difference in value between two actions to stand in such-and-such a ratio to the difference in value between two other actions. So I'll be heading towards a theory of the form:

> For the difference in value between actions A and B to be X times the difference in value between actions C and D is for P.

The second step will take us from an account of what it is for a cardinal ranking to

[122] See Maher (1993) for a helpful survey of the different definitions of "preference", including Maher's own norm-expressivist one, inspired by Gibbard (1990), which departs from the traditional understanding.

[123] Of course, there are plenty of philosophers who don't think normative judgments are cognitive attitudes, either, but I take it as a point of agreement between cognitivists and sensible non-cognitivists that normative judgments differ from mere dispositions to choose.

exist to an account of what it is for an agent to believe in a cardinal ranking. Here I'll argue that for an agent to have the belief that the difference between A and B is X times the difference between C and D is just for the agent to have the belief that P.

Something more will have to be said about this second step. It is, after all, generally illicit to move from:

1) S believes that P, and

2) For it to be the case that Q is for it to be the case that P,

to

3) S believes that Q.

Consider: Lois Lane believes that Superman is in love with her. For Superman to be in love with Lois Lane is for Clark Kent to be in love with Lois Lane. But from this it does not follow that Lois Lane believes that Clark Kent is in love with her. She doesn't know that Clark Kent is Superman.

However, this sort of move is licit when facts of the form of premise 2 are established through theoretical definition. And what I've been providing so far, and will continue to provide, are theoretical definitions of what it is for it to be the case that value differences compare rationally.

This makes theoretical definitions different from identities like "Clark Kent loves Lois Lane" and "Superman loves Lois Lane". It also makes theoretical definitions different from what are perhaps their closest cousins – translations. To borrow a case from Saul Kripke (1979), Pierre may believe that Londres est jolie without thereby believing that London is pretty. Perhaps he got his sense of "Londres" by visiting the city on a beautiful autumn day, but got his sense of "London" from reading 19th Century Social Realist fiction. So the first term's cognitive significance for him is different from that of the second term.

But theoretical definition rules out this difference in cognitive significance. When I stipulate a meaning for P, I say, in the manner of television hucksters, "Forget everything you know about P!" That is, treat P as having all and only the cognitive significance that Q has.[124] This provides a license for my general way of proceeding: First, give a stipulative definition of what it is for there to be such-and-such a cardinal ranking of actions. Then, use this definition to impute beliefs about cardinal rankings of actions to agents.

With that said, let's commence with the first step – giving an account of what it is for actions to fall along a cardinal evaluative ranking. I shall use a modification of what I earlier termed the <u>Asymmetric Three-Action Version</u>, with the understanding that modifications like this will apply <u>mutatis mutandis</u> to the other theorems stated so far:

<u>Asymmetric Three-Action Version for Objective Value</u>: The difference in Objective Value (OV) between outcome actions A and B is Y:X times the difference

[124] My treatment of stipulations is accompanied by a corollary about how stipulations can fail. Suppose I say, "I stipulate that by 'cool', which refers to that property shared by the likes of Miles Davis, James Dean, Elvis Costello, and Barack Obama, I mean 'hip'". Well, of course, "hip" has its own sense, whatever that may be. But my stipulation is saying, in one breath, that "cool" has precisely this sense, and that it has the sense of being a salient property shared by the foregoing eminences. And yet these sense might come apart; I may, for example, think that Barack Obama, whatever his virtues, just isn't hip. Good stipulations aren't "two-faced" in this way; they hone in on the semantic properties of one expression, and say that another expression has precisely those semantic properties.

> in OV between outcome actions B and C =df. The prospect actions 1) an action with an X subjective probability of being an instance of A, and a Y subjective probability of being an instance of C, and 2) an action with a 1.0 subjective probability of being an instance of B, are of equal Non-Normative Belief-Relative Value (N-NBRV).

This assumes that the N-NBRV of an action is linearly proportional to its EOV – an assumption similar to the one we’ve been making so far, but also an assumption that we shall call into question in the next section of this chapter. N-NBRV is the appropriate sort of value to assign to prospect actions as such, since it takes as inputs the agent's credences regarding, among other things, which outcome actions her actions are instances of, and these, of course, determine which prospect action an action is.

Now the move from cardinal rankings of actions to beliefs about cardinal rankings of actions:

> Asymmetric Three-Action Version for Normative Judgments: The subject believes that the difference in Objective Value (OV) between outcome actions A and B is Y:X times the difference in OV between outcome actions B and C =df. The subject believes that the prospect actions 1) an action with an X subjective probability of being an instance of A, and a Y subjective probability of being an instance of C, and 2) an action with a 1.0 subjective probability of being an instance of B, are of equal Non-Normative Belief-Relative Value (N-NBRV).

What if People Don't Have Beliefs Like These?

It may be objected, quite reasonably, that people tend not to have beliefs like the one just described. If they don't, then my attempt at cardinalization cannot succeed. So it'll be necessary for me to quell this worry.

A quick reminder before the serious argument begins: Don't be fooled by the baroque language – “prospect action”, “subjective probability”, “being an instance of A”, and especially “Non-Normative Belief-Relative Value”. While ordinary people tend not to use terms like these, that doesn't mean they lack the concepts these words express. It's highly plausible that our concepts are individuated much more finely than our public language terms are, and consequently, it often happens that different concepts get “mushed together” under the same term. Subjective probability and objective probability both just get called “probability” by the man on the street; mutatis mutandis for N-NBRV and OV. But that philosophers find it necessary to invent neologisms, and that the man on the street, with enough explanation, can understand these neologisms, attest to the fact that the distinct concepts of, say, subjective and objective probability were there all along, just waiting for names.

So what, then, are the different ways in which someone might lack the beliefs about prospect actions upon which I'm relying?

Possibility #1: Someone might lack a belief about the equality of two prospect actions because she had just never given the matter any thought, and hence hadn't formed any doxastic attitude regarding those two actions.

While I'm sure that this possibility sometimes obtains, two things may diminish the threat it represents. First, not all beliefs are occurrent. Most, I should think, are dispositional – not present before the mind, but rather playing behind-the-scenes roles in our theoretical and practical reasoning. So we should not assume that, simply because a thinker has never consciously entertained a proposition, that she lacks any beliefs regarding that proposition.

Second, even if someone lacks a belief about how two prospect actions compare, he may nonetheless consider the matter and form a belief when the necessity arises. He may "cardinalize on the fly", we might say, in situations where acting in accordance with EOV maximization requires at least partial cardinalization of the rankings over which is credence is divided. Suppose, for example, that I'm uncertain between a "consequentialist" ranking that says it's better to kill 10 than let 100 die, and a "moderate deontological" ranking that says it's better to let 100 die than to kill 10. Suppose further that, according to the consequentialist ranking, letting 100 die is equivalent to killing 100, but that, according to the deontological ranking, letting 100 die is equivalent to killing 1. Finally, suppose that both of the rankings agree that killing 100 is worse than killing 10, which is in turn worse than killing 1, since the relative disvalues of killing different numbers of people are not at issue between consequentialism and moderate deontology.

Finding myself with so-structured beliefs, I might be lucky enough to already have an additional belief of the form Killing 10 has the same N-NBRV as a prospect action with an X subjective probability of being an instance of killing 100, and a Y subjective probability of being an instance of killing 1, which will allow me to cardinalize the relevant parts of both of the rankings (by the Shared Ranking Method discussed in the last chapter, or the method of "background rankings" discussed in Sepielli (2009)). But if not, there's nothing stopping me from considering such a "10 vs.

100 or 1” scenario precisely for the purpose of forming beliefs about cardinal structure. Again, though, it's not obvious that in considering this matter and affirming to myself some proposition, I'm forming a new belief rather than bringing to occurrence one that already exists, and that guides my actions without my conscious attention to it.

Possibility #2: Someone might not have a full belief about how two prospect actions compare; rather, he may be uncertain about it.[125]

My way of dealing with this possibility is somewhat revisionary, but easy to explain. Consider two prospect actions: 1) An action that is certain to be an instance of B, and 2) An action that has an X subjective probability of being an instance of A, and a Y subjective probability of being an instance of C. I am uncertain which values of X and Y will make it the case that the two actions have equal N-NBRV's. My credence is .2 that the values are .3 and .7, respectively, .2 that they are .4 and .6, respectively, and .6 that they are .35 and .65, respectively. Each of these value pairs will yield a cardinalization, which will, in combination with a normalization and the agent's credences in the different objective value rankings, determine the EOV's of some actions. So an action performed under normative uncertainty – call it will have one EOV conditioned on the .3/.7 assignment, one conditioned on the .4/.6 assignment, and one conditioned on the .35/.65 assignment.

We can take the probability-weighted sum of these three different EOV's, which will yield the action's expected expected objective value, or EEOV. In a sense, though, the action's EEOV is just its true EOV, once we've considered all of the layers of probability involved. A simple illustration of this last point: Suppose I have a ticket that

[125] As I mentioned in Chapter 1, I'm ignoring imprecise credences throughout the dissertation.

will yield a prize with a value of 100 if a coin lands heads, and nothing if it lands tails. If my credence is .5 in heads, then the expected value of the ticket is 50. But now suppose that this coin will only be flipped if another coin lands heads, and that I have a .5 credence that it will do so. Then we might say that the expected expected value of the ticket is 25, but we might equally well say that the expected value, period, of the ticket is 25.

Generally, then, the strategy is to treat a credence regarding the equality of prospect actions as a credence in the cardinalization we get if those two prospect actions are equal. On each such cardinalization, the various actions performed under normative uncertainty will have EOV's. The true EOV's of the actions are just the sums of their EOV's on the various cardinalizations, weighted by the credences in those cardinalizations. Stated formally, the EOV of A = $\sum_i \sum_j$ p(Cardinalization$_i$) · p(Ranking$_j$|Cardinalization$_i$) · v(A given Ranking$_j$).

Possibility #3: Someone might not believe that the prospect actions are equal; instead, she may believe that they are on a par or incomparable.

This is the most philosophically interesting possibility, but it doesn't present much of a threat to our overall way of doing things. For someone with the beliefs in question, my approach will yield the result that his cardinal rankings will be indeterminate, and that this indeterminacy will be either what I called at the end of Chapter 2 "parity-induced indeterminacy" or what I there called "incomparability-induced indeterminacy". As we saw in that chapter, a limited degree of indeterminacy is consistent with there being a highest-EOV action. But indeterminacy can sometimes result in the non-suboptimal actions being what I'd called "rationally on a par" or "rationally incomparable". Recall that rational parity and rational incomparability exert different pressures on my behavior than rational equality does. For example, it is more common for the EOV of deliberating more about normative matters to be positive when there is rational equality than when there is rational parity or incomparability.

In summary: An agent's total lack of beliefs regarding prospect actions is rarer than one may think, once non-occurrent beliefs are taken into account, and may in any event be overcome through “cardinalization on the fly”; uncertainty regarding prospect actions can be dealt with by making only a small formal modification to our way of calculating EOV; and an agent with beliefs to the effect that prospect actions are on a par or incomparable will have indeterminate cardinal rankings of actions, but this is a possibility that may sit quite comfortably with the main arguments of the dissertation, as we saw in Chapter 2.

The Possibility of Misrepresentation

Now I want to extend my account in response to four challenges. Each of these challenges purports to take issue with the assumption that all agents are EOV maximizers under non-normative uncertainty. The first challenge claims that there may be agents who believe in maximizing not EOV, but rather some other quantity in which objective value is weighted. The second challenge alleges that there are agents who believe in maximizing not EOV, but rather some other quantity in which probability is weighted. The third challenge says that there are agents who are risk-sensitive in a particular sense that I discussed in Chapter 2, but that I'll remind you of in a moment. The fourth challenge claims that there are agents who regard intentions and motives, in addition to outcome actions, as relevant to the values of prospect actions. If any of these challenges succeeds, then my method of cardinalizing is mistaken. For I have been assigning these beliefs on the assumption that all agents are EOV maximizers under non-normative uncertainty. So my method may potentially misrepresent the cardinal rankings of all agents who are not EOV maximizers.

To see how these challenges complement one another, it will help to have a look at how the EOV of an action, A, is calculated. We take the probability of some possible state of the world, multiply that by the value of A given that state of the world, do the same for all the possible states of the world, and add them up:

$$\Sigma_i\, p(S_i) \cdot v(\text{A given } S_i)$$

According to the first challenge, there are some agents who believe that the action with the highest N-NBRV is the one that maximizes the following quantity:

$$\Sigma_i\, p(S_i) \cdot F(v(\text{A given } S_i))\text{; where } F(*) \text{ is non-linear}$$

This differs from EOV in that it replaces the value of an action given some state of the world – $v(\text{A given } S_i)$ – with a non-linear function of that value – $F(v(\text{A given } S_i))$.[126]

According to the second challenge, there are some agents who believe that the

[126] The formalizations in this section are similar to those discussed in Chapter 2, the difference being that I “separate out” normative and non-normative uncertainty in that chapter; here I do not, as I am discussing only non-normative uncertainty.

action with the highest N-NBRV is the one that maximizes this quantity:

Σ_i F(p(S_i)) · v(A given S_i); where F(*) does not equal *

This differs from EOV in that it replaces the probability of some state of the world – p(S_i) – with a differently-valued function of that probability – F(p(S_i)).

You should be able to see how these two challenges are complementary. To put it loosely, the first postulates agents with a different "value term" than the EOV maximizer; the second postulates agents with a different "probability term" than the EOV maximizer.

According to the third challenge, there are some agents who believe that the action with the highest N-NBRV is the one that maximizes this quantity:

v(A given S_2) + F(p(S_1)) · (v(A given S_1) – v(A given S_2)); where F(*) does not equal (*) and v(A given S_1) is greater than v(A given S_2)

To repeat what I said in Chapter 2: This is like the EOV function and the two functions just discussed in that it depends on probabilities of states of affairs, and the values of actions given those states of affairs. It differs in that it multiplies the probabilities of states not by the values of actions given those states, but rather by the differences between values of actions given different states: by v(A given S_1) – v(A given S_2), not by v(A given S_1), say.

According to the fourth challenge, there are some agents who believe that the action with the highest N-NBRV is the one that maximizes this quantity:

$F(\Sigma_i\ p(S_i) \cdot v(A \text{ given } S_i))$; where $F(*)$ does not equal $*$

This challenge postulates agents who believe that the N-NBRV of a prospect action depends, perhaps, on the probabilities and values of various outcome actions, but also on something else. I'm imagining that this "something else" is the intention or motive with which a prospect action is performed, but really, it can be anything that affects the value of a prospect action in a way that can't be explained by its effect on the values of the associated outcome actions. In a way, this challenge "outflanks" the other three; rather than simply claiming that agents can tote up the values of outcome actions in a way other than EOV maximization, it claims that in evaluating the N-NBRV of prospect actions, agents can do more than simply tote up the values of outcome actions.

Let's examine each challenge in greater detail.

The First Challenge

This challenge asks us to imagine a character who ranks prospect actions not in terms of their EOV, but in terms of their expected weighted value. For this character, there's a real distinction between the OV of an outcome action, and the contribution that value makes to the N-NBRV of a prospect action. The expected weighted value of a prospect action, again, is calculated as follows:

$\Sigma_i\ p(S_i) \cdot F(v(\text{A given } S_i))$; where F(*) is non-linear

If someone characterized by this function assigns the same N-NBRV to a prospect action with a .2 probability of being an instance of A and a .8 probability of being an instance of C, and a prospect action with a 1.0 probability of being an instance of B, then the difference between F(v(A)) and F(v(B)) will be four times the difference between F(v(B)) and F(v(C)).

But what about the value differences themselves? Well, since the F(*) function is non-linear, the difference in OV between A and B will not be 4 times the difference in OV between B and C. So treating the weighting agent as though he were maximizing EOV will get his value differences wrong. It will mistakenly treat his differences in weighted value between actions as though they were his differences in (just plain) value between actions.

It will help to consider, for a moment, a specific weighting function. Suppose the function that characterizes me is: $F(x) = x^{1/2}$. For me, then the marginal contribution of the OV of an outcome action to the N-NBRV of a prospect action diminishes as the OV of the outcome action increases. Consequently, I regard an increase in OV from 100 to 200 as, loosely put, more significant than an increase in value from 200 to 300. Contrast me with someone who maximizes EOV, and must therefore consider all increases in value by 100 to be equally significant.

The standard rebuke at this point is: “No, I’m defining differences in value as just those things assigned by the Ramseyan method. This person who you’re calling a value weighter is really an EOV maximizer, and what you're calling functions of her values – F(v(A)) and so forth – are just her values. On my picture, there's just no conceptual room for a difference between the value of an outcome action and the contribution of that outcome action to the value of a prospect action."[127]

But there must be more to the matter than that, for there are other ways of giving content to the notion of a difference in value. The probability-based method is but one option. With differences in value determined in another way, it may be possible for some agents, when acting under uncertainty, to be expected weighted value maximizers rather than EOV maximizers. That is, the availability of another method leaves room for a gap between what the agent believes to be OV of an outcome action, and the N-NBRV that outcome action contributes to a prospect action. Interestingly enough, Ramsey himself considered a non-probabilistic way of assigning value differences in a short paper written after "Truth and Probability". He is talking about differences in value between worlds, of course, not actions. Here's Ramsey: "A meaning [of "equal differences in value"] may, however, be given by our probability method, or by means of time: i.e. x-y = y-z if x for 1 day and z for 1 day = y for two days."[128]

The idea is that, rather than determining differences in value by comparing options, or goods distributed over epistemically possible worlds, we might determine differences in value by comparing goods distributed over intervals of time. This suggests the following modification to Ramsey's method:

[127] See, for example, Broome (1995).

[128] Ramsey (1929), pp. 256-7.

> Ramsey's Method (Time Version): The difference in value between A and B is equal to the difference in value between C and D =df. The agent has no preference between 1) A for an interval of time of length L and D for an interval of L, and 2) B for an interval of L and C for an interval of L.

But why stop at worlds and times? Why not do the same for goods distributed over persons? This might give us something like:

> Ramsey's Method (Persons Version): The difference in value between A and B is equal to the difference in value between C and D df. The agent has no preference between 1) A for one person and D for another, and 2) B for one person and C for another.

These are just some of the many ways of defining differences in value. Some of these ways will strike us as more plausible than others. For instance, I would be surprised if the following method attracted as many adherents as those above:

> Andrew Sepielli's Wealth Method: The difference in value between A and B is equal to the difference in value between C and D =df. The difference between AS's wealth, if A obtains, and AS's wealth, if B obtains, is equal to the difference between AS's wealth, if C obtains, and AS's wealth, if D obtains.

These are all ways of determining differences in value between states of affairs.

However, it is not hard to conceive of alternative ways of determining differences in value between actions. Consider an analogue of Ramsey's Method (Time Version)**:**

> Ramsey's Method (Sequence of Actions Version)**:** The subject believes that the difference in OV between outcome actions A and B is equal to the difference in OV between outcome actions C and D =df. The subject believes the sequences of actions 1) A followed by D, and 2) B followed by C, are of equal OV.

Hopefully you're getting the picture. There are all sorts of ways of defining differences in value. If something other than the probability-based method is right, then the probability-based method may sometimes get agents' value differences wrong, and vice versa. So what do we do now? How do we decide which method(s) to use?

There are two very extreme approaches that should be rejected. At one pole lies the view that we should just pick one method and stipulate that by "the subject believes that the difference between A and B is X times the difference between B and C", we're talking about the differences determined by this method. But this has the spirit of raw stipulation, of the sort I was accusing Lockhart of, rather than of a genuine attempt to get the semantics right for relevant concepts. For while it's out of place to criticize a stipulation as being wrong – people are free to stipulate whatever meanings they'd like for their terms – it's certainly in order to criticize a stipulation for taking us too far away from concepts that actually figure in our thoughts.

At the other pole lies the "grab bag" approach: We take the beliefs about value differences imputed by lots of different methods and average the value differences (or something like that), and impute to agents beliefs in the averages. This approach can't be made to work. First of all, why average the differences? Why use any way of combining the differences according to different methods rather than another way? I see no way of giving a principled answer to this sort of question. Secondly, which methods do we throw in the stew? You may be tempted to say "all of them". But this would be to include not only the sensible ones, but also silly ones like Andrew Sepielli's Wealth Method above.

I want to adopt an approach that lies between these two poles – one that is ecumenical enough to do justice to our pre-theoretical concept of a difference in value, but still able to exclude certain approaches on theoretically legitimate grounds. This approach will not tell us exactly which methods to use; rather, it will give us a recipe of sorts for determining which methods to throw in.[129]

Let me introduce this ecumenical approach by focusing on the possible relations between "dimensions" along which value may be located.[130]

Consider the following pairs of alternatives:

1.a. Matthew eats at a fancy restaurant now and a McDonald's tomorrow, or

1.b. Matthew eats at a decent restaurant now and tomorrow

2.a. Matthew eats at a fancy restaurant and Jeannie eats at a McDonald's, or

2.b. Matthew eats at a decent restaurant and Jeannie eats at a decent restaurant

[129] Thanks to Holly Smith for helping me to work out the details of this proposal.

[130] I borrow the language of "dimensions" along which value may be "located" from Broome (1995) and (2004).

3.a. Jeannie has a .5 probability of eating at a fancy restaurant and a .5 probability of eating at a McDonald's, or

3.b. Jeannie has a 1.0 probability of eating at a decent restaurant.

In each pair of cases, bearers of value – in this case, dinners – are spread out over some dimension. In the first pair of cases, the bearers of value are spread out over different times (i.e. a dinner for Matthew now, and a dinner for Matthew later); in the second pair of cases, the bearers of value are spread out over different people (i.e. a dinner for Matthew, and a dinner for Jeannie); in the third pair of cases, the bearers of value are spread out over different epistemically possible worlds (i.e. some chance of this dinner for Jeannie, some chance of that dinner for Jeannie).

There are two types of possible evaluative facts that concern us about these cases. First, there are facts about how the options within each pair compare to each other – about whether 1.a. is better than 1.b., about whether 3.a. and 3.b are equal, and so on. Second, there may be facts about how comparisons within one pair are related to comparisons within other pairs. It may be that, 1.a. is better than 1.b. if and only if 3.a. is better than 3.b. Or it may be that 2.a. and 2.b. are equal if and only if 3.a. and 3.b. are equal. I will call facts of this second sort "Cross-Dimensional Relations".

So why might people believe in cross-dimensional relations? Two reasons: First, one might have a view about the metaphysics of times, persons, and possible worlds that underwrites a certain view about cross-dimensional relations. I might, for example, be of the view that persons are just collections of similar-enough "person-slices" stacked end-to-end temporally. If this view is right, then we will want to treat the person-slices

"Matthew Now" and "Matthew Tomorrow" just like the person slices "Matthew Now" and "Jeannie Now". For example, we will want to say that Matthew Now dining at a fancy restaurant and Matthew Tomorrow dining at a McDonald's is better than Matthew Now and Matthew Tomorrow dining at decent restaurants if and only if Matthew Now dining at a fancy restaurant and Jeannie Now dining at a McDonald's is better than Matthew Now and Jeannie Now both dining at a decent restaurant. In other words, we will want to say that 1.a. is better than 1.b. if and only if 2.a. is better than 2.b. I am not necessarily endorsing this particular view about cross-dimensional relations; it is merely the kind of view one could hold.[131]

People may also believe in cross-dimensional relations because of more formal arguments. Here is such an argument, modeled on so-called "Dutch Book" arguments popular in decision theory: Suppose there are two alternatives for Jeannie right now: 1) We flip a fair coin. If it lands heads, Jeannie eats at a fancy restaurant now; if it lands tails, Jeannie eats at a McDonald's now; 2) Jeannie eats at a decent restaurant now. Suppose there are two alternatives for Jeannie tomorrow. 3) If the coin we flipped yesterday lands heads, Jeannie eats at a McDonald's tomorrow; if it lands tails, Jeannie eats at a fancy restaurant tomorrow; 4) Jeannie eats at a decent restaurant tomorrow. Now, suppose we say that 3.a. from above is better than 3.b. Well, then option 1) here is better than option 2), and option 3) here is better than option 4). So it's better for Jeannie to flip the coin both times. But since the outcomes in 1) and 3) depend in opposite ways on the same coin, Jeannie will eat at the fancy restaurant now only if she eats at the McDonald's tomorrow, and at the McDonald's now only if she eats at the fancy restaurant tomorrow.

[131] As you may recognize, it is a bowdlerized version of the position defended in Part III of Parfit (1984).

In cases like these, where successions of probabilistic outcomes are mutually dependent, two occurrences of 3.a. rather than 3.b. will yield 1.a. over 1.b. So 3.a. had better stand in the same comparative relation to 3.b. that 1.a. stands to 1.b., otherwise something has gone very wrong. Again, I'm certainly not alleging that this Dutch Book-ish argument is sound. Indeed, I've argued against Dutch Book arguments briefly in Chapter 2. But it is the sort of reason that might prompt one to believe in cross-dimensional relations.

Now suppose that, in addition to beliefs about basic comparative facts – how 1.a. compares to 1.b. and so forth – the agent has beliefs about cross-dimensional relations. This introduces the possibility that her perspective on probabilistic options is somewhat Janus-faced. She may believe a) that 3.a. is better than 3.b., and b) that 2.b. is better than 2.a., and c) that, if 2.b. is better than 2.a., then 3.b. is better than 3.a. These cannot all be true, since b) and c) imply that a) is false. This puts a bit of a spanner in the works: Our strategy has been to derive cardinal rankings of items from ordinal rankings of probabilistic options. But here there are two salient ordinal rankings – the one the agent actually believes in, and the one implied by his other beliefs. How can these be brought together?

My answer is to peg the agent's cardinal ranking of these items not to her actual beliefs about probabilistic options, but rather to her rationally idealized beliefs about probabilistic options. In the example just given, the agent has the belief, to whatever degree, that 3.a. is better than 3.b. But given that she has two other beliefs that together imply that 3.b. is better than 3.a., it would be theoretically rational for her to revise this belief. Perhaps she should instead think that 3.a. is equal to or worse than 3.b. Or perhaps her credences should be divided among these different hypotheses. The doxastic state she'd have after these revisions is her rationally idealized state.

Now, the foregoing discussion proceeded in terms of the values of items, not the values of actions, but the points are easily cross-applied. We might, for instance, combine the Sequence of Actions Version and the Asymmetric Three-Action Version. Suppose I believe that the sequences of actions 1) A followed by C, and 2) B followed by B again, are of equal OV, and that if these sequences are of equal OV, then the prospect actions 1) an action with an .5 subjective probability of being an instance of A and a .5 subjective probability of being an instance of C, and 2) an action with a 1.0 subjective probability of being an instance of B are of equal N-NBRV. In other words, I've got a view about the values of sequences, and believe further that the values of prospect actions should be computed "just like" the values of sequences.

Suppose I have the further belief that 1) an action with a .7 probability of being an instance of A and a .3 probability of being an instance of C, and 2) an action with a 1.0 probability of being an instance of B are of equal N-NBRV. Obviously, the content of this belief cannot be true if the contents of both of the aforementioned beliefs are true, so there will be a gulf between my actual beliefs regarding prospect actions and my rationally idealized beliefs – the ones I would have after I revised my beliefs in accordance with the tenets of theoretical rationality. This leaves us with the following adjustment to the Asymmetric Three-Action Version:

Idealized Asymmetric Three-Action Version: The subject has a cardinal ranking

> according to which the difference in OV between outcome actions A and B is Y:X times the difference in OV between outcome actions B and C =df. The subject would believe the prospect actions 1) an action with an X probability of being an instance of A, and a Y probability of being an instance of C, and 2) an action with a 1.0 probability of being an instance of B, are of equal N-NBRV, were she ideally theoretically rational.

...with the understanding that the same revisions will apply, mutatis mutandis, to the other methods discussed so far.

A few comments about these revisions. While I've been discussing agents who have beliefs in cross-dimensional relations, it's quite possible that some have absolutely no beliefs of this sort. (My suspicion – and take this with a pile of salt – is that those who see the normative world through a consequentialist lens are much more likely to have beliefs about cross-dimensional relations than those who don't.) For that matter, there's no requirement that an agent have evaluative beliefs about dimensions other than epistemically possible worlds, either. Someone might have the view, for example, that sequences of actions, as opposed to actions themselves, just don't have values at all. The claim is simply that, if an agent has some of these other beliefs, these may put constraints on the rankings of prospect actions she may rationally have.

Second, while I've been focusing on beliefs about cross-dimensional relations specifically, there are other beliefs that may play a similar role. For example, I might believe that a) One should spend twice as many years in jail for killing 10 people than for killing 5 people, and b) If a), then an action with a .5 probability of being an instance of killing 10 people and a .5 probability of being an instance of killing no one has the same N-NBRV as an action with a 1.0 probability of being an instance of killing 5 people. If I have these two beliefs, this will put some rational pressure on me to increase my credence in the consequent of b).

Third, this general strategy provides principled grounds for excluding methods like the Andrew Sepielli's Wealth Method. I might have beliefs about cross-dimensional relations, or about how appropriate levels of punishment for actions can serve as a guides to action under uncertainty; I am under no illusion, however, that there is any systematic connection between my own personal wealth and the reasons I have to perform prospect actions. As a general matter, conditions will be excluded from having any effect on an agent's cardinal ranking insofar as that agent finds them irrelevant to rankings of actions performed under uncertainty, which is as it should be.

Fourthly, I should say something about the "global rationality vs. local rationality" issue. I want to just stipulate that when I talk about a ranking of prospect action's being "ideally theoretically rational", I mean to invoke a local conception of rationality, relative only to the agent's beliefs about prospect actions, "linking beliefs" like those about cross-dimensional relations, and her beliefs linked via those beliefs to her beliefs about prospect actions. Whether my linking beliefs or the linked beliefs are themselves irrational given some further set of beliefs is irrelevant, once the subset of beliefs has been so circumscribed.

The Second and Third Challenges

The next two challenges will receive very similar treatments, so let me consider them in the same section. The second challenge asks us to imagine agents who are "probability weighters"; the N-NBRV's of their prospect actions are calculated as follows:

Σ_i F(p(S$_i$)) · v(A given S$_i$), where F(*) does not equal *

Just as the expected weighted value formula contained functions of the values of outcome actions rather than just the values of outcome actions, the present formula contains functions of probabilities rather than the probabilities themselves. (That is to say, we've just shifted the F(*) from the value term to the probability term.) To get a feel for how this formula works, consider an agent whose probability function is: F(x) = $x^{1/2}$. This agent will value a prospect action with this profile:

	Probability	Value
Outcome Action 1	.5	0
Outcome Action 2	.5	100

...more highly than a prospect action with this profile:

	Probability	Value
Outcome Action 1	0	50
Outcome Action 2	1.0	50

...even though both prospect actions have the same expected value: 50. That's because the first action's probability-weighted expected value is ~70.71, while the second action's is only 50. The sort of agent represented here cares more about differences in probability towards the "lower" end – i.e. closer to zero – than about differences in probability towards the "higher" end – i.e. closer to 1. (By contrast, an agent whose probability function was $F(x) = x^2$ would have the inverse view.) The Ramseyan method, however, assigns differences in value on the assumption that agents are EOV maximizers, and so will err when it comes to agents who are probability weighters.

The third challenge asks us to imagine agents who are risk-sensitive. Their N-NBRV's are, once again, calculated as follows:

$v(\text{A given } S_2) + F(p(S_1)) \cdot (v(\text{A given } S_1) - v(\text{A given } S_2))$; where $F(*)$ does not equal $(*)$ and $v(\text{A given } S_1)$ is greater than $v(\text{A given } S_2)$

Lara Buchak, from whom I've borrowed this challenge, explains the intuitive idea behind the formalism as follows:

> "In effect, the interval by which the agent might improve her lot above what she is guaranteed to get shrinks not merely by her probability of getting the better prize, but by a function of this probability, which reflects her attitude towards various probabilities. Thus the value of a gamble will be the minimum value guaranteed plus the amount by which the agent could do better, weighted by this function of the probability of doing that much better."[132]

[132] Buchak (ms #3), p. 11.

If A could have two values – one if S_1 obtains, a lesser one if S_2 obtains – then the risk-sensitive agent's N-NBRV of A will be the sum of A's value of S_2 obtains, plus some function of the probability of A's being better than that – in other words, some function of the probability of A's value given S_1 minus A's value given the "baseline", S_2.[133]

When F(p) is greater than p for at least one value of p, and at least as great as p for all values of p, we may say that an agent is risk-seeking. Such an agent will multiply the difference between the higher and lower possible values of A by a greater number than will the risk-neutral agent, if F(p) is greater than p. This will magnify the significance for her of the possibility of A's having a value higher than its "baseline". Such an agent will care more about doing very well, and less (comparatively speaking) about what the baseline is.

When F(p) is less than p for at least one value of p, and not greater than p for any value of p, we may say that an agent is risk-averse. Such an agent will multiply the difference between the higher and lower possible values of A by a lower number than will the risk-neutral agent, if F(p) is greater than p. This will shrink the significance for her of the possibility of A's having a value higher than its "baseline". Such an agent will care less about doing very well, and more (comparatively speaking) about what the baseline is.

[133] Contrast this quantity with the quantity:
$v(\text{A given } S_2) + p(S_1) \cdot (v(\text{A given } S_1) - v(\text{A given } S_2))$,
...which is equivalent to:
$p(S_1) \cdot v(\text{A given } S_1) + (1 - p(S_1)) \cdot v(\text{A given } S_2)$,
which is simply the EOV of the prospect action A, when either S_1 or S_2 will obtain.

One response to challenges like 2 and 3 is to build in the agent's love of or aversion to risk, or her attitudes regarding probabilities, into the values of outcome actions through post-action attitudes like regret and jubilation. On this approach, we’re all EOV maximizers after all. Those of us who seem risk-averse, for example, are really taking into account the regret they'll feel if they opt for a risky prospect action and "lose", while those who seem risk-seeking are really taking into account the jubilation they'll fee if they opt for a risky prospect action and "win". Once we take into account the extra value or disvalue added by the agent's experiencing these attitudes, we can explain what may seem like probability weighting or risk sensitive valuation of prospect actions solely in terms of the probabilities and values of outcome actions. This approach is borrowed from John Broome's defense of expected utility theory in Weighing Goods.[134]

Buchak has recently criticized this sort of response. While it is certainly possible, she says, for an agent’s post-action attitudes to affect the value of the outcome in which it arises, this phenomenon cannot explain all cases of putative probability weighting and risk sensitivity. The agent may simply prefer the absence or presence of risk ex ante, even if she knows she’ll experience no regret or jubilation whatsoever once all’s said and done. Perhaps she knows she’ll never find out how a risky prospect action ended up. Here’s Buchak:

> “[Values like certainty] will not be dispersed among the states; they will not truly be the values of the outcomes by themselves. Risk is a property of a gamble before its result is known, and risk need not, so to speak, leave a trace in any of the outcomes.”[135]

[134] Broome (1995), p. 98.

[135] Buchak (ms #3), p. 28

And later, criticizing Broome: “There is no room on Broome’s picture...for risk to enter into an agent’s feelings about a gamble but not about any particular outcome.”[136]

This line of criticism hits its mark against those who would seek to explain away risk-sensitivity or probability-weighting in terms of after-the-fact attitudes like regret, jubilation, and so on. However, this is not the only way to build the value or disvalue of risk into outcome actions. It’s not even the most natural way.

Consider for a moment the relationship between prospect actions and outcome actions. It is not as though a prospect action beginning at some time T causes an outcome action that begins at some later time T+N. Rather, the outcome action starts at exactly the same time as the prospect action, because the outcome action is simply one of the actions the prospect action has a probability of being. It's just not certain which outcome action the prospect action will be. Given that this is the relationship between prospect and outcome actions, it’s wrong to think of ex-ante features like risk as properties of the former but not of the latter. On the contrary, such features are properties of all outcome actions associated with a given prospect action, since they will be instantiated in the action “come what may”. As such, any value or disvalue associated with properties like risk and/or certainty should also be seen as inhering in all outcome actions. In summary: Buchak is right that an agent may have "feelings about a gamble" that are not about "any particular outcome". But that's not because they are not about outcomes at all. They are about all outcomes.[137]

[136] Ibid.

[137] This way of building risk sensitivity into outcome actions is similar to a strategy proposed by Paul Weirich (1986). Broome dismisses Weirich’s strategy on the ground that it trivializes the assumption of risk neutrality. But in the present context, I should think that trivialization is precisely our goal.

Since we’re now counting probabilistic properties of a prospect action as a component of its outcome actions, the latter way need to be individuated more finely than we’ve been doing so far. We shall need to countenance the possibility of a difference in OV between an the outcome action *destroying the Washington Monument when there was a .0001 probability that you would do so*, and the outcome action *destroying the Washington Monument when there was a .45 probability that you would do so*. There is, for this very reason, a theoretical casualty of the approach I’ve just articulated. It’s something called the *Rectangular Field Assumption*.[138] But because this assumption will also be threatened by my response to the fourth challenge, I’ll wait until the end of the next section to discuss it.

The Fourth Challenge

The fourth challenge asks us to imagine a character who believes that the N-NBRV of a prospect action is a function not only of the probabilities and values of outcome actions, but also of the intention or motive with which the action is performed.

Views on which intentions and motives affect the values of actions are, of course, very common in practical philosophy. Consider an example sometimes used to illustrate the “Doctrine of Double Effect”:

> *Pilot One* drops a bomb with the intention of destroying the enemy’s munitions plant. However, he knows that the blast will also kill some innocent civilians

[138] Thanks to Lara Buchak for raising this concern.

> whose homes are located near the plant. Pilot Two drops a bomb with the intention of killing innocent civilians. However, he knows that the blast will also destroy the munitions plant.

As I'm imagining the case, the effects of Pilot One's actions are the same as the effects of Pilot Two's actions – a destroyed munitions plant, and some number of civilian deaths. And in both cases, the pilot knows that these will be the effects. It's not as though, say, Pilot One is under the illusion that his bomb will strike only the munitions plant. Finally, the causal chains have the same structure. In both cases, it's the very same bomb blast that (more or less) simultaneously destroys the plant and kills the civilians. And yet, many people believe that Pilot Two's action is worse than Pilot One's, on account of the bad intention with which he acted.

A similar kind of argument may be advanced by those who think that an agent's motive in acting is relevant to the value of the action. If a judge sentences me to prison with the motive of seeing justice done, this is arguably a better than if he had sentenced me to prison out of personal hatred, even if all else is equal – deterrence effects, my degree of culpability, and so forth.

It does not matter, for my purposes, whether the foregoing views are correct. The fact is, there are people who believe in the normative relevance of intentions and motives, and my method of assigning beliefs in cardinal rankings would be faulty if it misrepresented such people. The Fourth Challenge alleges that it is faulty, in just this respect. For intentions and motives, it's argued, are features of prospect actions, but not features of any of the outcome actions, nor are they determined in any way by the distribution of credence over outcome actions. We cannot, then, simply assign OV to outcome actions based on ordinal rankings of prospect actions, since agents' valuations of intentions and motives may be affecting their ordinal rankings of prospect actions, too.

My response to this challenge is very similar to my response to the Second and Third Challenges. Outcome actions are actions that prospect actions have a probability of being, so all properties of prospect actions are properties of outcome actions, too. If I perform a prospect action with some intention and some motive, then that intention and that motive should be counted as features of the outcome action that eventuates, and their value should accrue to that outcome action. On this picture, we should not think of a prospect action as composed like this:

Prospect Action = Motive + Intention + P_1(Outcome Action 1) + P_2(Outcome Action 2)…

...but instead as composed like this:

Prospect Action = P_1(Motive + Intention + other features of Outcome Action 1) + P_2(Motive + Intention + other features of Outcome Action 2)

Some believe that intention and motive are different from the other features of outcome actions in that they are not, from the perspective of the deliberating agent, matters of probability. I may be unsure whether my action will harm someone or not, whether it will deceive someone or not, and so on. I will not, at the moment of my decision to φ, be unsure what my intention or motive is in φ-ing.

It is far from clear that we're certain of our own intentions. But even if we are, this poses no special difficulty for the present strategy. If doing some prospect action carries with it a 1.0 probability of acting with a certain intention or motive, then that intention or motive will simply be a part of each of the possible outcome actions. No matter which outcome action the outside world "selects", the intention and motive stick around and affect the value of that outcome action.

My discussion in this section generalizes in two ways. First, not only should we understand intentions and motives as components of outcome actions; we should construe outcome actions as including everything that an agent believes is relevant to the values of actions. So if, for example, the agent believes that the color of someone's ascot is relevant to the evaluation of her actions, then killing 25 people while wearing a red ascot will count as a different outcome action than killing 25 people while wearing a green ascot. Of course, this will mean that the very brief descriptions I've been using – "killing 25 people", and so forth – will be inadequate. For most agents, at least, action descriptions that refer to all of the relevant features will be much, much longer. For committed particularists, they may be infinite! In the interest of brevity, however, I'll stick with the "bare bones" descriptions I've employed up to this point.

Second, I will include a feature in the descriptions of an agent's actions not only when the agent fully believes that the feature is relevant, but also when the agent has any credence greater than zero that the feature is relevant. So, for example, if an agent's credence is .2 that intention matters and .8 that it doesn't, I will treat her as individuating actions in such a way that actions performed with different intentions, but that are otherwise qualitatively identical, will count as different actions. Such actions can in principle occupy different places on a value ranking (while, by contrast, two actions that differ only in terms of features that the agent has credence of zero are normatively relevant cannot occupy different places on a rational agent's value ranking).

The Rectangular Field Assumption

Now I want to discuss the aforementioned Rectangular Field Assumption. The name is a bid intimidating, but the underlying principle is very simple: Any two outcome actions can be outcome actions of the same prospect action. For even two wildly different outcome actions – for example, Squashing a fly and Vaporizing a cantaloupe – there is some conceptually possible prospect action that has a non-zero probability of being an instance of either.

My responses to the Second, Third, and Fourth Challenges threaten this assumption. The response to the Second and Third Challenges required the inclusion of probabilities as features of outcome actions, leaving us with outcome actions like Destroying the Washington Monument when there was a .45 probability that you would do so. This, obviously, cannot be an outcome of the same prospect action as the outcome action Chipping off a bit of the Washington Monument when there was a .50 probability that you would destroy it. For the probability of a prospect action's being an instance of destroying the Washington Monument cannot be both .45 and .50. The response to the Fourth Challenge required the inclusion of the intention or motive as a feature of outcome actions, leaving us with outcome actions like Destroying the Washington Monument with the intention of harming some nearby tourists. This cannot be an outcome of the same prospect action as the outcome action Chipping of a bit of the Washington Monument without the intention of harming anyone. For the prospect action's accompanying intention cannot be both to harm people and not to harm people. The more we build into outcome actions, the more difficult it becomes to "combine" them into prospect actions; if we build in enough, it becomes conceptually impossible to combine them.

The problem with violating the Rectangular Field Assumption is that it makes it more difficult to construct a cardinal ranking of outcome actions. Suppose that there are three outcome actions, A, B, and C, and that they are ranked in that order from best to worst. If A and C cannot be outcome actions of the same prospect action, then we can't determine how the difference in value that the agent believes there to be between A and B compares to the difference in value that the agent believes there to be between A and C – at least, not using the methods I've been developing so far.

I propose to compensate for this in two ways. The first is by using other actions that can be ranked vis a vis both A and C to put them on the same cardinal scale. Suppose I'm certain that another action, D, has the same value as C. Furthermore, D has an advantage in that it can be an outcome action of the same prospect action as A. Then C and A can be put on the same cardinal scale. Specifically, the difference between B and C will be the same size as the difference between B and D, which will stand in some ratio or other to the difference between A and B.

Here's a concrete example of that. Suppose I want to place complementing someone with the intention of conning him and complementing someone with the intention of helping his confidence on the same scale. Let's suppose, for argument's sake, that these cannot be outcome actions of the same prospect action. (And this is not an uncontroversial supposition; it’s quite possible that I might be uncertain whether my intention in complementing someone is to flatter him, or to help his confidence.) Still, an agent might believe that *complementing someone with the intention of conning him* has the same OV as *criticizing someone harshly with the intention of helping his confidence*. So if we can place the latter on the same scale as *complementing someone with the intention of helping his confidence*, then so can we place the former. And it does seem that we can do this with the latter. For it’s possible for me to do an action with the intention of helping someone’s confidence, and yet be unsure whether that action will be an instance of complementing or criticizing.

The second way is by separating outcome actions that cannot be part of the same prospect action into different prospect actions, and then comparing those prospect actions. Suppose once again that A and C cannot be outcome actions of the same prospect action. Still, we might compare the prospect action that has some probability of being an instance of A, and some probability of being an instance of *D*, with the prospect action that’s certain to be an instance of C, and locate A and C on the same cardinal ranking that way. So even if, for instance, *complementing someone with the intention of helping his confidence* and *complementing someone with the intention of conning him* can’t be parts of the same prospect action, they can still be parts of different prospect actions which can then be compared ordinally to one another.

Conclusion

This chapter concludes the defense of the thesis that it's most rational to maximize

EOV when acting under normative uncertainty. Viewed at a finer grain, it concludes my response to the Problem of Value Difference Comparison. It shows in detail how the actual cardinalization involved in the three cardinalization procedures I discussed in the last chapter is supposed to work. In a way, though, that characterization of the chapter gives it less credit than it's due. It's common for some philosophers to appeal to their normative theories' cardinal structures – to say, for example, that the difference in value between producing 100 utils and producing 50 utils is 5 times the difference in value between producing 90 utils and producing 80 utils. Other philosophers are skeptical that these intra-theoretic comparisons can be made sense of in a way that doesn't come off as ham-handed. If the arguments I've supplied here have been successful, then this chapter demonstrates just such a way.

CHAPTER SIX: STATUSES UNDER UNCERTAINTY

Introduction

So far I've focused on the question of what it's most rational to do under normative uncertainty. But you may be interested in an answer to a different, but related, question: What are the rational statuses of actions performed under normative uncertainty? That is, when is an action performed under normative uncertainty rationally required, or rationally forbidden, or rationally supererogatory? In this chapter, I'll sketch some possible answers to that question.

I say "answers" in the plural because there are several different ways of characterizing statuses, and of relating them to value rankings and other features. The answer to our question will depend on which approach the agent under evaluation adopts, and on which approach we as assessors adopt. And on certain well-known views, it will be nonsense to speak of rational statuses. I'll survey all of these views in the next section.

As a prelude to all of this, let me just flag two data points that a theory of rational statuses should help us to explain. The first is the anti-demandingness intuition we encountered in Chapter 3: Suppose you have some credence in a set of hypotheses such that, if those hypotheses are true, you are forbidden from doing A. (It's often thought that the abortion case falls under this schema; abortion may be forbidden, but there's no chance it's required.) Only an overly demanding theory of rationality would say that I'm rationally required not do do A whenever this is true. It's much more plausible to say that this requirement will be "dampened" by the significant possibility that B is permitted, leaving us with the result that it's rationally permissible to do either A or B.

The second, related intuition is that the rational status of an action should depend on the relative sizes of the value differences according to the normative hypotheses in which the agent has credence. For example, we ought to be able to say, in response to a question like, "Is it rationally forbidden to have an abortion?", that it depends whether abortion, if forbidden, is as bad as murder or only as bad as interrupting someone. Recall from Chapter 1 that this is the insight that the Compensationists were able to capture, but that Laxists, Rigorists, Probabilists, Equiprobabilists, and Probabiliorists were not.

The Different Characterizations of Statuses

Since what we say about rational statuses will depend on how statuses in general are characterized, it makes sense to begin with some attempts at such characterization. There are, as far as I can see, three different methods of relating statuses to value rankings. I'll illustrate these ways using the distinction between the obligatory and supererogatory as a recurring example.

The first is the Reasons-Plus approach: For an action to have a status is for it to occupy a position on some value ranking, and to have another feature that is not constituted by the action's position on a value ranking. Consider what is perhaps the most popular view of statuses: an action's status is determined by its position on a value ranking, and by the reactive attitudes towards the agent that her performance of the action would tend to make appropriate.[139] This view might say that an obligatory action is one

[139] Why only "tend to"? On our everyday concept of the obligatory, someone may fail to do an obligatory act without being thereby blameworthy. See, for example, cases involving psychotic aggressors. Conversely, someone may be blameworthy even when he

that has a high enough value *vis a vis* the other available actions, and is such that the nonperformance of it would tend to make the agent blameworthy. A supererogatory action also has a high enough value *vis a vis* the other available actions, but the nonperformance of it would not tend to make the agent blameworthy.

Note: “blame*worthy*”. On this particular view, the status of an action is a function of normative features only – the reasons to do the action, and the reasons to blame someone that its performance engenders. It illustrates the general feature of Reasons-Plus views that the “Plus” part needn't be non-normative, and in particular, needn't be something other than the strength of reasons for something or other. The only restriction is that it can't be the strength of reasons *to do the action*. So the “Reasons” part of “Reasons-Plus” refers to the reasons for or against the action; the “Plus” part might refer to any other feature, normative or non-normative.[140] This, in turn, is not to say that the “Reasons” part is explanatorily irrelevant to the “Plus” part. That would be silly. Taking the view above as an example, the strength of reasons to do an action is obviously relevant to whether one would be blameworthy for not doing it. It would be odd to say of the lowest-valued action in a situation that one could be blameworthy for not doing it.

Another possible Reasons-Plus view is one on which an action's status is a function of its position on a value ranking, plus the normative features of the act of deliberation that might precede the action.[141] On one elaboration of such a view, an

meets his obligations. A white supremacist jury member may be obligated to vote a black defendant guilty because of the evidence presented, but may nonetheless be blameworthy on the grounds that the defendant's race was among *his* motivating reasons for voting guilty.

[140] Thanks to Ruth Chang for comments that forced me to clarify this.

[141] On the more interesting reading of Raz's (1975) view, it works like this. What he calls “Second-order reasons” are not reasons to, say, go to the store rather than to the

obligatory act is one that has a high enough value *vis a vis* the alternatives, and is such that deliberation between the act and the alternatives has a very *low* value (i.e. it's great to do the action, and not so great to even consider not doing it). By contrast, a supererogatory act is one that one has a high enough value *vis a vis* the alternatives, and is such that deliberation between it an the alternatives has a high enough value as well (i.e. it's great to do the action, but not so terrible to consider not doing it).

Another major example is Susan Wolf's view, on which an action's status is determined by a combination of the strength of reasons to do it and the social expectation regarding its performance. On this view, an obligatory action is one that has a high enough value, and that it's socially expected that one will do. A supererogatory action is one that has a high enough value, but that it's not socially expected that one will do.[142] This Reasons-Plus view differs from the previous two in that its "Plus" feature is entirely non-normative. For this reason, it is also, to my ear at least, a much less plausible view.

The second general approach is the *Kinds of Reasons* approach: For an action to have a status is for it to occupy a position on at least two different kinds of value rankings. Both of these will be rankings in terms of kinds of value that go into determining an action's *overall*, "all-things-considered" value. Indeed, one of them may

pool. They're reasons to disregard certain of the reasons to go to the store rather than to the pool. Such a view can also be extracted from Bernard Williams' "One Thought Too Many" observation about a man deciding whether to save his wife or several strangers from drowning. See Williams (1981). It's error not only for the man to save the strangers, but to even consider the question of what it's right to do, rather than simply judging that his wife is drowning and saving her. Deliberation-regarding reasons are also invoked during the opening credits to the 1980's sitcom *Amen*, when the viewer is shown a sign above Deacon Frye's parking spot that reads "Don't Even *Think* About Parking Here!" (emphasis in original)

[142] See Wolf (2009).

just be the overall value ranking (and the other a subsidiary value ranking), as is the case on Doug Portmore's view of statuses. According to Portmore, an obligatory act is morally better than the available alternatives and all-things-considered better, but a supererogatory act, while morally better than the alternatives, is not all-things-considered better.[143]

Another example of the Kinds of Reasons approach is the one I favor, which makes use of a distinction between “voluntary” and “non-voluntary” value. The distinction, which is perspicuously employed in a recent paper of Chang's[144], is as follows: The amount of voluntary value that various considerations provide is determined by the agent's will; the amount of non-voluntary value that various considerations provide is not. Chang's view is that deliberation involves two stages: one in which we try to find out how much non-voluntary value actions have, and another one in which we exercise our wills in the form of “taking” things to be reasons (or to be stronger reasons), and thereby contribute extra, voluntary value to some actions. Building on this account, we might say that an obligatory act is one that has a greater non-voluntary value than any of the alternatives, and this difference in value is so large that none of the alternatives can be rendered better by an exercise of will. By contrast, a supererogatory act is one that has a greater non-voluntary value than any of the alternatives, but this difference is small enough that at least one of the alternatives can be rendered better by an exercise of will.

I like this view for two reasons: First, it gives voice to the idea that you have to do the obligatory, and not the supererogatory, or that the obligatory is (in a deontic sense)

[143] See Portmore (2003). Michael Zimmerman (1996) utilizes the distinction between “deontic value” and “non-deontic value” in service of a structurally similar account.

[144] See Chang (2009).

necessary while the supererogatory is not. An obligatory act is one that is better, and there's nothing the agent can do to change that. A supererogatory act is better, but the agent can change, or could have changed, that, through the act of taking considerations to be reasons. Second, it yields statuses that are actually relevant to practical deliberation. On accounts like – just to pick an easy target – Susan Wolf's, it's not clear why an agent should care, for the purposes of deciding what to do, whether actions are supererogatory or obligatory. Both are better than the alternatives, and that's all that seems relevant, in the final account, for deliberation. (I might have reasons to do what's socially expected of me as such, but a) I probably don't, and b) if I do, these will be reflected in the overall balance of reasons, just like the reasons provided by any old feature of the world that doesn't show up in an analysis of statuses.) But on the voluntary/non-voluntary view I favor, the obligatory/supererogatory distinction is deliberatively relevant. When I come to believe that an act is obligatory, I will see no point to the second, will-involving, phase of practical deliberation. When I come to believe that an act is supererogatory, though, I will see a point to this second stage, and will have cause to engage in it.

The third general approach to statuses is less complicated than the others, and I think, less plausible, but it's worth including it for the sake of completeness. This is the Strength of Reasons approach: For an action to have a status is for it to have some set of overall value-ranking features. On one such view, an obligatory action is one that's better than any of its alternatives to a degree less than N; a supererogatory action is one that's better than any of its alternatives to a degree greater than N. (This view expresses, perhaps too literally, the idea that the supererogatory is "above and beyond" the dutiful.)

On a more sophisticated view of this sort, an obligatory act is one with a higher value than any of the alternatives, and the ratio of the difference between it and the mean-valued alternative and the difference between it and the best alternative is less than some value M. A supererogatory act has a higher value than any of the alternatives, and the ratio of these differences is greater than M. In other words, an obligatory act is one such that the alternatives are clustered around some value that is a good deal lower than its value; a supererogatory act is one such that the alternatives are spread out fairly evenly below it.

A few comments about this division of approaches: First, depending on how we characterize some of the notions employed in the definitions above, a single method of defining statuses might fall under more than one of the approaches. Suppose that we define moral reasons as those reasons such that one tends to be blameworthy for acting against the balance thereof. Then the blame-based view and the moral reasons-based view may come out as the same view; this view will count as both a Reasons-Plus and a Kinds of Reasons approach. Notwithstanding possibilities like this, it's important to keep the approaches conceptually separate, since there are other ways of characterizing notions like that of a moral reason, for example, on which these two views of statuses are not the same.

Second, it's possible that a view might instantiate more than one of these methods, even in the absence of this sort of internotional characterization. It might do so by being a hybrid of one or more methods. For example, if we say that a supererogatory act is a) one that's better than the available alternatives by at least N, and b) one such that the act of deliberating about whether to do it has high enough value, then we've offered a view of supererogation that's both a Strength of Reasons and Reasons-Plus view.

Third, I should say that I take this array of approaches and their hybrids to exhaust the ways of relating statuses to value rankings. The Reasons-Plus approach makes actions' statuses functions of features other than those actions' value rankings (which again, may include the value rankings of things other than the actions); the Strength of Reasons approach makes statuses functions of values of at least two different qualitative kinds; the Kinds of Reasons approach makes them functions of purely quantitative features of an overall value ranking. Other than quantity and quality of reasons, and non-reason features, I can't see what other resources we could use to characterize statuses.

Fourth, it's important to see that the approaches above take no stand on whether statuses are conceptually prior to rankings, or rankings are conceptually prior to statuses, or whether neither is conceptually prior to the other. They are simply views about how statuses metaphysically depend on rankings (and perhaps other features). It may help to consider examples of conceptual priority and mere metaphysical dependence, respectively: The concept SOUND is prior to the concept AMPLIFIER. We think of an amplifier as something that makes sounds louder, rather than a sound as something that an amplifier makes more intense (and a muffler makes less intense, and deaf person can't detect, etc.) So we would give an understanding of AMPLIFIER (and MUFFLER, and DEAF) in terms of SOUND, not the other way around.

However, we can often explain how one thing metaphysically depends on another without coming to a view about which of the associated concepts is prior. For example, I can say that something's being red depends on its being colored, regardless of whether RED or COLOR is the prior concept; I can say that something's being a face depends on it's having enough of the following: eyes, a nose, a mouth, etc., without a view about whether FACE is prior to EYES, NOSE, MOUTH, etc., or the the other way around. Similarly, you are not barred from thinking that an act's being obligatory depends on its occupying such-and-such a position on a value ranking and its having such-and-such other properties, even if you think that status concepts are prior to ranking concepts. For my own part, I think that such a view about priority is nuts, but it's important to see that you can hold it and still accept everything in the foregoing, and in the rest of this chapter.

How Rational Statuses are Determined

Before we see how rational statuses are determined on each of these views, it's important to clear up an ambiguity. When we ask "What is the rational status of an action performed under normative uncertainty?", we're asking a question that's not maximally specific. One can be uncertain, after all, about all sorts of normative claims, and the local rational status of an action relative to uncertainty about one set of claims may differ from its rational status relative to uncertainty about another set. So what are the relevant sorts of uncertainty here?

Two answers suggest themselves. We might fill out our question as, "What is the local rational status of an action relative to an agent's credences regarding objective statuses?" Or we might fill it out as, "What is the local rational status of an action relative to an agent's credences regarding what, on some-or-other theory of statuses, are the determinants of objective statuses?" If we go with the first way, we will focus on the beliefs the agent expresses by saying, "Murder is forbidden," or "Playing two encores is supererogatory." If we go with the second way, we will focus on the beliefs the agent expresses by saying, “Murder's value vis a vis the alternatives is low, and one is blameworthy for murdering,” (if the blame-based view of statuses is assumed), or “Playing two encores is morally better than the alternatives, but not better overall,” (if Portmore's view is assumed).

Now, if the view we assume is the agent's own view of statuses, then pace an important condition, the answers we will give to the first “filled out” question will match the answers we will give to the second “filled out” question. That condition is that the agent is not theoretically irrational such as to believe, for example, a) that an action, A, has some status, b) that an action has that status only if a set of conditions, S, obtains, but also c) that S does not obtain. That is, an agent's beliefs about statuses must hook up properly with her beliefs about theories of the determinants of statuses, and her beliefs about what, on those theories, are those determinants.

If the agent is not theoretically rational, or if we, as assessors, assume a view of statuses that is not the agent's own, then our answers to the two questions will diverge. The rational statuses of actions relative to the agent's credences regarding objective statuses will differ from the rational statuses of actions relative to the agent's credences regarding what, on a view about the determinants of statuses (whether ours or the agent's), those determinants are. Either way, it will be important to settle on conclusions about what the rational statuses of actions are relative to beliefs about what, on some view or other, are the determinants of statuses. For some such view will be ours, as assessors; another will be the one that the agent believes; another, if the agent is irrational, is the one that is consistent with the agent's beliefs about the statuses of particular actions and her beliefs about the different possible determinants of those statuses. When assigning rational statuses to actions, it's worth doing justice to each. With that, let's see what happens on these different views.

The Reasons-Plus Approach

As it concerns rational status assignments under normative uncertainty, this approach to statuses will yield the quirkiest results. The reason is that, on this approach, statuses of actions depend on features other than those actions' values, and these features may not come in belief-relative forms. Consequently, it will be unclear exactly what a belief-relative or rational status consists in.

Consider the blame-based view of statuses. On this view, an action's status is determined by a) its value, and b) whether one tends to be blameworthy in virtue of doing it. Now, the "a)" feature, value, comes in all the different varieties we've been discussing – objective, evidence-relative, non-normative belief-relative, normative belief-relative, fully belief-relative/rational, and several "intermediate" grades. So for example, there's a kind of value that may not depend at all on the agent's beliefs, and a kind that depends only on the agent's beliefs.

But it's overwhelmingly plausible that blameworthiness doesn't work like this. It doesn't seem like there's one variety of blameworthiness such that whether an action renders one blameworthy on this variety depends not in the least on one's beliefs, and another variety of blameworthiness such that whether an action renders one blameworthy depends only on one's beliefs. Not only does blameworthiness not come in objective, belief-relative, and evidence-relative varieties; the view that best captures common sense is that the single variety it does come in is extremely gerrymandered relative to this classificatory scheme. Here's just a bare sketch of what I take to be common sense:

a) The distribution of available evidential support over certain non-normative hypotheses is an element of a set of conditions (the others involving conative and affective states) sufficient for S's performance of an action rendering S blameworthy. For example, my doing an action may render me blameworthy if the available evidence supports its being harmful, even if I don't believe it's harmful, and even if it's not in fact harmful.

b) The distribution of the agent's credence over certain non-normative hypotheses is also an element of a set of conditions sufficient for S's performance of an action rendering S blameworthy. For example, my doing an action may render me blameworthy if I am certain that it is harmful, even if the available evidence doesn't support it's being harmful, and even if it's not in fact harmful.

c) That certain normative hypotheses are actually true is an element of a set of conditions sufficient for S's performance of an action rendering S blameworthy, if those normative hypotheses are what we might call “core” normative hypotheses (e.g. it is wrong to inflict some amount of harm on an innocent person if the only good consequence of doing so is preventing less harm from being inflicted on another person). For example, someone's doing an action may render her blameworthy if, according to the core normative hypotheses, it is wrong, even if she did not believe it was wrong, and even if her available evidence did not suggest that it was wrong.

d) The distribution of *each of credence and available evidential support* over certain *normative*, but not "core" normative, hypotheses, is an element of a set of conditions sufficient for S's performance of an action rendering S blameworthy. A non-core normative hypothesis *par excellence* is that a form of prioritarianism with precisely such-and-such a weighting function is the right theory of distributive justice. I am not blameworthy for distributing goods in accordance with a slightly different theory, even if that theory is mistaken. However, someone may be blameworthy for distributing in accordance with a theory he believes is mistaken, even if the evidence doesn't suggest this, and even that's not the case. Similarly, someone may be blameworthy for distributing in accordance with a theory that the evidence suggests is mistaken, even if she doesn't believe this, and even if it's not the case.

Now, I doubt that I've captured everyone's blameworthiness intuitions perfectly. Among the reasons for this are, first, that I simply skated over the conative, affective, or intentional determinants of blameworthiness (e.g. I may be blameworthy for doing the best action if I'm motivated in doing it by one of the value-*reducing* features of the action), and second, I haven't considered the effects on blameworthiness of my credence distribution's (or my evidential distribution's) being the result of culpable action or inaction on my part.[145] But let's suppose that the foregoing picture is more or less

[145] On the conditions under which ignorance is culpable, see Smith (1983).

acceptable as is. What we've got, then, are all different kinds of value, each relative to a natural class of features, where a "class of features" is something like _the agent's beliefs regarding the determinants of objective value_ (in the case of belief-relative value), or _the evidence regarding the determinants of objective value_ (in the case of evidence-relative value). But we only have _one kind_ of blameworthiness, relative in screwy, convoluted ways to some features of each class, as we saw in the "a) through d)" list above. How do we construct the different kinds of statuses out of the different grades of value, plus the only kind of blameworthiness there is?

To make things much easier, let's just call the features of the gerrymandered class to which blameworthiness is relative "C". As we saw above, common sense says that C includes, at the very least, credence and evidence regarding non-normative and non-core normative hypotheses, and which core normative hypotheses are actually true. This gives us two options for constructing actions' statuses out their values and the blameworthiness one would incur for doing them. First, we might say that there's only one grade of status – "C-relative" status. An action is C-relative obligatory, for example, if a) its performance tends to render the agent blameworthy, and b) it has a high enough C-relative value _vis a vis_ the other available actions in a situation. In that case, there's likely no such thing as an action's rational status or its objective status or its evidence-relative status, since C cuts across all of these categories. It's also reasonable to say that there's no such thing as C-relative status, since C-relative _value_ is of dubious ontological vintage; what possible use could there be for introducing this notion? Anyhow, if we take this first route, then the problem to which this chapter is addressed vanishes.

The other option is to say that there are belief-relative, evidence-relative, and all the other grades of statuses, but that each sort depends on blameworthiness, which is relative to C, and the relevant sort of value – e.g., belief-relative value for belief-relative obligation, permission, and so forth. For example, an action will count as rationally obligatory if it has a high enough rational value relative to the other available actions, and one tends to be blameworthy (in the only sense there is) for not doing it.

An upshot of combining blameworthiness, which is C-relative, and any of the grades of value we've been discussing, is that there will now be a status corresponding to actions that have the highest value of a certain grade, but the performance of which nonetheless renders one blameworthy. This should not be surprising. As moral philosophers, we've all considered cases of someone's doing the action that there are, as it turns out, the strongest objective reasons to do, but that the agent is nonetheless blameworthy for doing. On the other end, we've all considered cases in which someone does what is belief-relative best or most rational, but is nonetheless blameworthy. Consequently, there will end up being perhaps more rational statuses on this approach than we may have thought, since we'll now have a category of actions with very high, even optimal, rational value, but whose performance nonetheless renders the agent blameworthy. The rational statuses, then, will be:

- Rationally Obligatory – high enough rational value, blameworthy for not doing
- Rationally Supererogatory – high enough rational value, not blameworthy for not doing (and now we'll have to add) and not blameworthy for doing
- <u>Rational But Blame-Meriting</u> – high enough rational value, blameworthy for

doing

...and so forth.

If this is how we construct rational statuses, then it's rather easy to see how to determine an action's rational status on the basis of other features: We simply figure out it's rational ranking in the accustomed manner, and this along with its blameworthy-renderingness or lack thereof (which, I shall assume, we have some way of figuring out) gives us the action's status. We need make no grand modification to our system, as rational statuses are simply conjunctions of rationality rankings and blameworthiness.

For what it's worth, I'd want to say something similar about Susan Wolf's view of statuses. That is, we simply determine an action's rational value vis a vis other available actions, and whether the action is socially expected or not. These two facts together give us the action's rational status, if we assume Wolf's picture of statuses.

It's worth mentioning that there's a very different approach to rational statuses that someone with a broadly blame-based view might adopt. For while my blameworthiness for doing something depends on features of the world other than my beliefs, the propriety of other reactive attitudes directed at me may depend on precisely those beliefs. The sort of attitude I have in mind is the one (or one of the ones) we might take towards someone who does what may well be the objectively right thing, but in doing so acts irrationally in light of his own beliefs. We need to be careful in specifying what exactly this attitude is. It's not the attitude that we'd take toward the weak-willed person. Someone who acts irrationally in light of his own beliefs needn't suffer from a defect of will. The attitude is more like the one we'd have toward, for example, the person who believes that abortion is murder, but who condemns the slightest use of violence to prevent abortions. Perhaps he is objectively right to condemn such violence, but he is still a fitting object of some kind of “con”-attitude. Let's call this kind of attitude “rebuke”.

We may want to say that an action's being rationally obligatory, say, is for it to have a high enough rational value vis a vis the alternatives, and for an agent to be rebuke-worthy for not doing it. This seems like an acceptable definition, but it leaves us with a mystery to solve before we can find out which actions are rationally obligatory: on which features, exactly, does the propriety of rebuke depend, and how does it depend on them? Specifically, does it depend on the agent's beliefs about the features upon which blameworthiness depends, or on the agent's beliefs upon which objective value depends, or on a combination of the two, or on something else?

The conditions under which a sui generis reactive attitude is fitting is most certainly not a question I plan to take up here. What I've given here is simply a recipe for determining rational statuses, on a certain Reasons-Plus conception of statuses, once this question about rebuke-worthiness is answered.

I'll want to say something different about rational statuses on the Raz/Williams/Amen-inspired view than I said about them on either the blame-based or Wolfian views. There is no such thing as belief-relative blame, strictly speaking, and certainly no such thing as belief-relative social expectation. But there is certainly belief-relative or rational value as concerns the separate actions of deliberating, disregarding, and so forth. So just as the rational value of an action depends on the agent's credences regarding the determinants of the objective value of the action, the rational status of an action might depend on that, plus the agent's credences regarding the determinants of the objective value of the pre-action deliberation. An rationally obligatory act, then, might be one that has a high enough rational value, and is such that deliberation prior to performing it has a low rational value.

Were there to be another view of statuses on which the status of an action depends on its value and the value of another of the same agent's behaviors, I'd want to offer a similar treatment. For example, if we had an account that substituted "regret-worthiness" for "blameworthiness", it might be amenable to the present treatment, since regret, unlike blame, is an attitude that only the agent who performed an act may have.

The Kinds of Reasons Approach

It's relatively straightforward, on this approach, how rational statuses are determined. Let's have a look at two specific views.

The Portmorian view exploits the distinction between moral and all-things-considered practical reasons. An action is obligatory if it has a higher moral value than any other available action and a higher overall value than any other available action. By contrast, an action is supererogatory if it has a higher moral value than any other available action, but not a higher overall value. This second definition may strike you as bizarre. We typically conceive of supererogatory actions as better than the alternatives, rather than equal to (or more typically, under this definition) worse that some alternatives. The shock of this may be softened if we're willing to countenance parity and incomparability, and say that the overall value of a supererogatory action is often on a par with or incomparable with that of other available actions, but its moral value is greater. Anyhow, since the point of this exercise is not to evaluate these definitions of statuses, but rather to see how rational statuses would be determined on each, I'll press on.

If we opt for the Portmorian approach, then just as an objectively supererogatory action is one that's objectively morally best but not objectively best overall, we can say that a rationally supererogatory action is one that's rationally morally best, but not rationally best overall.

The term “rationally morally best” could use some explanation. Just as the rational overall value of an action is a function of the agent's credences in hypotheses about objective overall value, the rational moral value is a function of the agent's credences in hypotheses about objective moral value specifically. Now, I've been arguing that the rational value of an action is an increasing function of the action's expected objective value (EOV). It makes sense to say, along the same lines, that the rational moral value of an action is just an increasing function of the action's expected objective moral value. But nothing in what we've said so far demands this. We might instead say that, as it concerns rational moral value, Credence Pluralitarianism is correct. For what it's worth, I think the consequences of saying so are far less severe than the consequences of saying that Credence Pluralitarianism (or some other theory other than EOV maximization) is correct as it concerns rational value overall.

To determine an action's rational status, assuming the Portmorian view of statuses, we simply determine its rational overall value in the standard way, determine its rational moral value in whatever way we settle on, and then plug these answers into the formula I've suggested. The only tricky part is separating the value that counts as moral value from the value that doesn't. But of course, this is a challenge for someone who defends this view of deontic statuses, not for the theorist of normative uncertainty who's trying to find out how the rational status of an action is determined if this view of statuses is assumed.

Now for the view I favor – the one that relies on the distinction between voluntary value and non-voluntary value. It is not difficult to see how, on this view, we'd determine what actions' rational statuses are. A function of agents' credences in hypotheses, and actions' objective voluntary values on those hypotheses, gives us those actions' rational voluntary value; a function of credences in hypotheses, and actions' objective non-voluntary values on those hypotheses, gives us those actions' rational non-voluntary value. We then plug these rational values into the scheme above to yield an action's rational status. For example, a rationally obligatory action is one that is so much more rational than the alternatives such that that I cannot, by an act of will such as "taking" something to be a reason, render any alternative more rational overall.

If there's a general method of determining rational statuses on Kinds of Reasons approaches, then, it's this: separate the objective values into types. Calculate the rational values of those types based on the agent's credences in hypotheses, and objective values of those types on those hypotheses. Determine rational statuses in the way that the view of statuses says that statuses are, in general, determined by the different kinds of value.

Strength of Reasons Approach

Our method of determining rational statuses, if we opt for this approach to statuses in general, is the most transparent of all: however statuses depend on value rankings in general, rational statuses depend on rationality rankings. So long as we have a way of moving from credences in objective rankings to rationality rankings – which I should hope we do! – we're able to apply this method straightforwardly.

This is true whether the dependence of statuses on rankings is simple or complex. On the simple theory we canvassed at the outset, an obligatory action is one that's better than its alternatives by a comparatively small margin, while a supererogatory action is one that's better than its alternatives by a comparatively large margin. This is a weird theory, of course. We don't typically think of the distinction between the obligatory and the supererogatory in terms of a straightforward quantitative difference like this. But again, let's table that worry.

On this simple view, it's natural to say that the relationship of rational statuses to rational value is just like the relationship of objective statuses to objective values. The rationally obligatory action is rationally better than the alternatives by a small margin, and the rationally supererogatory action is better by a large margin.

On the more complex Strength of Reasons view we looked at, an obligatory act is one with a higher value than any of the alternatives, and one such that the ratio of the difference between it and the mean-valued alternative and the difference between it and the best alternative is less than some value M; a supererogatory act has a higher value than any of the alternatives, and the ratio of the aforementioned differences is greater than M. Rationally obligatory or supererogatory acts, respectively, are just those for which the aforementioned differences are differences in rational value.

Back to the Intuitions

Let me bring this discussion to a close by returning to the two intuitions I flagged in the introduction.

There is ample room, given each of the views of statuses we canvassed, within which to accommodate the anti-demandingness intuition raised at the outset. An act may be rationally best even though it is not rebuke-worthy, even though it is not socially expected, even though it is not rationally morally best, even though the difference in rational value between it and the next-best act is small enough to be overcome by an act of will, and even though the ratio between the difference between it and the mean-rationally-valued alternative and the difference between it and the rationally best alternative is very large, and so on. Therefore, an act may be rationally best without being rationally required.

There is also ample room to accommodate the Compensationist intuition. On all of these views about statuses, an action's status and, and specifically, an action's rational status, depend partly on sizes of value differences on the different hypotheses over which an agent's credence is distributed. On all of these views, you might say, “stakes matter”. But of course, the objection to Compensationism was never that it's counterintuitive to suppose that stakes matter – that a chance of a material venial sin counts the same as a chance of a material mortal sin. It was that, however strong the intuition that such distinctions do matter, it doesn't make sense to compare degrees of sinfulness across different views of what's sinful. However, this is simply the Problem of Value Difference Comparisons, albeit couched in starker terms. If I've managed to solve this problem over the course of the previous two chapters, then Compensationism should come into view as the clearly correct view about Reflex Principles. And on any of the approaches to rational statuses we've surveyed, a contemporary version of it will likewise turn out to be true.

CHAPTER SEVEN: UNCERTAINTY ABOUT RATIONALITY

Introduction

Once we countenance normative uncertainty, and rules of rational action under normative uncertainty, it seems as though we must also concede the possibility of uncertainty regarding those rules, and of second-order rules of rational action under that uncertainty. But of course, one might be uncertain regarding those second-order rules, and there might, in turn, be third-order rules of rational action under this sort of uncertainty. We can imagine more and more iterations of the same.

It's clear that such a scenario raises difficulties for my project.[146] However, it's not obvious exactly what those difficulties are. I claim that there are two of them: First, there's what I shall call the Problem of Action Guidance. The problem is that it may be impossible for an agent who is uncertain not only about objective normative rules, but also about all “orders” of rationality, to guide her behavior by norms. Second, there's what I shall call the Problem of Mistakes About Rationality. The problem here doesn't concern action guidance directly. It starts from the thought that an agent who is uncertain regarding the rules of rationality is therefore less than certain in the correct theory of rationality, and so is, to some degree at least, mistaken about rationality. We'll see that it's not clear what to say about the rationality of actions performed by someone who is subject to rational norms, but is mistaken about what those norms are. In answering these challenges, we'll learn more about how rationality works, and about what it takes for

[146] Several philosophers have raised this point in conversation: David Chalmers, Richard Fumerton, Mark van Roojen, George Sher, Ernest Sosa, Wayne Sumner, and Sergio Tenenbaum.

one's behavior to be norm-guided, as opposed to simply norm-conforming.

Before we get to all of that, though, let me make a few preliminary remarks. Notice first that this particular sort of concatenation of uncertainties can only occur if there's such a thing as normative uncertainty. For only if there is normative uncertainty can someone be uncertain regarding what it's rational to do under uncertainty. The absence of normative uncertainty is, of course, consistent with my being subject to rational norms, but not consistent with my being unsure which rational norms I'm subject to.

Second, it's important to see, even if only to appreciate the broad applicability of the arguments in this chapter, that it's possible to exhibit uncertainty about the tenets of theoretical rationality, not just of practical rationality. I might be uncertain, for example, whether it's rational to update one's beliefs in accordance with the reflection principle or – and this is relevant to at least one strand of the "peer disagreement" literature – to what extent I should modify my views about the evidence for a proposition in light of my views about higher-order evidence (that is, evidence regarding the evidential status of the first-order evidence).[147] Much of what I say in this chapter will be relevant to uncertainty about theoretical rationality, too.

Finally, it's worth getting straight on what the problems raised by uncertainty about rationality are not. It can't be among the problems that we're unable to act when we're uncertain about all orders of rationality. My guess is that many people are uncertain "all the way up", as it were. And yet we don't see these otherwise healthy people

[147] Weatherson (ms) raises a problem of this sort. Elga (forthcoming) provides a response that, for reasons stemming from my arguments in the second half of this chapter, I think is mistaken. I suggest a more satisfactory response to Weatherson in Sepielli (ms #2).

completely paralyzed, or locked in a jittery limbo like characters in a stalled video game. Their minds are pervaded by uncertainty, and yet they act.

Nor can one of the problems be that an agent's lacking a full belief in a norm is inconsistent with her being subject to that norm, such that someone who is uncertain about all orders of rationality is subject to no rational norms whatsoever. There may well be something it's most locally rational for me to do given my uncertainty between utilitarianism and deontology; my uncertainty regarding the applicable norm of rationality does not disparage that. Furthermore, there may well be something it's most locally rational for me to do given my uncertainty between utilitarianism and deontology, *and* my uncertainty between different first-order rational norms. My uncertainty regarding rationality relative all of these does not disparage that. And as we reach the boundary between inner and outer space, there may well be something it's *globally* rational for me to do, given *all* of my beliefs. This is so even though we've exhausted the contents of my mind, leaving me with no beliefs whatsoever regarding this global norm of rationality – not certainty, not uncertainty, nothing. So my being certain in a norm of rationality is not a condition on that norm's applying to me. One would have to have one's head stuck firmly in one's shell to think otherwise.[148]

The Problem of Action Guidance

The first problem we'll consider is whether uncertainty regarding objective normative features, coupled with uncertainty regarding all orders of rationality, is inconsistent with one's conduct being norm-guided. The animating idea is as follows: If

[148] See the celebrated discussion of a related phenomenon in Carroll (1895).

you are uncertain between normative hypotheses A and B, you cannot guide your behavior by either. Rather, you will need to guide your behavior by a norm of rationality under such uncertainty. But if you are uncertain between accounts of rationality C and D, then you cannot guide your behavior by either of these. By the same reasoning, if you are uncertain between accounts of rationality of all orders, you will be unable to guide your behavior by any of these accounts, and therefore will be unable to guide your behavior by norms, period.

Two assumptions are operative here. The first is that one can only guide one's behavior by norms about which one is certain. The second is that it makes sense, when one is uncertain among objective norms, to guide one's conduct by belief-relative norms (of which rational norms are the purest instances). In what follows, I'll chip away at these assumptions. I'll start with the latter, showing why it is really very odd to think of ourselves as guiding our conduct under uncertainty by belief-relative norms. There is, however, a thought in the neighborhood that's promising, and that's that we are guided under uncertainty by what, in Chapter 1, I called "epistemic probability-relative" norms. That this view is rarely, if ever, articulated, is due to a deficiency in the semantics of probability statements for which I suggested a remedy in Chapter 1. Unfortunately, this view is not going to be totally satisfactory, either, and once we see this, we'll find ourselves face-to-face with the first assumption – that we cannot guide ourselves by norms about which we're uncertain – with no options other than showing that it is mistaken, or admitting that our behavior is, to an alarming degree, unguided by norms. I'll go with the first option, and suggest how we might guide our conduct by norms about which we're uncertain. But if that's the case, you may ask, what's the point of developing a theory of rational action under uncertainty about norms? Aren't we supposed to be using theories like EOV maximization to guide our conduct? I'll provide a few answers to these questions before concluding.

Let's have a look at that first assumption, then. Philosophers sometimes think we need belief-relative norms, in addition to objective ones, if we're to guide our actions by normative hypotheses at all. There are two reasons they might think this. First, they might think that, since we can only act on the basis of what we believe, we must guide our actions by norms that are relative to our beliefs. If I believe, for example, that my train is arriving at 9:30, then I must guide my behavior by a norm like *If you believe your train is arriving at 9:30, you should be at the station at 9:25*.

This, however, is just a case of vehicle/content confusion. It's true that we must act on the contents of our beliefs, and as such, guide ourselves by norms that make reference to those contents. But we do not need to guide ourselves by norms that make reference to the beliefs themselves – norms like the one italicized above. Rather, if I believe the train will arrive at 9:30, I may guide my conduct by the norm *If the train will arrive at 9:30, you should be at the station at 9:25* – no "If you believe..." anywhere.

Not only do we not *need* belief-relative norms to play this role; it would be odd if they did. To guide myself by the belief-relative norm above, I'd need to form a belief *about my belief* that the train will arrive at 9:30. This sounds implausibly inefficient; why would nature have designed our minds to include this otiose meta-belief-forming mechanism? Furthermore, it would stand in tension with the natural view of practical reasoning, on which I look "outward" at the world in deciding what to do – well, most of the time, at least – rather than "inward" at my own mental states. Finally, it would hold my decisions about arriving at the train station hostage to considerations that had nothing to do with the train's arriving at 9:30. Suppose I read some articles on psychological "connectionism" and came to believe that there were no such things as beliefs.[149] If I were really reasoning using the belief-relative norm above, I would resist concluding that I ought to arrive at the train station at 9:25. But that's just absurd.

That's all well and good, you might say, as it pertains to cases of agents acting on the basis of full belief or certainty, but there is another, more legitimate motivation for introducing belief-relative norms for the purposes of action guidance: We must guide our conduct by belief-relative norms when we are uncertain about what's the case.[150] When I'm not sure whether the train will arrive at 9:30, 9:00, or 1:42, a norm of the form If the train will arrive at 9:30... will do me very little good in guiding my behavior. Rather, I'll need to guide my actions by a norm that looks something like If your credence is X that the train will arrive at 9:30, Y that the train will arrive at 9:00, Z that the train will arrive at 1:42...then you should.... This is a belief-relative norm, but it is relative to credences rather than to full beliefs. The idea is that, when I'm uncertain, there's no one way I believe the world is, and so an objective norm that makes crucial reference to the world being this way or that will be unhelpful.[151]

Mutatis mutandis for cases of normative uncertainty. If I'm certain that

[149] See, e.g., Garon, Ramsey, and Stich (1990).

[150] Holly Smith seems to favor the view that we may guide our action under certainty by objective norms, but must guide our action under uncertainty by belief-relative norms. See Smith (ms), p. 4.

[151] See Smith (ms) for a clear statement of this thought; Kagan (1998), Feldman (2006), and Jackson (1991) arguably rely on it as well.

utilitarianism is right, then I may use this theory to guide my actions. But if I'm uncertain whether utilitarianism, desert-sensitive consequentialism, or some form of non-consequentialism is right, I can't guide my action by any of these three theories. Instead, I'll need to guide my action by a norm of rationality under normative uncertainty like EOV maximization.

While the appeal to belief-relative norms is much more plausible here, it yields a picture of practical reasoning that is no less odd. Again, the manner of reasoning suggested is inefficient, inward-looking, and beholden to considerations that seem irrelevant to the matter at hand. Moreover, it is especially implausible to think that we would have essentially two modes of practical reasoning – one outward-looking one for cases of certainty or full belief, and one inward-looking one for cases of uncertainty.

What defenders of this view may have in mind, though, is a subtly different thought. Recall my discussion of epistemic probability-relative value (EPRV) from Chapter 1. The theory introduced there was that epistemic probability statements (and their more colloquial paraphrases) stood to states of uncertainty in the same way that statements without mention of epistemic probabilities stood to full beliefs. I express my belief that snow is white by saying, “Snow is white”, and my credence of .5 that snow is white by saying, “There's an EP of .5 that snow is white.” EP-relative norms, then, are just norms that are relative to epistemic probabilities. Given that such an apparatus is available, we might be very tempted towards the view that, under conditions of uncertainty, we may guide our actions by EP-relative norms.

The parallelism alone should make this view attractive. I express my belief that the kettle is boiling by saying “The kettle is boiling,” and may guide my behavior, given that belief, by a norm such as If the kettle is boiling, I ought to throw in the lobster. Along the same lines, since I express my credence of .8 that the kettle is boiling by saying, “There's a .8 EP that the kettle is boiling,” it makes sense to say that I may guide my behavior given that credence by a norm such as If there's a .8 EP that the kettle is boiling, I ought to throw in the lobster. In reasoning like this, I'm not looking inward at my credences, but rather looking outward at the world as it's (fracturedly) represented by those credences.

That the EP-relative view is available makes the first assumption – that we must guide our conduct by norms about which we're certain – more plausible than it would otherwise have been. Consider: If, in order to guide our conduct under uncertainty by norms in which we were certain, we needed to do so by belief-relative norms, then, given the strangeness of doing that, we might be led to give up the assumption that we needed to guide our actions by norms about which we were certain. But if instead we're able to guide our conduct by EP-relative norms in which we're certain, then we have no reason yet to give up the assumption that we must guide our conduct by norms in which we're certain.

However, if this first assumption is true, then we must ultimately give up the view that we may, in all cases, guide our actions by EP-relative norms. For just as I may be uncertain regarding belief-relative norms of all “orders”, I may also be uncertain regarding EP-relative norms of all “orders”. The regress problem haunts not only the implausible belief-relative view of action guidance, but also the more plausible EP-relative one.

To solve the problem, then, we'll have to give up the first assumption, and accept a view on which we can guide our conduct by norms about which we're uncertain. These can be EP-relative norms, certainly, but they can also be – and I suspect that, in quotidian cases, typically will be – objective norms. On this picture, an agent can move right from credences divided among objective norms to an intention to act that is based on those credences, and then from there to action. She doesn't need an intermediate certainty or full belief about which action is belief-relative or EP-relative best in light of her credences in these objective norms.

There are two grounds for accepting such a view. The main ground is that it appears as though we're able to guide our conduct by norms even under uncertainty about all orders of rationality, and this view is the only one that vindicates this appearance.

The other ground is parity of reasoning. Suppose I'm certain that I objectively ought to do A. My action may be guided by the norm that is the content of this certainty. So there must be a relation R, constituted by at least a sort of non-deviant causal influence, that obtains between my doxastic state and my action, A, such that A counts as being guided by the content/information of my doxastic state. Okay, now suppose my credence is .6 that I objectively ought to do A and .4 that I objectively ought to do B. Why not suppose that this relation obtains between each of these doxastic states and, let's just say, A? Why can't my .6 credence bear R to A, and my .4 credence also bear R to A? Perhaps an analogy will help. There's some way I can behave aboard a canoe such that my behavior counts as guiding the canoe. Now suppose we are both aboard a canoe. There is some way I can behave, and also some way that you behave, such that we both count as guiding the canoe. This is true even if we're paddling in different directions – that is, even if the canoe would go a different way than it's going if one of us were absent.

We might say, analogically, that my paddling the canoe alone : reasoning under certainty :: both of us paddling the canoe : reasoning under uncertainty.

“That picture makes sense,” you might say, “But then what was the point of developing a theory of rationality under normative uncertainty, and trying to convince people that it's right? If we can move straight from credences regarding objective hypotheses to intentions, without taking belief-relative or EP-relative norms as intermediate premises, then norms of the latter sort are otiose.”

This is an excellent question, but it's not unanswerable. For one thing, your credences in belief-relative or EP-relative norms may play a causal but non-inferential role in determining which actions you perform. Therefore, it may be useful to form a high credence in the correct theory of rational action because doing so will make it more likely that you'll move from your credences in objective norms to the action that is in fact rational relative to those latter credences. For example, it may be that raising one's credence in EOV maximization from, say, .3 to .8, will make it more likely that one will transition from credences in objective rankings to intentions to do EOV-maximizing acts.

For another thing, it's perfectly possible for me, as a defender of EOV maximization, to give you advice that accords with my theory without your having to token a belief in the theory every time you act. Suppose you are uncertain whether A is better than B, or B better than A. And suppose that, given your credences and value differences, A's EOV is higher. I can tell you, “Do A,” and you can follow this advice, without my having to say anything about EOV maximization, and without your having to token a belief about it. My advising you in this way is due to my high credence in EOV maximization, though, and this credence is due in large part to the sorts of considerations adduced throughout this dissertation. The theory, then, is doing some work with regard to your behavior, even though only I have a belief about it.

A deeper point, though, is that it's useful to have in hand a theory of rational action because, even if you don't need to guide your actions by it, you often may guide your actions at least partly by it. Suppose that you are uncertain between utilitarianism and some form of non-consequentialism, and that the theories deliver different verdicts in the present situation. Now, as I've said, there's no problem with your moving right from those theories to an intention to act. In most quotidian cases, that's what does happen. That's why the problem of action under normative uncertainty doesn't ring familiar with most people. But often, when we make more momentous decisions, we more consciously consider the advisability of actions given the contents of our credences. One might say to oneself, “Well, there's some chance utilitarianism is right, and some chance this form of non-consequentialism is right, so what should I do?” and then answer by saying, “Hmmm...I should probably do the action with the highest expected value [or whatever heuristic for this that we come up with], although there's a small chance that's wrong, too.” At this point, you might just move from your credences in the theories of rationality to an intention. But the higher your credence in the correct theory of rationality, the more likely it is that you'll act on that theory. So it's worth developing and defending the correct theory of rationality under normative uncertainty for those more reflective moments when we do guide ourselves by such theories. The higher your credence in the correct theory of rationality, the more likely it is that you'll do what is in fact rational.

Finally, theories of rationality play a role in specifying which behavior counts as norm-guided in the first place. It's worth pursuing this at some length, as it will shed light on the nature of norm-guidedness in general.

In other work, I defend a conception of action guidance that is partly normative.[152] It's easiest to see this in cases of certainty. Suppose I'm certain that it's objectively better to do A than to do B. As I suggested earlier, there must be a non-deviant causal relationship between this belief and my action in order for the action to count as guided by the norm that is the belief's content. But that's not all that's required. It's also a condition on my action's being guided by the norm that I do what it's rational to do given the belief. That is, I must do A. If I do B, I have not guided my action by the norm, even if my doing B is caused by my having a belief in the norm.[153]

Nobody, I'm sure, would resist the view that when one is certain that A is objectively better than B, it's locally rational relative to this certainty to do A. Rationality under uncertainty, though, is more controversial. My own view, of course, is that it's rational to do the action with the highest EOV. So suppose my credence is .7 that A is objectively better than B, and .3 that B is objectively better than A. On my normative account of action guidance, my action can only count as guided by these normative hypotheses if I do what's rational given this credence distribution – that is, if I do whichever of A and B has the higher EOV in this situation.

So while we needn't guide our actions by norms of rationality, it's only when we act in accordance with those norms that our actions count as guided by norms at all. The

[152] See Sepielli (ms #1).

[153] It may be that doing what's locally rational relative to a belief in a norm is among the requirements for that belief's causal relationship to my action's being non-deviant. This doesn't tell against my account. Rather, it means that I've managed to shed light not only on the nature of norm-guidedness, but also on the nature of the relevant sort of non-deviance.

norm of rationality applicable in a situation determines at least one of the conditions on an action's being norm-guided when performed in that situation.

This may strike you as an extreme conclusion – that only the EOV-maximizer's conduct counts as norm-guided. Of course, you might doubt it simply because you doubt that EOV maximization is the correct theory of rationality under uncertainty. But in that case, you have no quarrel with the connection between rationality and action guidance I'm sketching now; you simply think that rationality consists in something other than maximizing EOV. We can also take the edge off of this conclusion by allowing, as I'll want to do, that action guidance, like the rational value that partly undergirds it, can come in degrees. So if I do the action with an EOV somewhere between the highest possible and the lowest possible, then my behavior will count as partly guided by the contents of those credences. Finally, it's consistent my view that there are important distinctions among actions that are less than fully norm-guided. A low-EOV action is not fully norm-guided, but neither is it tantamount in all crucial respects to a spasm.

You might also object to this view of action guidance on the basis of examples like this famous one of Frank Jackson's:

> A doctor must decide whether to give a patient Drug A, Drug B, or Drug C. A is certain to almost completely cure the patient's illness; there is a .5 chance that B will completely cure the patient's illness and that C will kill the patient; and there is a .5 chance that C will completely cure the patient's illness and that B will kill

the patient.[154]

Suppose the doctor is certain of a reasonable normative theory. Such a theory will say that there's no chance whatsoever that the doctor ought to choose Drug A, and a 1.0 chance that the doctor ought to choose either B or C. So it seems that there's no way for the doctor to be guided by the objective theory to choose A, even though choosing A seems like the belief-relative best thing to do. This result is inconsistent with my view of action guidance, on which agents may be guided by objective norms, and that their being so guided consists partly in their doing the belief-relative best actions.[155]

But I'll want to resist this way of interpreting examples like Jackson's. Instead, what I'll want to say about the doctor will depend on which of two possible ways we “fill out” her psychology. I'll consider these in turn:

Suppose first that the doctor really only has beliefs about what she objectively ought to do, and no beliefs about how nearly curing, curing, and killing, respectively rank cardinally. Then I think her credence distribution over the different possibilities is not enough to support any of the actions as being belief-relative better than any of the others. There's simply no fact of the matter about which is belief-relative best. This is the nature of simple “ought” judgments: in cases of uncertainty among them, they radically underdetermine what it's rational to do. Along the same lines, none of choosing A, choosing B, or choosing C counts as either guided or unguided by the contents of these credences. As Wittgenstein (1953) would have said, it will often be that many actions can be “made out to accord with” “ought” judgments, and so none counts as guided by them.

[154] See Jackson (1991).

[155] Thanks to Richard Chappell for raising this objection.

To summarize, then, if the doctor has only “ought”-beliefs, then neither do her actions occupy places on a rationality ranking, nor do any of them count as more or less guided than the others, so my link between rationality and action guidance is undisturbed.

Suppose on the other hand that the doctor also has beliefs about at least rough cardinal rankings of choosing A, choosing B, and choosing C. (This is, I suspect, the supposition we're making when we intuit that choosing A is belief-relative best.) Specifically, suppose she believes that nearly curing has quite a high value, curing a slightly higher value, and killing a much lower value. If she has those beliefs, then it will turn out that it's belief-relative best to prescribe A. But then she will also be able to guide her choosing by the contents of these beliefs. So this way of filling out the case presents no threat to my claim that acting rationally is a condition on one's action being norm-guided.

What we cannot do, however, is to say, on assumption that she has only “ought”-beliefs, that she cannot be guided to choose A rather than B or C, and in the same breath say, on the assumption that she has beliefs about cardinal rankings as well, that choosing A is more rational than choosing B or choosing C. We must say one thing or the other about the agent's psychology. Either she has only “ought”-beliefs, and so it will neither be the case that it's more rational to choose any of the drugs over the others nor will it be the case that any of the choices will count as norm-guided; or she has beliefs about cardinal value rankings, too, in which case it may be more rational to choose one of the drugs and that choice will count as norm-guided. So there is no obstacle either way to my identifying rational action under uncertainty about objective norms with action that is guided by those norms.

The Problem of Mistakes About Rationality

To get a feel for this problem, consider an extreme case that gives rise to it – not of someone who is uncertain regarding the correct rule of rationality, but of someone who is certain in a rule of rationality that is, unbeknownst to her, mistaken. This is highly unrealistic, but it puts the problem in the highest relief possible. Suppose that Tory is certain that it's objectively better to do A than to do B, and (but?) is also certain it's more rational to do whatever is worse in the objective sense. Call the first belief “B1” and the second “B2”. It is most locally rational, relative to B1, to do A. But if she does A, she will do something that is less locally rational relative to B2. Put more generally, there's something, A, that it's rational for Tory to do relative to a set of beliefs, but she has a further belief that something else is rational to do relative to this first set, and it would be rational relative to this further belief to do B.[156]

[156] There's room to deny this very last bit – that there's something it's rational to do relative to one's beliefs about rationality (as opposed to one's beliefs about objective reasons). If I believe that the evidence overwhelmingly supports believing that Oswald acted alone, then it would seem irrational for me to then form the belief that he didn't act alone. But can the same be said if I believe that, relative to my other beliefs, it would be rational for me to believe that Oswald acted alone? Is my mental maneuvering in response to the belief about rationality really reasoning – really the sort of thing that's subject to standards of rationality, rather than something more like, say, taking ecstasy to alter my mistaken paranoid beliefs, or for that matter, cutting my hair based on the judgment that I'd look better with short hair? Ruth Chang suggested the negative answer in comments on an earlier draft of this chapter, and Niko Kolodny provisionally endorses this answer in his (2005). My hunch is that the positive answer is correct, but I don't have a great argument for it at the present time.   Consider, though, that every action falls under multiple descriptions. Perhaps an action might be sensitive to one category of judgments under one description, but sensitive to another category of judgments under another description. (This talk of judgment-sensitivity is borrowed from Scanlon (1998).) Specifically, we can think of an action performed under normative uncertainty as, among other things, an attempt, or a try, to do what's right or what's best. It seems natural to think that doing some action – turning the trolley, say – is sensitive qua “try” to

Those who are uncertain about rationality may not always run into rational conflicts of this sort. Suppose I am uncertain between utilitarianism and deontology, and suppose it's rational to maximize EOV in cases of normative uncertainty. Let's say the numbers work out so that doing what utilitarianism recommends has the highest EOV, and so that's in fact the most rational thing to do. Suppose further that I am uncertain whether EOV maximization or Credence Pluralitarianism is the correct theory of rationality. It may be that the numbers work out such that what it's locally rational to do relative to my credences in these theories is what utilitarianism recommends. In that case, there's no rational conflict; problem averted. But the numbers may work out so that what it's locally rational to do relative to my credences in theories of rationality is what deontology recommends. That'll happen, perhaps, if my credence in Credence Pluralitarianism is very high. Then we have rationality pulling in two directions, just like in the more extreme case above. This sort of rational conflict has the potential to arise in any case in which my credence in a correct theory of rationality is less than 1.

So that's the scenario I have in mind – where less-than-certainty about rules of rationality gives rise to a particular sort of rational conflict. The Problem of Mistakes about Rationality is simply that it's not obvious, to me at least, what it's rational to do relative to all of the beliefs involved – the beliefs to which the rational rule is relative, and the beliefs about that very rational rule. In what follows, I'm going to search for an adequate account of rationality in these circumstances.

judgments about rationality, even if it is sensitive qua something else only to judgment about reasons. This is because judgments about practical rationality can be construed as judgments about what constitutes the best try. EOV maximization says the best try to do what's best is the action with the highest expected value; Credence Pluralitarianism says the best try is the action that's most likely to be best.

On some approaches, what it's rational to do relative to a set of beliefs, S1, and a set of beliefs, S2, the contents of which are propositions about what it's rational to do relative to S1, depends on either only S1 or only S2.

One such option is to disregard S2; what it's rational to do relative to this entire set will depend only on the beliefs that comprise S1 (plus the rule of local rationality that actually applies to these beliefs, whatever the agent's credence in it). Argument: It’s theoretically irrational to have mistaken beliefs about rationality. A credence of less than 1 in the correct norm of rationality relative to S1 is a mistaken belief. Theoretically irrational beliefs should be treated as though they didn't exist for the purposes of determining requirements of practical rationality. So in this case, the beliefs in S2 get “quarantined off”, and we get the result that, necessarily, it's rational relative to both sets of beliefs to do what it's rational to do relative to S1 alone.

This answer is troubling in several ways. First, it relies on the premise that it's theoretically irrational to have false beliefs about rationality. But this is not totally obvious. It's not necessarily irrational to have false beliefs about, say, physics, or psychology, or morality, so why should practical rationality be any different? Second, it's far from obvious that theoretically irrational beliefs should have no influence on what it's practically rational to do. We have no problem saying that it's locally practically rational to do A relative to your certainty that A is the best thing to do, even if it's irrational to have that certainty in the first place. That is to say, there are better and worse ways to respond to a theoretically irrational belief. It's hard to see why a theoretically irrational belief's influence on what it's practically most rational to do should disappear once it becomes only one of a set of beliefs that we're evaluating rationality relative to.

Another option is to disregard S1, and say that what it's rational to do relative to the beliefs in S1 and S2 together depends only on the latter. This position seems meritless. There's no case to be made that S1 is theoretically irrational, though, and it would be absurd to suggest that norms of local rationality relative to sets of beliefs are somehow eviscerated once an agent forms beliefs about those norms.

This pushes us towards what seems like the more sensible family of approaches, according to which what it's rational to do relative to a set of beliefs is a function of all of the beliefs in that set. Illustrating the specific approach I favor will require making a distinction between two different types of rational conflict: what I shall call hierarchical and non-hierarchical types of conflict. In hierarchical conflicts, it's impossible to do what's most locally rational relative one set of mental states, S1, and also what's most locally rational relative to another set of mental states, S2, in virtue of S2's being a set of mental states the contents of which are propositions about what's locally rational relative to S1. In non-hierarchical conflicts, it's impossible to fully satisfy both norms of local rationality, but not in virtue of the contents of one set of states being hypotheses about what's locally rational relative to the other set.

Tory's case is just about the clearest case of hierarchical conflict you will find. It's locally rational relative to B1 to do A; the content of B2 is a hypotheses about what it's locally most rational to do relative to B1. Since it's a false hypothesis, Tory's in a state of rational conflict.

Here is a case of non-hierarchical conflict: Suppose Henry is certain of the hypothesis that action A is objectively better than action B. However, he also believes that B has more of feature F than A, and furthermore, he is certain that, for any two actions X and Y, X is better than Y if X has more of feature F than Y. Relative to one hypotheses, it is locally rational to to A; relative to the other, it is locally rational to do B. Notice that neither normative hypothesis is about what's rational relative to a belief in the other hypothesis; it is simply that different things are rational relative to beliefs in the two hypotheses, respectively.

I'll want to treat these types of rational conflict differently. Of particular interest for the current project are hierarchical conflicts, of course, but it's worth considering non-hierarchical conflicts in order to note the contrasts.

In a case like Henry's, it makes sense to say this: We add up the rational value of A relative to each of the conflicting hypotheses, add up the rational value of B relative to each of the conflicting hypotheses, and then do the action with the highest total rational value. Now, of course, it's theoretically irrational to be in Henry's situation in the first place. Specifically, it's theoretically irrational for one's credences over inconsistent propositions to sum to more than 1. But in determining what it's practically rational to do relative to one's beliefs in these hypotheses taken together, it seems correct to treat degrees of belief just as we'd treat them if the agent were theoretically rational – even if there are, so to speak, "more degrees" than there should be. So, for example, if we go with my theory that the most practically rational action is the one with the highest EOV when the credences over inconsistent propositions add up to 1, then I see no reason not to think that the most practically rational action is also the one with the highest EOV when the credences over inconsistent propositions add up to more than 1.

It seems to me that a similar treatment of hierarchical conflict is ham-handed. For example, it would be odd to take the rational value of one of Tory's actions relative to her belief that A is objectively better than B, and then simply add that to the rational value of her action relative to her belief regarding what it's rational to do relative to her first order belief. If you find it difficult to arrive at that intuition, consider a somewhat analogous case, involving theoretical rationality. Suppose Tristram is certain that A, and that A → B, but believes it's most rational to *retain both of these beliefs, but avoid forming the belief that B*. Doing so would be irrational relative to the first two beliefs, but rational relative to the third, "meta" belief. The idea that we simply add up the rational values here seems strange.

I think we can offer a theory of what undergirds that intuition. In Henry's case, the rational value relative to the first hypothesis is a function of the agent's credence in that hypothesis, and the objective values that obtain on that hypothesis. *Mutatis mutandis* for the second hypothesis. We might call rational value that's a function of credence and objective value *first-order* rational value (since it's the kind of rational value that's not relative to other beliefs about rational value; it's the kind that's "just above" objective value).

Things are different in Tory's case. There, the rational value relative to B1 is, of course, first-order rational value. But the rational value relative to her belief that it's more rational to do B than to do A if one believes that A is better than B (i.e. B2) is not first-order rational value. Rather, it's a function of her credence in that hypothesis, and of the sort of value that hypothesis is about, which is not objective value, but rather first-order rational value. In other words, it is the sort of value that is a function of the agent's credences regarding functions of the agent's credences in hypotheses about objective value. In still other words, it's the sort of value that's a function of the agent's credences about “how to put together” objective values and credences in order to yield first-order rational values.

Since there was only one order of rational value at stake in Henry's situation, it made sense to simply add it up and say that it'd be most rational for Henry to do whatever had the most of it. But since there are two orders of rational value at stake in Tory's situation, this same move makes much less sense. As a variation on this theme, suppose there's really water in a glass, but Craig's credence is higher that it's gasoline. It'd be crazy to simply add up the objective value (which drinking has more of) and the non-normative belief-relative value (which not drinking has more of), arrive a total for each action, and then say that Craig should in some “overall” sense, do the action with the highest value total. To counsel Tory to do the action with the highest aggregate rational value is to commit the same sort of mistake. But this is obscured if we fail to keep in mind the distinction between local rational value that's a function of the agent's credences regarding objective value, and a kind of higher-order local rational value that's a function of the agent's credences regarding first-order local rational value. The different orders of rationality, like objective and belief-relative value, are incomparable.

Does this imperil the plausibility of an overall rationality-ranking of actions in cases of hierarchical conflict? I think not, for there is a way to “collapse” cases of hierarchical conflict so that they look more like cases of non-hierarchical conflict: There is, on one hand, the actual first-order rational value of various actions, relative to the agent's credences regarding objective value. But when the agent has a credence distribution over propositions about first-order rational value, there is also a kind of “mock first-order rational value” – first-order rational value according to the beliefs about first-order rational value in which she has credence. My proposal is that the rational value of an action relative to both levels of belief is simply the sum of the first-order rational value of that action, and a value given by a function of first-order rational values of the action according to each of the theories of first-order rationality, and the agent's credences in the respective theories. We might call this total rational value the Collapsed Rational Value of the action.

It's easiest to see how this would work in an extreme case like Tory's. The action with the highest rational value relative to her first-level beliefs is, of course, A. But according to the theory of first-order rationality she believes in, the action with the highest rational value relative to these first-level beliefs is B. To find out the Collapsed Rational Value of one of these actions, we simply add together its actual first-order rational value, and its mock first-order rational value – that is, its first-order rational value according to Tory's own theory thereof. We can see how this makes Tory's case sort of like Henry's. In Henry's case, what it was rational to do relative to one set of beliefs differed from what it was rational to do relative to another non-hierarchically conflicting set. So we just added up the rational values of each action relative to the different sets and arrived at an overall most rational action. In Tory's case, what it was rational to do relative to one belief differed from what it was rational to do relative to a hierarchically conflicting belief. So we added up the rational value of each action relative to the first belief and what the rational value relative to the first belief would be, if the second belief were correct.

Tory's case is very easy to handle for two reasons. The first is that she's certain of some (false) theory of first-order rationality, so it makes sense to speak of what would happen “if the second belief were correct”. But obviously, if someone is uncertain regarding a proposition, not all of her partial beliefs can be correct. If she were uncertain regarding first-order rationality, then as I said before, we would add to the actual first-order rational value of each action the value of some function of a) the first-order rational value of that action according to the theories of first-order rational value, and b) the agent's credences in those theories. What is this “some function”? It is the function given by the correct theory of second-order rationality – the correct theory of “what to do when you don't know what to do when you don't know what to do”.

So, for example, if she had a credence of .5 that, given her belief that it's objectively better to do A than to do B, it's more rational to do A, and .5 that it's more rational to do B, then to get the rational value of A, the correct theory of second-order rationality – call it expected rational value maximization – might have us add:

1. the actual first-order rational value of A to
2. (.5) □ (the first-order rational value of A if the first theory of first-order rationality is right) to
3. (.5) □ (the first-order rational value of A if the second theory of first-order rationality is right).

It also makes Tory's case easier that she only has two levels of belief. But suppose she had more. I see no reason why the procedure just adumbrated could not be re-iterated as follows:

We'd add:

1. The first-order rational value of A to
2. A function of a) the first-order rational values of A according to theories of first-order rationality, and b) the agent's credences in those theories, to
3. A function of a) the values of the sort of functions mentioned in step 2, and b) the agent's credences in those functions, to
4. A function of a) the values of the sort of functions mentioned in step 3 – which, in turn, are functions of the values of the sort of functions mentioned in step 2 – and b) the agent's credences in those functions...

...and so on to, get a the local rational value of A relative to all levels of credence.

To summarize: when an agent is beset with hierarchical rational conflict, we don't disregard her credences at all levels except one, and say that it's always most rational relative to all of the levels to do what's most rational relative to that level. Nor do we simply add together the rational values of different orders, and say that it's most rational to do the action that maximizes this sum. Rather, we add first-order rational value to what we might call "mock" first-order rational value, which depends on the agent's beliefs about first-order rational value, her beliefs about second-order rational value, her beliefs about third-order rational value, and so on. This allows us to do justice to all of her levels of belief in a way that seems principled rather than ham-handed.

Nozick and Ross on Hierarchical Rational Conflict

As far as I know, only two other writers have considered the question of practical rationality in cases of hierarchical rational conflict: Robert Nozick in The Nature of Rationality and Jacob Ross in his doctoral dissertation Acceptance and Practical Reason. (Actually, this is only true in the extensional sense. Nozick discusses rationality under uncertainty about rational norms, but he never actually diagnoses instances thereof as cases of rational conflict.) In this section, I'll discuss their approaches to the topic and contrast them with mine.

Nozick considers specifically the case of someone uncertain as to whether evidential decision theory or causal decision theory is the correct theory of rationality. (The former, recall, counsels "one-boxing" in Newcomb's example; the latter counsels "two-boxing".) Here is Nozick's suggestion:

> "Let CEU(A) be the causally expected utility of act A, the utility that act as it would be computed in accordance with...causal decision theory; let EEU(A) be the evidentially expected utility of act A, the utility of that act as it would be computed in accordance with evidential decision theory. Associated with each act will be a decision-value DV, a weighted value of its causally expected utility and its evidentially expected utility, as weighted by that person's confidence in being guided by each of these two kinds of expected utility.
>
> $$DV(A) = Wc \times CEU(A) + We \times EEU(A)$$
>
> And the person is to choose an act with maximal decision-value."[157]

It's important to note that Wc and We are not simply the agent's credences in

[157] Nozick (1994), p. 45.

causal and evidential decision theory, respectively. Nozick suggests that they will be influenced by those credences, but that they needn't be identical to them.[158]

There are several problems with Nozick's approach. To begin with, he doesn't seem to countenance local rationality, and is therefore unable to appreciate that different actions can be most rational relative to different sets of mental states. Certainly, relative to the agent's beliefs about how much money is in the boxes, and about how his choice is related to the perfect predictor's actions, it's most rational to do what the correct decision theory says, whatever one's credence distribution over the decision theories is. This makes his discussion a bit difficult to interpret.

On the most charitable reading, though, it seems that Nozick is offering an account of rationality relative to the beliefs mentioned in the previous paragraph *as well as* the agent's credences in the different decision theories. That is, he's giving an account of rationality relative to a set of beliefs that may stand in hierarchical rational conflict. Now, the most obvious feature of Nozick's answer is that it's hardly an answer at all. What it's rational to do is given by *some weighting or other* of the values according to the two decision theories? Not exactly stepping out on a limb.

But for all its non-specificity, Nozick's answer is still determinate enough to count as determinately *wrong*. For it is simply a version of the second approach to hierarchical conflict we surveyed earlier, according to which what it's locally rational to do in cases of hierarchical conflict depends only on the agent's highest-level credences. We simply ignore the fact that one or other decision theory *may be the correct theory of rationality* relative to the agent's beliefs about outcomes and the predictor's actions. What matters is

[158] Ibid.

only the agent's credences over the decision theories. An analogous answer as regards theoretical rationality would say that the rationality of conditionalizing on evidence depends only on my beliefs about the merits of conditionalizing, or that the rationality of inferring in accordance with modus ponens depends only on my beliefs about the merits of that inference rule. Whether conditionalizing or inferring in accordance with modus ponens is actually rational relative to the underlying beliefs is irrelevant, on this view, once I've ascended a level and formed beliefs about rationality. As I said before, I see little to recommend this blinkered approach.

Ross's treatment of the problem exhibits a greater clarity of reasoning than does Nozick's. However, there is a sense in which, rather than offering an answer to our question about rational action under hierarchical conflict, Ross pushes us further away from such an answer. This is due not to a flaw in his approach, but instead to its focus. As you'll recall, Ross's project is not to develop an account of rational action, but rather to develop a account of rational theory acceptance. Our projects converge at many points; for example, the Problem of Value Difference Comparisons is a problem for both of us. But here, pursuing a theory of rational acceptance actually makes it more difficult to develop an account of rational action. The reason for this is obvious upon reflection: The challenge in cases of hierarchical conflict is to say what's locally rational relative to all of the beliefs that give rise to the conflict. Ross tells us which theory of rationality to accept in such a case, not what to do. But this makes determining what it's rational to do all the more difficult. For now we must say of an agent who follows his advice what it's locally rational for her to do relative to all of the beliefs just mentioned plus this new mental state: acceptance. Now, one might say that, once a set of mental states includes the state of acceptance, none of the other states in that set play any role whatsoever in determining the rationality of actions relative to the set. But I can't see any merit in such a view.

Rather than try to solve the problem as it's complicated by this new mental state of acceptance, I want to see if Ross's view about acceptance in cases of hierarchical conflict has an action-concerning analogue we might fruitfully assess. I believe that it does. Ross argues that it's rational relative to a credence distribution over hypotheses about objective value and hypotheses about first-order rational value to accept the hypothesis about objective value such that acting on one's acceptance would yield the highest expected objective value.[159] One's credences in hypotheses about first-order rational value – hypotheses like EOV maximization, Credence Pluralitarianism, and so forth – play no role in determining what it's overall most rational to accept. The analogous theory about rational action would be the very first approach we considered, on which it's most rational to do the action that's most locally rational relative to one's credences in hypotheses about objective value; one's credences regarding theories of rationality relative to these first credences are completely irrelevant. Once we translate Ross's view about rational acceptance into a view about rational action, then, it turns out to be diametrically opposed to Nozick's.

As we observed earlier, though, this view seems to find support only in the thought that it's irrational to have mistaken views about rationality, coupled with the further thought that it's irrational to act on theoretically irrational views. As it turns out, Ross does argue for something like the irrationality of holding mistaken views about

[159] Ross (ms), p. 305-308. Ross says that when one is uncertain regarding rationality, there is a "context-independent requirement to accept what one does not believe" – namely, the correct theory of rationality.

rationality. Specifically, he argues that it's irrational to be under conflicting rational norms because being in such a state condemns one to practical irrationality, and since having mistaken views about rationality puts one in a state of rational conflict, it's irrational to have mistaken views about rationality.[160] Now, I agree that being under conflicting rational norms condemns one to practical irrationality in some sense. Someone who is mistaken about first-order rationality will not have available an action that's both as first-order rational and as second-order rational as at least one of the actions available to someone who's certain in the correct theory of first-order rationality. Put simply, a "purely" practically rational action is unavailable to someone who is mistaken about rationality, but is available to someone who is certain in the correct view of rationality.

What I deny, though, is that having a set of beliefs such that, if you had a different set of beliefs, you would could have done a more practically rational action than any action you in fact have available, renders the first set of beliefs irrational. The practical rationality of the actions that a belief enables does not reflect the rationality of the belief itself. (Similarly, the strength of the epistemic reasons that I have for a belief is not a function of the strength of the practical reasons supporting the actions I might perform based on that belief.) Instead, the theoretical rationality of a belief depends on precisely what you'd expect – its accordance or discordance with other beliefs and intentional states an agent has. And if the beliefs that condemn one to practical irrationality are not thereby irrational themselves, then it is a mystery why they should not play some role in determining what it's rational to do. And I should again insist again that, even if a belief is irrational, it's a mystery why it should not play a role in determining what it's rational to

[160] Ibid., p. 289-305.

do.

The lack of dependence between practical rationality and theoretical rationality, then, runs in two directions: A belief is not theoretically irrational simply because the actions to which it gives rise are less practically rational than the actions to which some other belief would've given rise, and an action is not practically irrational simply because the belief that gave rise to it is theoretically irrational. These two insights should lead us to resist the action-concerning analogue of Ross's theory of rational acceptance.

My suspicion is that Nozick and Ross are drawn to their rather severe views because the alternative – a ecumenical position on which the rationality of actions performed under hierarchical conflict depends on all of the beliefs that ground that conflict – barely shows up as a live option. It may be thought that such an approach to rationality under hierarchical conflict will either ham-handedly bulldoze that conflict, or else preserve it in the form of rational incomparability, either of which is unacceptable. But this is the sort of incomparability we can live with, once we realize that we can translate higher-order rational value into "mock" first-order rational value and thereby settle the matter of what, overall, is most rational to do.

Conclusion

In summary, there are two legitimate worries about those who are uncertain "all the way up" about practical rationality: First, how can their behavior be norm-guided if there are no norms of behavior of which they are certain? The answer is that we can guide our behavior by norms about which we're uncertain. These may be EP-relative norms. But they may also be objective norms; the function of rational norms, then, is not necessarily to guide action, but rather to specify the circumstances in which belief-based action counts as truly guided at all.

Second, what do we say about the rational value of their actions, given that different actions are most locally rational relative to different “levels” of their normative thinking? The answer is that there are orders of rational value corresponding to each level of normative thinking, and that the degrees of value of one order are incomparable with the degrees of value of the other orders. But this is okay, because it's not exactly the values of the different orders that we add to determine the rationality of actions; rather, we use the values of the different orders to determine what the first-order values of actions would be relative to higher-level beliefs, and add these “would be” first-order values to the actual ones.

AFTERWORD

Rather than conclude this dissertation by explicitly summarizing its arguments, I'll instead note five areas where more work is needed. These are not the only such areas, but they are, in my view, the most important and the most interesting of them.

Meta-taxonomy

There is a meta-taxonomical elephant in the room that I've only barely acknowledged. There are two stances we might take towards the distinction between objective value, rightness, and so forth, and all of the different subjective variants of the same. One stance is that this is a distinction between substantive answers to univocal questions about which actions have the highest value, or which actions are right, etc. That is, there's the "objectivist" position on such questions, and the "subjectivist" position, and they clash in the same way that utilitarianism and deontology clash. Another stance is that this is a distinction between different concepts. That is, one might consistently provide different answers to the questions of what one objectively ought to do, and of what one belief-relatively ought to do, in some situation; as a corollary, the simple question, "What ought one to do?" is underspecified.

One might adopt one of these stances towards some distinctions of this sort, and the other stance towards other such distinctions. For example, I might hold the "different concept" view as it regards the broad distinction between objective and subjective value, but then also think that the "belief-relative view" and the "evidence-relative view" provide conflicting answers to the question of what one subjectively ought to do. (I think this particular hybrid position is very common, actually.) One might also adopt the second stance towards a distinction, but think that some of the concepts distinguished are somehow less-than-legitimate. You might say, for example, "Sure, you can define up 'non-normative evidence relative rightness', or whatever you want to call it, but that concept plays no important role in our practices [or plays no role in the practices we should have, or fails to carve nature at the normative joints, etc.]." This is what I said about "C-Relative Value" in Chapter 6. There's a distinct concept there, but it's useless, and there's a decent chance that nothing falls under it.

It should be clear where I've stood throughout this dissertation. I've tended to regard the distinctions in question as marking off different concepts, rather than different substantive positions. My credence is fairly high that I've been right in doing so, but plenty of smart people seem to adopt the contrary position. Papers with titles like "Ought: Between Objective and Subjective"[161] and "Is Moral Obligation Objective or Subjective"[162] attest to that. And this contrary position in practical philosophy is mirrored by an even more widely-held position in epistemology – that the "internalist" and "externalist" are offering conflicting answers to perfectly well-formed, univocal questions about justification and knowledge.

So why take my side on this meta-taxonomical issue? Two reasons. First, we should think of these distinctions as conceptual because there are distinct roles that each of these putative concepts seems to play. Belief-relative normative concepts play at most a minor role in our practice of giving advice, for example. However, they are indispensable when it comes to specifying which actions count as norm-guided, or which

[161] Kolodny and MacFarlane (ms).

[162] Zimmerman (2006).

actions count as faithful to our normative beliefs. Second, we should not think of these distinctions as substantive, on the grounds that we don't experience them that way in deliberation. I can find myself torn between doing what I think utilitarianism suggests and what I think non-consequentialism suggests, but I cannot find myself torn between doing what I think I have objective reasons to do, and doing what I think I have belief-relative reasons to do. For it's by doing what I have belief-relative reasons to do that I make my best attempt at doing what I have objective reasons to do. Judgments about belief-relative reasons are "transparent" to judgments about objective reasons.

It will be worth saying more about this issue in the future, especially because it's just one battle in the intensifying war over which philosophical clashes are substantive and which are terminological.

The Roles of Normative Concepts

There's another reason for us to inquire into the nature of normative concepts. I argued in Chapter 4 that we cannot get different normative hypotheses "on the same scale" without adverting to the roles in thought of normative concepts as such. I made some suggestions about what these roles may be, but that's all they were – suggestions – and I didn't say anything about how these roles interact – how they might be weighed off against each other in determining when a particular normative concept has been tokened.

There are still several big-picture questions about concepts that will have to be answered before we can responsibly turn our attention to normative concepts specifically. As far as I can see, the going view in cognitive science is that the primary bearers of intentionality are concepts qua mental entities. But many philosophers still think primarily in terms of Fregean concepts that we may or may not “have” or “grasp”, or else regard intentionality as inhering originally in public language terms, and only derivatively in mental entities. Even after we explain why mentalistic concepts have the semantic information they do, there's a further question of how this information is stored – as a set of necessary and sufficient conditions, as prototypes that items falling under the concept must resemble, or as a set of inferential relationships to other concepts and non-inferential relationships to motivations, feelings, experiences, and so forth.[163]

After we're satisfied with answers to these questions, we can ask about the concept-constitutive features of OUGHT, MORE/LESS REASON, A RATIO OF VALUE DIFFERENCE A TO VALUE DIFFERENCE B OF 3:1, and the rest. I surmised in Chapter 4 that these features will be manifold and complexly-related, and I stand by that. It would be shocking if even the simplest of these concepts could be characterized in terms of a small set of necessary and sufficient conditions. (Crazy confession: I suspect that most of the fundamental normative concepts are partly phenomenal, and that psychopaths, say, lack PHENOMENAL WRONG, for example, just as Frank Jackson's “Mary” lacked PHENOMENAL RED.[164]) Admittedly, my sense of this is colored by my view that concepts understood as mental entities are the primary bearers of meaning, and that their meanings are determined in accordance with what cognitive scientists labelled “theory theory” – a theory of conceptual semantics with notoriously holistic implications.[165]

[163] The literature here is vast, but many of the key contributions are anthologized in Laurence and Margolis (1999). I found the editors' introduction very helpful as well.

[164] See Jackson (1982).

[165] Ibid., Chapters 19 and 20. For criticism of theory theory and other holism-supporting views, see Fodor and Lepore (1992).

The bad news for those trying to solve the PVDC is that these issues are unlikely to be resolved any time soon. The good news is that they're issues of incredibly general interest, and so there are a lot of talented people working on them.

The Thinker's and Agent's Perspectives

This dissertation emerged from the vague idea that there are certain relations to my thoughts and my actions that only I may bear, and that normative theory is incomplete unless part of it plays a role that's engaged in these relations. Asking about rational action under normative uncertainty is just one path that leads us to a confrontation with the questions, “What are these relations?” and “What is it about me qua thinker and qua agent such that I bear these relations uniquely?”

The last chapter's section on action guidance was an attempt at a partial answer to the first question. There I offered a theory about how one must be related to a set of cognized norms and an action such that one's performing that action counts as guided by those norms. My theory, you'll recall, is that the relation has a normative component. I must do the action that is rational given my beliefs regarding those norms in order for my action to count as guided by them. But I think this theory requires further defense than I've been able to give it. At any rate, it's incomplete. For one thing, I followed my discussion of action guidance with a discussion of hierarchical rational conflict that added further, and unaddressed, complications. When I'm under such conflict, there may be no action that's rational on all “orders”, as I was calling them. Does this mean that I can't in that case perform an action that counts as fully norm-guided? I'm not sure. The theory is also incomplete because there are clearly elements to action guidance other than normative and brute causal relationships between beliefs and actions. There's a further element of control or direction of one's actions that I'm still struggling to characterize rigorously.

Then there was the problem, raised by Ruth Chang in conversation and Niko Kolodny in print, of whether I'm bearing this sort of privileged relationship to my beliefs when I modify them in accordance with my further beliefs about rationality (as opposed to my further beliefs about reasons). I'm inclined to think that I am, but I'm not certain, which is why I tucked this issue away in a footnote. There's some plausibility to the thought that, in modifying my beliefs in accordance with my beliefs about rationality, I'm not "changing my mind" in the privileged sense of the phrase, but rather "changing a mind over which I have an especially high degree of control". If that's true, then this sort of belief revision isn't truly reasoning, and so there are no rational norms that govern it. Again, I had a suggestion about how to answer the Chang-Kolodny challenge that invoked the possibility of an item's being "judgment-sensitive under a description", and of the description of a belief or an action as a "try" as the one under which it is sensitive to judgments about rationality. But of course, that's not an answer to this challenge, but a mere gesture towards one.

Expected Value Maximization

I wish I were able to come up with more convincing positive arguments that maximizing expected value is the uniquely rational thing to do under uncertainty. As I said in Chapter 2, we can't simply assume this is right, and arguments that would conclusively show as much if they were successful – and here I'm thinking of Dutch

Book arguments – don't seem to work. So I think at this stage, all we have in our arsenal are intuitions about particular cases, which don't seem to favor expected value maximization over other reasonable theories, as well as more impressionistic “theoretical intuitions” like the ones to which I appealed in the “Long Run, Wide Run, etc.” and “Winning Percentage” arguments. Of course, once we see that expected value maximization is just another kind of distribution insensitivity and that its competitors are varieties of distribution sensitivity, it shouldn't surprise us that we lack conclusive arguments either way. After all, it doesn't look as though the major debates in population ethics – between utilitarianism, prioritarianism, egalitarianism, sufficientarianism, etc. – will be settled any time soon. Anyhow, it would be nice if we could develop some firmer ground on which to prosecute this sort of debate.

Sin, Conscience, and the Will

Finally, there's the historical matter of how the Catholic theologians' debate about Reflex Principles might be used to shed light on their conceptions of sin, conscience, and the will. Recall that Reflex Principles take probabilities of material sinfulness as inputs, and yield verdicts about formal sinfulness as outputs. But a formal sin is supposed to be a sin of the conscience, and it seems wrong to say that practically irrational actions are always performed with a defective conscience. In saying that, though, I'm subtly taking one side of perhaps the most important debate in medieval moral philosophy – the debate between the intellectualists, who conceived of bad actions as reflecting defects of the intellect alone, and voluntarists, who conceived of bad actions as always reflecting defects of the will as well.[166] The side I'm taking is the voluntarist side, at least if we conceive of formal sin as the sort of bad action over which the intellectualists and voluntarists were feuding. On this view, there is a gulf between following the wrong Reflex Principle and acting badly in the relevant sense, that can only be bridged by the operation of a bad will. The participants in this debate obviously seem to have held the opposite view.

If my diagnosis is right, then it suggests a tension in their thinking that deserves further consideration. For the story of medieval moral philosophy is of the retreat from intellectualism – on which, to put it bluntly, you could go to Hell for mere cognitive impairment – and the consequent embrace of voluntarism.[167] That the Rigorists thought you could commit a formal sin merely by following the Probabilist Reflex Principle, and *vice versa*, arguably represents intellectualism's last stand, manifested in the thought that a defect *in* oneself is necessarily a defect *of* oneself. If this is wrong, then we're left with the interesting question of what *does* constitute the latter sort of defect. But if this lingering intellectualism is right, then the stakes associated with the present project are higher than even its author had supposed.

[166] This debate is helpfully summarized and contextualized in Schneewind (1998), Chapter 2.

[167] Ibid.

BIBLIOGRAPHY

Adams, R.M. (1995), 'Moral Faith', Journal of Philosophy 92 (2), 75-95.

Allais, M. (1953), 'Criticisms of the postulates and axioms of the American School." In Rationality in Action: Contemporary Approaches, ed. Paul K. Moser. Cambridge University Press, 1990.

Anderson, E. (1995), Value in Ethics and Economics, Harvard University Press.

Arendt, H. (1965), Eichmann in Jerusalem: A Report on the Banality of Evil, Penguin Books.

Arrow, K. (1951), Social Choice and Individual Values, Wiley and Sons.

Ayer, A.J. (1952), Language, Truth and Logic, Dover.

(1957), 'The Concept of Probability as a Logical Relation', in S. Korner (ed.), Observation and Interpretation in the Philosophy of Physics, Dover Publications.

Blackburn, S. (1984), Spreading the Word: Groundings in the Philosophy of Language, Oxford University Press.

Broome, J. (1995), Weighing Goods, Blackwell.

(2004), Weighing Lives, Oxford University Press.

Buchak, L. (ms #1), 'Risk Sensitivity'

(ms #2) Risk and Rationality, Ph.D. Dissertation, Princeton University.

(ms #3), 'Risk Without Regret'

Bunyan, J. (1887), The Pilgrim's Progress, Hodder and Stoughton.

Carroll, L. (1895), 'What the Tortoise Said to Achilles', Mind 104 (416), 278-80.

The Catholic Encyclopedia (1913), available at http://www.newadvent.org/cathen/

Chang, R. (1997), Incommensurability, Incomparability, and Practical Reason, Harvard University Press.

(2001), Making Comparisons Count, Routledge.

(2002), 'The Possibility of Parity' Ethics, Vol. 112, No. 4, , 659-688.

(2009), 'Voluntarist Reasons and the Sources of Normativity', In D. Sobel and S. Wall (eds.), Reasons for Action, Cambridge University Press.

Chernoff H. (1954), 'Rational selection of decision functions', Econometrica 22 (4), 422–43.

Dancy, J. (2004),'Enticing Reasons', in R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (eds.) Reason and Value: Themes from the Moral Philosophy of Joseph Raz, Oxford University Press, 91-118.

De Finetti, B. (1937), 'La Prévision: Ses Lois Logiques, Ses Sources Subjectives', Annales de l"Institut Henri Poincaré 7, 1-68, in Studies in Subjective Probability, H. E. Kyburg, Jr. and H. E. Smokler (eds.), Robert E. Krieger Publishing Company.

de Medina, Bartolomé (1577), Expositio in 1am 2ae S. Thomae.

Dreier, J. (1993), 'Structures of Normative Theories', The Monist 76, 22-40.

Elga, A. (forthcoming) 'How to Disagree about How to Disagree', in R. Feldman and T. Warfield (eds)., Disagreement, Oxford University Press.

Elster, J. and Roemer, J.E. (eds.) (1991), Interpersonal Comparisons of Well-Being, Cambridge University Press.

Feldman, F. (2006), 'Actual Utility, The Objection from Impracticality, and the Move to Expected Utility', Philosophical Studies 129 (1), 49-79.

Fodor, J. and Lepore, E. (1992), Holism: A Shopper's Guide, Blackwell Publishers.

Foot, P. (1958), ‘Moral Beliefs’, Proceedings of the Aristotelian Society 5, 983-104.

Gärdenfors, P. and Sahlin, N.-E. (1982), ‘Unreliable Probabilities, Risk Taking, and Decision Making’, Synthese 53, 361-386.

Gert, J. (2007), ‘Normative Strength and the Balance of Reasons’, Philosophical Review, 116 (4), 533-62.

Gibbard, A., (1990), Wise Choices, Apt Feelings, Oxford University Press.

(2003), Thinking how to Live, Cambridge University Press.

Good, I.J. (1967), ‘On the Principle of Total Evidence’, British Journal for the Philosophy of Science 17 (4), 319-21.

Graves, M. (1989), ‘The Total Evidence Theorem for Probability Kinematics’, Philosophy of Science, 56 (2), 317-24.

Guerrero, A. (2007) ‘Don’t Know, Don’t Kill: Moral Ignorance, Culpability and Caution’, Philosophical Studies 136 (1), 59-97.

Hajek, A. (2005), ‘Scotching Dutch Books?’, Philosophical Perspectives 19 (1), 139–51.

Hammond, P. (1976), ‘Why Ethical Measures of Inequality Need Interpersonal Comparisons, Theory and Decision, 7 (4), 263-74.

Harsanyi, J.C. (1967), ‘Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility’, Journal of Political Economy, 75 (5), 765-66.

Hegel G.W.F. (1943), Philosophy of Right, T.M. Knox (trans.), Oxford University Press.

Hesse, H. (1981), Siddhartha, Bantam Classics.

Howard-Snyder, F. (1994), **‘**The Heart of Consequentialism**’,** Philosophical Studies, 76 (1), 107-29.

Hudson, J.L. (1989), ‘Subjectivization in Ethics’, American Philosophical Quarterly, 26 (3), 221-29.

Hurley, S. (1992), Natural Reasons: Personality and Polity, Oxford University Press.

Jackson, F. (1982), ‘Epiphenomenal Qualia’, Philosophical Quarterly 32, 127-36.

(1991), ‘Decision-Theoretic Consequentialism and the Nearest and Dearest Objection’, Ethics 101 (3), 461-82.

(1998), From Metaphysics to Ethics: A Defense of Conceptual Analysis, Oxford University Press.

Jeffrey, R. (1990), The Logic of Decision, University of Chicago Press.

Kagan, S. (1998), Normative Ethics, Westview.

Kamm, F.M. (1993), Morality, Mortality: Death, and Whom to Save From It (vol.1), Oxford University Press.

(1996), Morality, Mortality: Rights, Duties and Status (vol.2), Oxford University Press.

Kant, I. (1972), Groundwork of the Metaphysics of Morals, H.J. Paton (trans.), Hutchinson University Library.

Kolodny, N. (2005), ‘Why be Rational?’, Mind 114 (455), 509-63.

Kolodny, N. and McFarlane, J. (ms) ‘Ought: Between Subjective and Objective’

Kripke, S. (1979), ‘A Puzzle about Belief’, in A. Margalit (ed.). Meaning and Use, Dordrecht.

Kyburg, H. (1978), ‘Subjective Probability: Criticisms, Reflections, and Problems’, Journal of Philosophical Logic, 7 (1), 157-80.

Laurence, S. and Margolis, E. (1999), Concepts: Core Readings, The MIT Press.

Levi, I. (1974), ‘On Indeterminate Probabilities’, Journal of Philosophy 71 (13), 391-418.

(1982), ‘Ignorance, Probability, and Rational Choice’, Synthese 53, 387-417.

(2000), ‘Money Pumps and Diachronic Dutch Books’, Philosophy of Science, 69, 235-47.

Lillehammer, H. (1997), ‘Smith on Moral Fetishism’, Analysis, 57 (3), 187–95.

Lindley, D.V. (1971), Making Decisions, Wiley and Sons.

Liguori, St. A. (1852), Theologia Moralis, 2nd ed. (1755), R.P. Blakeney (trans.), Reformation Society.

Lockhart, T. (2000), Moral Uncertainty and its Consequences, Oxford University Press.

Loewer, B. (1993), ‘The Value of Truth’, Philosophical Issues 4, 265-80.

Luce, R.D. and Raiffa, H. (1957), Games and Decisions: Introduction and Critical Survey, John Wiley and Sons.

Maher, P. (1993), Betting on Theories, Cambridge University Press.

Matthen, M. (2009), ‘Chickens, Eggs, and Speciation’, Noûs 43 (1), 94-115.

McClennan, E. (1990), Rationality and Dynamic Choice: Foundational Explorations, Cambridge University Press.

McIntyre, A. (1981), After Virtue, University of Notre Dame Press.

Moller, D. (ms), ‘Abortion and Moral Risk’

The New Catholic Encyclopedia, 2nd ed. (2002), The Catholic University Press.

Norcross, A. (1997), ‘Good and Bad Actions’, Philosophical Review 106 (1), 1-34.

Nozick, R. (1977), Anarchy, State, and Utopia, Basic Books.

(1994), The Nature of Rationality, Princeton University Press.

Oddie, G. (1995), ‘Moral Uncertainty and Human Embryo Experimentation’, in K.W.M. Fulford, G. Gillett, J.M. Soskice, Medicine and Moral Reasoning, Oxford University Press.

Parfit, D. (1984), Reasons and Persons, Oxford University Press.

(1997), ‘Equality and Priority’, Ratio 10 (3), 202–21.

Pascal, B. (1853), Provincial Letters, T. M’Crie (trans.), Robert Carter and Brothers.

Peacocke, C. (1995), A Study of Concepts, MIT Press.

(2004), The Realm of Reason, Oxford University Press.

Portmore, D. (2003), ‘Position-Relative Consequentialism, Agent-Centered Options, and Supererogation’, Ethics, 113 (2), 303-32

(2009), ‘Consequentializing’, Philosophy Compass 4 (2), 329-47.

Prummer, D. (1957), Handbook of Moral Theology, Roman Catholic Books (Revised: 1995).

Putnam, H. (1986), ‘Rationality in Decision Theory and in Ethics’, Crítica 18 (54), 3-16.

Quinn, W. (1989), ‘Actions, Intentions and Consequences: The Doctrine of Double Effect’, Philosophy and Public Affairs 18 (4), 334-51.

Ramsey, F.P. (1926), ‘Truth and Probability’, in F. Ramsey (1931), The Foundations of Mathematics and Other Logical Essays, R.B. Braithwaite (ed.), Harcourt, Brace and Company, 156-98.

(1929), ‘Probability and Partial Belief’, in F. Ramsey (1931), The Foundations of Mathematics and other Logical Essays, R.B. Braithwaite (ed.), Harcourt, Brace and Company, 256-7.

Ramsey W., Stich S., and Garon, J. (1991), ‘Connectionism, Eliminitavism and the Future of Folk Psychology’, Philosophical Perspectives 4, 499-533.

Raz. J. (1975), ‘Permission and Supererogation’, American Philosophical Quarterly

12,161–8.

(1988), The Morality of Freedom, Clarendon Press.

Resnick, M. (1987), Choices: An Introduction to Decision Theory, University of Minnesota Press.

Robbins, L. (1938), ‘Interpersonal Comparisons of Utility: A Comment’, The Economic Journal 48, 635-41.

van Roojen, M. (1996), ‘Expressivism and Irrationality’, Philosophical Review 105 (3), 311-35.

Ross, J. (2006), ‘Rejecting Ethical Deflationism’, Ethics 116, 742-68.

(ms), Acceptance and Practical Reason, Ph.D. Dissertation, Rutgers University.

Scanlon, T.M. (1998), What We Owe to Each Other, Harvard University Press.

Schneewind, J.B. (1998), The Invention of Autonomy: A History of Modern Moral Philosophy, Cambridge University Press.

Schick, F. (1986), ‘Dutch Bookies and Money Pumps’, Journal of Philosophy 83 (2), 112-19.

Schroeder, M. (2006), ‘Not So Promising After All: Evaluator-Relative Teleology and Common-Sense Morality’, Pacific Philosophical Quarterly 87(3), 348-56.

(2007), ‘Teleology, Agent-Relative Value, and ‘Good’’, Ethics 117(2), 265-95.

(2008), Being For: Evaluating the Semantic Program of Expressivism, Oxford University Press.

Sen, A. (1970), Collective Choice and Social Welfare, Holden-Day.

Sepielli, A. (2006), ‘Review of Ted Lockhart’s Moral Uncertainty and its Consequences’, Ethics 116 (3)

(2009), ‘What to Do When You Don't Know What to Do’, in R.Shafer-Landau (ed.) Oxford Studies in Metaethics, vol. 4, Oxford University Press.

(ms #1), ‘Apriority, Analyticity, and Normativity’

(ms #2), ‘Evidence, Rationality, and Disagreement’

(ms #3), ‘Subjective Norms and Action Guidance’

Sidgwick, H. (1893), ‘My Station and its Duties’, International Journal of Ethics 4 (1), 1-17.

Skyrms, B. (1990), The Dynamics of Rational Deliberation, Harvard University Press.

Smith, H. (1983), ‘Culpable Ignorance’, Philosophical Review 92 (4), 543-71.

(1988), ‘Making Moral Decisions’, Noûs 22 (1), 89-108.

(forthcoming) ‘Subjective Rightness’, Social Philosophy and Policy.

Smith, M. (1994), The Moral Problem, Blackwell.

(2002), ‘Evaluation, Uncertainty and Motivation’, Ethical Theory and Moral Practice 5 (3), 305-20.

Smith, M. and Jackson, F. (2006), ‘Absolutist Moral Theories and Uncertainty’, Journal of Philosophy 103, 267-83.

Stocker, M. (1990), Plural and Conflicting Values, Oxford University Press.

Strawson, G. (2004), ‘Against Narrativity’, Ratio 17 (4), 428-52.

Taurek, J. (1977), ‘Should the Numbers Count?’, Philosophy and Public Affairs 6 (4), 293-316.

Taylor, C. (1989), Sources of the Self: The Making of Modern Identity, Cambridge University Press.

Temkin, L. (1987), ‘Intransitivity and the Mere Addition Paradox’, Philosophy and Public

Affairs, vol. 16, no. 2, 138-87.

(1996), ‘A Continuum Argument for Intransitivity’, Philosophy and Public Affairs, Vol. 25, No. 3, 175-210.

(2005), ‘A “New” Principle of Aggregation’, Philosophical Issues 15 (1), 218–34.

(forthcoming) Rethinking the Good: Moral Ideals and the Nature of Practical Reason.

Thomson, J. (2008), Normativity, Open Court.

Weatherson, B. (2002), ‘Review of Moral Uncertainty and its Consequences’, Mind, 111 (443), 693-96.

(ms) ‘Disagreeing about Disagreement’

Wedgwood, R. (2001), ‘Conceptual Role Semantics for Moral Terms’, Philosophical Review 110 (1), 1-30.

Weirich, P. (1986), ‘Expected Utility and Risk’, British Journal for the Philosophy of Science 37 (4), 419-42.

Wiggins, D. (1979), ‘Truth, Invention and the Meaning of Life’, Proceedings of the British Academy LXII, 331-78.

Williams, B. (1981), ‘Persons, Character, and Morality’, in A. Rorty (ed.), The Identities of Persons, University of California Press.

Wittgenstein, L. (1953), Philosophical Investigations, G.E.M. Anscombe (trans.), Basil Blackwell.

Wolf, S. (2009), ‘Moral Obligations and Social Commands’, in L. Jorgenson and S. Newlands (eds.), Metaphysics and the Good: Themes from the Philosophy of Robert Merrihew Adams, Oxford University Press.

Yalcin, S. (forthcoming), ‘Non-Factualism about Epistemic Modality’, in A. Egan and B. Weatherson (eds.), Epistemic Modality, Oxford University Press.

Zimmerman, M. (1996), The Concept of Moral Obligation, Cambridge University Press.

(2006), ‘Is Moral Obligation Objective or Subjective?’, Utilitas 18, 329-361.

(2009), Living With Uncertainty, Cambridge University Press.

Curriculum Vita

Andrew Christopher Sepielli

Education

1997-2001
Princeton University, A.B. (Philosophy), 2001

2001-2004
Yale University, J.D. (Law), 2004

2004-2009
Rutgers University, Ph.D. (Philosophy) 2010

Employment History

2009-
University of Toronto, Assistant Professor (Philosophy)

List of Publications

2006
"Review of Ted Lockhart's Moral Uncertainty and Its Consequences," Ethics 116: 601-3.

2009
"What to Do When You Don't Know What to Do," in Oxford Studies in Metaethics, Vol. 4 (Russ Shafer-Landau, editor), Oxford University Press.