Maximising Expected Value Under Axiological Uncertainty

An Axiomatic Approach

Stefan Riedener

St John's College

University of Oxford

February 2015

Thesis for the Degree of Doctor of Philosophy

Word Count: 74'925 words

Meinen Eltern

und meinem Bruder

Philosophy is, today, not a pastime. It is inescapable, because we no longer believe to know what is good. ... [We] are unable to sneak out of the moral point of view, yet there is nobody who tells us what it is.

Ernst Tugendhat

in: Steve Pyke, *Philosophers.*

Abstract

The topic of this thesis is axiological uncertainty – the question of how you should evaluate your options if you are uncertain about which axiology is true. As an answer, I defend Expected Value Maximisation (EVM), the view that one option is better than another if and only if it has the greater expected value across axiologies. More precisely, I explore the axiomatic foundations of this view. I employ results from state-dependent utility theory, extend them in various ways and interpret them accordingly, and thus provide axiomatisations of EVM as a theory of axiological uncertainty.

Chapter 1 defends the importance of the problem of axiological uncertainty. Chapter 2 introduces the most basic theorem of this thesis, the Expected Value Theorem. This theorem says that EVM is true if the betterness relation under axiological uncertainty satisfies the von Neumann-Morgenstern axioms and a Pareto condition. I argue that, given certain simplifications and *modulo* the problem of intertheoretic comparisons, this theorem presents a powerful means to formulate and defend EVM. Chapter 3 then examines the problem of intertheoretic comparisons. I argue that intertheoretic comparisons are generally possible, but that some plausible axiologies may not be comparable in a precise way. The Expected Value Theorem presupposes that all axiologies are comparable in a precise way. So this motivates extending the Expected Value Theorem to make it cover less than fully comparable axiologies. Chapter 4 then examines the concept of a probability distribution over axiologies. In the Expected Value Theorem,

this concept figures as a primitive. I argue that we need an account of what it means, and outline and defend an explication for it. Chapter 5 starts to bring together the upshots from the previous three chapters. It extends the Expected Value Theorem by allowing for less than fully comparable axiologies and by dropping the presupposition of probabilities as given primitives. Chapter 6 provides formal appendices.

Acknowledgements

Writing this thesis would have been impossible without all the help that I have received. I am very grateful for it.

The Clarendon Fund and the Swiss Study Foundation have both provided very generous financial support. For many helpful conversations on the topic of this thesis, I thank Samuel Hughes, William Jefferson, Harvey Lederman and Trevor Teitel. I am grateful to the audience at the 2014 Ujué Workshop on Topics of Practical Philosophy, and particularly to Anna Goppel for inviting me there. And I thank the audience at the Oxford DPhil Seminar, and Ralf Bader for commenting on my presentation of material from this thesis.

Brian Hedden, Robyn Kath, William MacAskill, Lukas Naegeli, Bastian Stern, Aron Vallinder and Silvan Wittwer have generously read parts of the thesis. They have provided countless excellent comments on sometimes painfully sketchy drafts. Nick Davies, Edi Karni, Robert Nau, David Schwartz and Teruji Thomas have kindly helped me with some formal aspects of the thesis. I want to thank them all.

It is difficult to say how grateful I am for having had John Broome and Hilary Greaves as my supervisors. They have been unbelievably generous with their time, and have offered me more encouragement and philosophical insights than I could possibly have hoped for. Beyond alls its rigour and clarity and its breathtaking scope, there is something in John's writing that has always moved me. His work is so *sincere*, and somehow it is animated with a rare humanity. With all of these qualities, philosophical and otherwise,

he has supported me. The impact of his ideas has been profound, but he has taught me much more than ideas. Hilary has supervised me during the last period of writing. Her brilliant ability at spotting difficulties with my thoughts or ways to take them further sometimes brought me to the point of desperation. But her kind encouragement and keen philosophical guidance always brought me back. I will ever be very deeply grateful to both.

I thank my friends in Zurich and Oxford for being there for me even after all my absences. The absences have been sadly frequent and sadly long for me. I hope they know how much their friendship meant to me, and how much they have actually been present during all the days at lonely desks.

For that invaluable sense of being unconditionally supported, on which everything was built, I thank my parents Sabine Furthmann and Hanspeter Riedener, and my brother Lukas.

Finally, I thank my fiancée, Anna Koim. I am at home with you, with your warmth and joy. And with your questions and your open eyes you keep me moving and awake. You make my life grow; you fill it and yet you make it light. How wonderful that this should only be the beginning, that there are houses and journeys awaiting us, unheard music and quiet feasts, so many mornings and winters and summers.

Contents

Chapter 1

The Problem of Axiological Uncertainty

1.1 The Basic Question

Introduction

Our lives are rife with uncertainty, and yet we constantly have to act. So uncertainty is the condition of almost any decision we make.

Part of this uncertainty concerns *non-normative* questions. When ordering a dinner at a restaurant, we are often uncertain about what exactly we will get; in starting a new job, we rarely know precisely how that will turn out; and in trying to alleviate climate change, we will not be certain about the efficacy and ramifications of different ecological measures. Indeed, we can rarely predict the long- or even the short-term effects of our actions. For practical philosophy, this raises the question about what you ought to do if you are non-normatively uncertain.

But often, we are also uncertain about purely *normative* questions. In response to climate change, for example, we may not be certain about whether we morally ought to be impartial in weighing the relevant different interests, or whether we ought to give more weight to currently living people over future generations, or to people we know over distant strangers, or to human beings over animals. More generally, we often do not know what the correct norms of morality, or rationality, or any other system of norms, are. Such uncertainty will often occur along with non-normative uncertainty, but it could even obtain if we were non-normatively certain. It raises the question about what you ought to do if you are normatively uncertain.

This thesis is about this latter, normative kind of uncertainty. More precisely, it is about a particular kind of normative uncertainty: uncertainty about moral value – about which options are *morally better* than which. I shall call this *axiological uncertainty*, since the part of morality that deals with moral value is called axiology. So my core question is about what you ought to do if you are axiologically uncertain. Or more precisely still, it is about which options are morally better than which if you are axiologically uncertain.

Axiological uncertainty is only a very specific form of normative uncertainty. Not only are there, plausibly, normative facts other than those of morality; but even morality itself may be broader than axiology. Perhaps morality also features facts about what we *ought* to do, and indeed facts that have nothing to do with moral value – obligations of fairness or justice, say. For reasons that will emerge, I take my narrower question to be somewhat less difficult than the more general questions of moral or normative

uncertainty. So I shall ignore uncertainty about all these other normative facts, and focus on axiological uncertainty only.

The Meaning of the Question

Some terminology will be helpful. I said that my core question is which options are morally better than which if you are axiologically uncertain. In one sense, there seems to be a trivial and indisputable answer to this question: an option is better than another, it seems, if and only if it is better according to the *true* axiology.

However, this is not the answer I am concerned with. To see that there might be a different one, consider first a case of ordinary non-normative uncertainty:

> **Example 1**. Blue is suffering from a mild form of pain. Red has a pill that he could give him, but Red is very uncertain about whether the pill is a pain reliever or a lethal form of poison.

Suppose that, as a matter of fact, the pill is a pain reliever. Is it then better to give this pill to Blue, or better not to give it?

Cases like this have caused an extensive debate with respect to the deontic concept 'ought' – about whether there are two distinct concepts of 'ought' ('subjective' and 'objective') or just a single one, about which one it would be, or which of the two is ultimately more basic or important.[1] I shall not go into the details of that debate. It seems undeniable to me that *something*

[1]Cf. e.g. Hudson (1989), Jackson (1991), Howard-Synder (1997), Wiland (2005), Feldman (2006), Zimmerman (2008), Bykvist (2009b), Broome (2013, ch.3), Kolodny and MacFarlane (ms).

has to be said, and can be said, both in favour of giving and in favour of not giving the pill. On the one hand, we can evaluate Red's options on the basis of their actual outcomes, without taking his uncertainty into account. When we do that, we will say that giving the pill is better. After all, Red will thereby simply relieve Blue's pain, and nothing else will happen. On the other hand, we can evaluate Red's options on the basis of the prospects that they represent, taking his uncertainty into account. When we do that, we will say that *not* giving the pill is better. After all, Red will otherwise risk Blue's death for the sake of a mild pain. We may be interested in either of these judgments, and both uses of 'good' seem familiar. I shall remain neutral about which of them is ultimately more important, or more in line with our common concept of goodness (or 'ought').

A parallel example can be given concerning axiological uncertainty:

> **Example 2**. Red has a pill, with which he can either relieve Blue of a mild pain, or relieve a non-human animal of a significantly greater pain, but not both. Red is very uncertain about the moral value of animal welfare – uncertain, that is, between a speciesist and a non-speciesist axiology.

Suppose that according to the speciesist axiology it would be better to benefit Blue, and according to the non-speciesist axiology it would be better to benefit the animal. Furthermore, suppose for the sake of argument that the speciesist view is right.

Then again, we can evaluate Red's options on the basis of their actual value, without taking his uncertainty into account. When we do that, we will say that it is better to benefit Blue. After all, this is what the true

axiology says. However, we can again evaluate Red's options on the basis of the prospects that they represent, now taking his *axiological* uncertainty into account. And if we do that, it at least remains an open question which act is better, under Red's state of uncertainty. By benefiting Blue he after all risks performing an act that is comparatively bad. Judgments of this kind are perhaps less familiar from our ordinary practice, but they are no less distinct than in the non-normative case.

I am concerned with judgments of this latter kind. To distinguish them conceptually, I will use the concept '*uncertainty-relative value*', or '*u-value*' to denote the goodness we refer to by taking your axiological uncertainty into account. I will use the simple terms '*value*', '*goodness*' or '*betterness*' to denote the goodness we refer to *without* taking into account your axiological uncertainty. In Example 2, I shall thus say that it is better to benefit Blue, but (so far) an open question which of the two options is u-better. And this is the core question of my thesis: which options are u-better than which, if you are uncertain about which options are better than which? This question has no trivial and indisputable answer.

The Importance of the Question

Axiological uncertainty is a very important phenomenon. Normative value theory is difficult, and hence many – if not all – of our most momentous decisions are decisions under axiological uncertainty. Let me illustrate this with a few examples.

Take questions of population ethics. For instance: can a sufficiently large

population of people with lives barely worth living be better than a smaller population with people living wonderful lives? Some people find it very plausible that this is the case,[2] while others find it almost impossible to believe,[3] and there are arguments for both sides. More basically: can it even be good to bring people into existence? Most people have a strong intuition that it can normally not, but that intuition faces serious objections.[4] Or relatedly: to what extent are the values of populations involving different numbers of people even comparable? While some have argued that there is only some vagueness in how they compare,[5] others claimed that such populations are never comparable,[6] and still others defended a middle ground between these views.[7] Population ethics is notoriously difficult. In fact, there are numerous impossibility theorems showing that a number of intuitively plausible principles of population ethics are incompatible.[8] We should be far from certain about any answers in that field.

However, each of these questions is key to evaluating our potential responses to the largest challenges for humanity – such as global poverty, climate change, or (other) catastrophic risks like plagues and pandemics, artificial intelligence or terrorism and war. In all of these cases, among else, the number and identity of future people is at stake, and so they raise population-ethical questions. For the same reason, and closer to home, the above questions are of primary importance when we decide on any policy of public

[2]Cf. e.g. Huemer (2008).
[3]Cf. e.g. Temkin (2012) and Parfit (1984).
[4]Cf. Broome (2004, ch.10).
[5]Cf. Broome (2004, ch.12).
[6]This was suggested to me by Ralf Bader in conversation.
[7]Cf. e.g. Parfit (ms).
[8]Cf. e.g. Arrhenius (forthcoming).

health, security or migration, say. And they also matter greatly for very personal decisions each of us faces – such as whether or not we use our resources to save or improve the lives of other people, thereby perhaps allowing them to have enormous chains of children, grand children, great grand children, and so on.[9] From the largest decisions of humanity to the most personal questions of how we conduct our lives, our decisions plausibly or potentially affect the population, and thus involve vexed population-axiological questions, and axiological uncertainty.

Moreover, in all these examples, there is no way to avoid making a decision under axiological uncertainty. We cannot somehow wait until we acquired axiological certainty or knowledge. Or more precisely, even to 'wait' and do moral philosophy to acquire moral certainty or knowledge would be a decision. So this only raises the question whether *that* would be the u-best thing to do.[10]

But population ethics is just one example. Consider the question of what is good *for* a being. Very plausibly, this question is relevant to assessing the moral value of options. But we are by no means certain about it. For example, would it be bad for sentient beings to have extensive pleasure artificially stimulated in their brains, to have wonderful lives simulated in experience machines, or to undergo serious enhancements through drugs or genetic engineering? While to some such scenarios sound horrific,[11] others have decidedly embraced them.[12] And these questions are expectably of great practical im-

[9]For the practical importance of population ethics, cf. particularly Broome (2004, ch.1) and Beckstead (2013, ch.1).

[10]Cf. MacAskill (2014, ch.6) for a discussion of this question.

[11]Cf. e.g. Nozick (1974, 42ff.).

[12]Cf. e.g. Ng (1997, 1849ff.) and Crisp (2006).

portance, as pleasure stimulations and enhancements will plausibly become a serious possibility in the future. Or take questions about goodness in time: does the wellbeing of future beings have the same value as that of currently existing ones, or should we discount the value of wellbeing over time? There seem to be very good arguments against discounting,[13] but not discounting future lives has implications that contradict common sense morality radically.[14] This issue too is extremely important, as very many of our actions potentially have ramifications and ripple effects until a very distant future. Finally, consider questions about the moral value of animal welfare again. For example, is it bad that animals suffer in the wild, are being eaten by predators or die from starvation and diseases? Again, our common sense practice seems to disregard almost all of this suffering. But there are very good arguments that such suffering is bad, indeed *very* bad, given the astronomical number of wild animals.[15] Examples can be multiplied with ease. Decision making under axiological uncertainty is unavoidably very common.

So at least *prima facie*, it is very important that we reflect on how to evaluate options in the face of axiological uncertainty. I say '*prima facie*', because even granting that axiological uncertainty affects our most momentous decisions, one might raise several objections against the importance of u-value. And it is in fact not altogether easy to say why exactly it is important to reflect about u-value. I shall discuss some of these objections in section 1.3. But before I turn to them, let me outline in more detail what I shall do in this thesis.

[13]Cf. e.g. Broome (2004, ch.4).
[14]Cf. e.g. Beckstead (2013).
[15]Cf. e.g. McMahan (2010).

1.2 This Dissertation

I shall now first state the main goal, the approach and main results of this thesis in general terms; I then provide a synopsis of the later chapters; and then I briefly mention some problems that I shall set aside.

Goals, Approach, and Relationship to the Existing Literature

Again, my core question is which options are u-better than which. As the title of the thesis indicates, the answer I shall give to this question is *Expected Value Maximisation*, or *EVM*. This view says, roughly, that an option is u-better than another if and only if it has the greater expected value – where the expected value of an option is a weighted sum of the values it is assigned by the axiologies, with weights representing the probabilities of these axiologies. So, roughly, the main goal of this thesis is to defend EVM as the correct answer to my core question.

As the subtitle of the thesis indicates, I shall pursue an axiomatic approach to this goal. I shall introduce formal theorems to the effect that EVM follows from a certain set of axioms; and I shall claim that these results are philosophically important in various ways. If my arguments are sound, these theorems can help us formulate what EVM means and help us understand its nature and normativity; they can help us vindicate that EVM is true; and they can also help us assess the prospects of formulating and defending EVM as a general theory of normative (rather than merely axiological) uncertainty. So more precisely, I might say that the goal of this thesis is to explore the *foundations* of EVM.

Unfortunately, there are various versions of EVM. I do not think that we can be certain of any particular one; nor indeed do I think that we can be certain that EVM – rather than some other theory of axiological uncertainty – is correct. It is beyond the scope of this thesis to provide full-blown axiomatic foundations for different versions of EVM, let alone for different theories of axiological uncertainty. So for most part, I shall simply elaborate on a form of EVM that I find very convincing, and that lends itself most readily to an axiomatisation. But wherever possible, I shall also indicate how alternative versions of EVM could be axiomatised, and start to explore the axiomatic foundations of theories of axiological uncertainty other than EVM. So in most precise terms, the goal of this thesis is to outline possible axiomatic foundations for one version of EVM, and to indicate possible axiomatic foundations for other versions of EVM as well as for alternative theories of axiological uncertainty. Thus I hope to argue for one specific theory of axiological uncertainty, while at the same time providing at least the beginnings of an overview over the space of possible views and their respective foundations.

To pursue this axiomatic approach, I shall rely as much as possible on the existing literature about non-normative uncertainty. More specifically, I shall rely on representation theorems from decision theory. These theorems are standardly interpreted as grounding facts about what your preferences should be under non-normative uncertainty. In this thesis, I shall take such theorems from decision theory, extend them in various ways and reinterpret them accordingly, and thus apply them to the context of axiological uncertainty. In this context, I shall claim, they can ground facts about the u-value relation. In some respects, this will naturally raise similar questions and

problems as it does in the non-normative context. But it also raises various very specific issues. So as I hope my thesis will show, it is worth investigating the axiomatic foundations of Expected Value Maximisation under axiological uncertainty specifically.

This has not been done. Quite generally, the theory of normative uncertainty is still in its infancy.[16] In the wake of Ted Lockhart's (2000) pioneering work, various people have endorsed EVM; some obvious similarities between the problem of normative and that of non-normative uncertainty have been mentioned; and various authors have expressed the idea that 'decision theory' might vindicate EVM.[17] However, as far as I am aware, the theory has not received a decision-theoretically sophisticated, or axiomatic exploration. The reason for this may be that philosophers thought the application of expected utility theory completely trivial and unproblematic,[18] or that they took it to be philosophically uninteresting.[19] Be it as it may, the possible axiomatic foundations of EVM – or indeed the problem of axiological or normative uncertainty generally – have not been thoroughly investigated. I shall try to fill this lacuna in this dissertation.

Synopsis

After this introductory chapter, the thesis will feature four main chapters,

[16] As far as I am aware, less than a dozen publications address the general problem; cf. Hudson (1989), Gracely (1996), Lockhart (2000), Weatherson (2002), Sepielli (2006; 2009), Ross (2006b), Guerrero (2007), MacAskill (2013), Gustafsson and Torpman (2014); cf. also Hedden (forthcoming), Harman (forthcoming); Ross (2006a), Sepielli (2010), MacAskill (2014) for PhD theses. Moral uncertainty is sometimes mentioned in specific debates in applied ethics; cf. e.g. Pfeiffer (1985), Oddie (1994), Moller (2011), Broome (2012, 183ff.).

[17] Cf. e.g. Ross (2006b, 753ff.); also MacAskill (2014, ch.1).

[18] This seems to be true for Ross (2006b, 753ff.).

[19] This seems to be true for Sepielli (2009, 27; 2010, 169).

followed by a chapter of appendices. In the remainder of the present chapter, I shall answer some objections against the importance of u-value, and thus further motivate the project of formulating and defending a theory of axiological uncertainty.

In chapter 2, I shall first introduce the decision-theoretic framework that I employ in this thesis – that of so-called state-dependent utility theory – and show why this framework is particularly congenial to the theory of axiological uncertainty. I shall then reproduce a simple representation theorem from state-dependent utility theory. Together with certain background assumptions and extended by an additional axiom this result will imply the most basic theorem of this thesis – the Expected Value Theorem. This theorem says, roughly, that EVM is true if the u-value relation satisfies the von Neumann-Morgenstern axioms and a Pareto condition with respect to the underlying axiologies. I shall suggest that, given certain restrictions and simplifications, and *modulo* the problem of intertheoretic comparisons, this theorem presents a powerful means to formulate and defend EVM.

In chapter 3, I shall then discuss the problem of intertheoretic comparisons – the question whether sizes of value differences or heights of value levels can be compared across axiologies. I first state an argument to the effect that absolute scepticism about intertheoretic comparisons has unacceptably implausible implications, and thus that at least some intertheoretic comparisons must somehow be possible. Assuming that this is true, I then discuss different accounts of *why* it is true, and begin to outline for each of these accounts how the Expected Value Theorem can be interpreted or extended so as to be compatible with it. On the account I shall ultimately find most convincing,

some plausible axiologies are not comparable in a precise way. The Expected Value Theorem presupposes that all axiologies are comparable in a precise way. So this motivates the search for weaker conditions than those of the Expected Value Theorem, so that our theory of axiological uncertainty can cover less than fully comparable axiologies.

In chapter 4, I then focus on the concept of a credence distribution over axiologies. In the Expected Value Theorem, this concept figures as an unexplained primitive. I argue that this is ultimately unsatisfying, and that we need an account of what this concept means. In line with my general approach, I then employ and extend a representation theorem from decision theory to provide an explication for it. According to this explication, roughly, the credence you have in an axiology is the weight you give that axiology in your u-value judgments. Like other explications that are based on representation theorems, this account has important implications for the normative structure of EVM. In light of this, I shall discuss various recent objections against the significance of representation theorems for the purposes of defining credences. At least with respect to axiological uncertainty, I shall answer these objections and argue that representation theorems provide the best account of credences, and thus that they can and should play an important role in grounding EVM.

In chapter 5, I start to bring together the upshots from the previous three chapters. I shall combine results from two branches of decision theory – state-dependent utility theory, and utility theory without the Completeness axiom – and state the most comprehensive theorem of this thesis. The result is a representation theorem for state-dependent utilities without the Com-

pleteness axiom, which at least goes some way towards dropping the presupposition of axiological credences as primitives. This will allow the theory of axiological uncertainty to cover axiologies that do not compare in a precise way.

At the end of each of these main chapters I shall pursue a 'further exploration': I shall discuss the possibilities of axiomatising views other than the main theory outlined in the thesis. Thus I shall consider theories of axiological uncertainty other then EVM (in chapters 3 and 4); I shall consider whether we can defend expected value maximisation as a theory of uncertainty about theories of axiological uncertainty (in chapter 2); and I shall examine whether the axiomatic framework of this thesis can be extended beyond axiological uncertainty to cover general *moral* uncertainty (in chapter 5).

Problems that I Ignore

Let me briefly flag three problems that I shall not discuss in this thesis. First, I shall not address any metaethical issues underlying the problem of normative uncertainty. My questions presuppose a notion of axiological 'truth', and a notion of an agent's 'credences' in an axiology. As far as the metaethics is concerned, I shall simply take these notions for granted. It is a separate and difficult question to what extent my questions arise, or can be made sense of, within non-cognitivist or anti-realist metaethical views. Some people think that non-cognitivists cannot account for the phenomenon of normative uncer-

tainty;[20] others think they can.[21] And in any case, the ability to make sense of normative uncertainty is generally treated as a desideratum on metaethical views: if such a view cannot account for normative uncertainty, that is standardly taken to be a problem for this view, rather than a problem for the project of normative uncertainty.[22] I think that this is a plausible view of the dialectic. But be that as it may, I shall ignore any metaethical dimension in this dissertation.

Secondly, I shall not discuss the question about the deontic status of your u-best option. It might be that, *ceteris paribus*, if one of your options is u-best, you are under a normative (narrow-scope[23]) requirement to choose that option. Or you might be under a normative (wide-scope) requirement to the effect that, if one of your options is u-best, *ceteris paribus*, you choose it. You might also be under both kinds of requirement, or under neither. And perhaps these requirements would be requirements of morality, or of rationality, or something else again. This involves large questions. They are beyond the scope of this thesis.

Thirdly, as I already indicated, I shall not say anything more about the conceptual questions surrounding different orders of goodness. In particular, I shall not discuss the existence of a putative notion of overall-goodness. Suppose that your best option is a, but that given your axiological uncertainty, your u-best option is b. Indeed, suppose you are also uncertain about theories of axiological uncertainty, and that according to the true theory of

[20] Cf. e.g. Smith (2002), Bykvist and Olson (2009).
[21] Cf. e.g. Sepielli (2012).
[22] Cf. e.g. Smith (2002).
[23] Cf. Broome (2013, ch.7) for the terminology of 'wide-' or 'narrow-scope requirements'.

uncertainty about theories of axiological uncertainty, your u^2-best option (as we might call it) is c. We might imagine that the orders of value go ever higher up. This raises the question about whether there is some overall-best option, beyond or besides the different orders of value. I think that there is no such notion of overall-betterness.[24] But this is a conceptual debate on its own. I shall not enter this debate in this thesis (though I shall briefly discuss a related issue arising through a potential regress on page 25).

1.3 Objections

Before moving on to the main chapters, let me address some objections concerning the importance of u-value.

The Fetishism Objection

Some people have argued that we do not need a theory of u-value (or of moral uncertainty more generally) because concern with u-value is *fetishistic*. To see what this means, suppose that you have a choice between a vegan meal and a steak, and that you are uncertain between a speciesist and a non-speciesist view about the value of animal welfare. Suppose that you find the speciesist view much more plausible. However, you believe that *if* the non-speciesist view is right, killing animals is *very* bad, and so you think it is u-better to get the vegan meal. Suppose finally that you do get the vegan meal *because* you think it is u-better. Then your vegan diet does not spring

[24] Cf. Sepielli (2013a, 13ff.) for a defence of a related view.

from a core concern that you have for *animals*, it seems. Rather, you seem to care about (u-)value *as such* – about whatever turns out to be (u-)valuable, simply because it is (u-)valuable. And as Michael Smith pointed out in a similar context, we might think that there is something inappropriate – 'fetishistic' – about that. Smith says:

> Good people care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality, and the like, not just one thing: doing what they believe to be right, where this is read *de dicto* and not *de re*. Indeed, common-sense tells us that being so motivated is a fetish or moral vice, not the one and only moral virtue. (1994, 75)

Smith objected to *de dicto* concern with what is morally right. But similar considerations arise in the context of axiological uncertainty: concern with u-value might seem inappropriately fetishistic.

Brian Weatherson, for example, believes this. Concering the case of meat-eating I just described, he says: 'it would be perverse for [you] to turn down the steak. To do so [you] would have to care about morality, whatever it is, not about the list of things that Smith rightly says a good person will care about' (2014, 13). And this is one of the main reasons for why he concludes that 'being uncertain about the physical consequences of your actions should matter both to what you do, and how you are assessed. [...] But a mere probability that meat eating is immoral should not change one's actions, or one's evaluations of meat eaters' (2014, 2).[25]

However, it is not entirely clear what the 'inappropriateness' of being motivated by u-value should consist in, and what conclusion we should draw

[25]The fetishism objection is also raised in Hedden (forthcoming, 22).

from it. So let me suggest three interpretations of this fetishism-objection, and answer each of them in turn.

On the first interpretation, the inappropriateness of being motivated by u-value is a matter of first-order value: it is *better* to help animals (say) out of true and deeply felt concern with their wellbeing, rather than to help them because one believes that their welfare matters morally or dominates u-value in a particular case. According to the true axiology, concern with u-value is comparatively bad.

If this is the claim, I do not deny it. It might be better to be moved *de re* rather than *de dicto* by what is valuable. This is simply an axiological claim, and nothing I say contradicts it. In fact, it can simply be accommodated in the theory I outline. What I shall defend is only a criterion of u-betterness. It is not a decision procedure, an account of how to consciously make or be motivated to make decisions.[26] If an action can be performed under different motivations, we can treat these motivations simply as different options – on a par with the option of doing some completely different action. So if you believe that being motivated by u-value is bad, my theory will imply that it is comparatively u-bad to be consciously motivated by it. This is perfectly consistent. Of course, if we could be *certain* that *de dicto* concern with u-value is extremely bad, then my theory will be self-effacing: it will imply that we should almost or literally never consciously act on considerations of u-value. And although that would still not render it inconsistent, it might raise the question why one should spend much time thinking about it. But

[26]For the classic distinction between a criterion of rightness and a decision procedure in utilitarianism, c.f. e.g. Bales (1971), Mill (1861, ch.2, Par.19) or Sidgwick (1907, 413).

we should clearly not be *certain* that this is so – let alone that *de dicto* concern with u-value is *so* bad as to dominate all other possibly bad things. So if the inappropriateness is understood as a standard claim about value, it does by no means establish that we do not need a theory of u-value.

On a second interpretation, this inappropriateness has a slightly different form. In the above quote, Smith says that 'good people' are motivated by what is valuable *de re*. Perhaps this does not mean that such motivation is bad. Rather, it might mean – in a virtue ethical spirit – that we are not ideal moral agents if we are motivated by u-value. Moral philosophy does not need a theory of u-value, this objection goes, because ideal moral agents are not motivated by u-value.

Again, perhaps this is true. Perhaps an ideal moral agent will always consciously be motivated by what is valuable, understood *de re*. However, the entire project of examining normative uncertainty is based on the fact that we are *not* ideal moral agents. We are inescapably uncertain about what is valuable, *de re*; and our project is to determine how *we* non-ideal agents should evaluate our options. So the question about what an ideal moral agent will do is simply a different question from the one we are addressing. Again, nothing I say contradicts any answer to that question. Presumably, an ideal moral agent would always (*ceteris paribus*) choose the objectively best option. But that in itself does not show that *we* need not reflect about u-value. Similarly, perhaps an ideal moral agent would be motivated by what is valuable, *de re*. But that in itself does again not show that *we* do not need to reflect about u-value. So for all that this second interpretation says, it still seems that u-value is important for us non-ideal agents.

But here is a third interpretation: perhaps the objection is not only that ideal agents are not motivated by u-value, understood *de dicto*. Perhaps the objection is that even non-ideal agents should not be so motivated. Understood thus, the objection is not that we do not need *any* theory of what to do under normative uncertainty. Instead, the objection advocates a specific such theory, according to which, under uncertainty, we should always intuitively follow something that we care about *de re*. This claim does indeed address the question of this thesis, and it contradicts the theory that I outline.

But I do not think that this claim is convincing. To begin with, this alternative theory is not very useful. If we are axiologically uncertain, there might not *be* anything that we unqualifiedly care about *de re*. Very often, we will be torn between different values, and then the present theory will not tell us what to do. But more importantly, the present idea simply seems rather crazy. As I outlined in section 1.1, we have to make enormously important moral decisions in the face of uncertainty – about existential risks, global poverty, human enhancements, and so on. Making a morally bad decision concerning any of these issues might have enormous ramifications, involving vast numbers of beings until some very far future. It just seems very implausible that we should run the risk of incurring such astronomical badness, just to avoid a particular kind of motive.

I cannot think of any other interpretation of the fetishism objection. So I conclude that, although being motivated *de re* by what is good may in some ways be preferable to being motivated *de dicto* by u-value, this does not imply that we do not need a theory of u-value.

Harman's Objection

Let me turn to another objection against the importance of u-value, which is due to Elizabeth Harman.[27] Harman argues against what she calls 'Uncertaintism', and defends 'Actualism' instead. She says various things about these views, and I am not sure whether I understand what she means by them. Her paper is entitled 'The Irrelevance of Moral Uncertainty', and thus the core dispute between these two views seems to be about the importance of u-value or moral uncertainty more generally. Thus she says that 'according to Uncertaintism, an agent's moral uncertainty (and specific moral credences) are crucially relevant to how the agent should act' (forthcoming, 1). In contrast, according to Actualism, an agent's 'moral beliefs and moral credences are usually irrelevant to how she (subjectively) should act' (forthcoming, 5).

Harman's main argument for the 'irrelevance of moral uncertainty' is this:

(A) Uncertaintism implies that if you are certain of a false moral theory and act in accordance with that theory, you are not blameworthy;

(B) if you are certain of a false moral theory and act in accordance with that theory, you are blameworthy; therefore

(C) Uncertaintism is false.

To illustrate premise (B), she considers someone who works for a Mafia family and is certain that he has a moral obligation of loyalty that requires him to kill innocents when it is necessary to protect the family. Harman claims that he would be blameworthy if he in fact killed an innocent person to protect the

[27] Hedden (forthcoming, 20ff.) tentatively raises the same objection.

family, notwithstanding his belief in the respective obligation. This is so, she argues, because '[a] person is blameworthy for her wrongful behavior just in case it resulted from her failure to care *de re* about what is morally important – that is, from her failure to care adequately about the non-moral features of the world that in fact matter morally' (forthcoming, 13). Hence, an agent's moral beliefs are generally irrelevant for whether he is blameworthy. And in spite of his belief – or perhaps as his belief shows – our agent fails to care adequately about the wellbeing of innocents.

In response, let me first say that I do not share the view that this person is blameworthy – or at the very least, that we know remotely enough about him to say that. I think that one may blamelessly acquire false moral beliefs, and then blamelessly act upon them. This is so, I think, even if the beliefs and the resulting acts are as bad as those of Harman's Mafia member. But it is particularly so with regards to the difficult axiological questions I mentioned in section 1.1. Given the widespread disagreement, very many people will have false moral beliefs about these issues. I do not think that these people are necessarily blameworthy if they act upon their beliefs. So I think that (B) is false.

This permissive view about blameworthiness would be enough to block Harman's objection. However, I shall not defend it here. It raises large questions,[28] and I think Harman's argument fails even if we grant (B) and her rigorist view of blameworthiness. That is, I think I *can* grant (B): whatever Harman exactly means by 'Uncertaintism', nothing in *my* view implies that if you are certain of a false moral theory and act in accordance with that

[28]Cf. particularly Rosen (2003; 2004) for criticism of (B).

theory, you are not blameworthy. This is a claim about blameworthiness, and thus an entirely separate issue. I am simply not making any claims about blameworthiness in this thesis, nor am I committed to any such claims.

Harman thinks that Uncertaintism is committed to the permissive view of blameworthiness because she believes that

(D) you are blameworthy for some behaviour only if you should not have behaved in that way, given your beliefs and credences.

And at one point, she seems to define Uncertaintism as the view that, 'given [your] beliefs and credences', you should maximise the expected moral value of your actions (forthcoming, 4). Together with principle (D), this claim seems to imply that if you are certain of a false moral theory and act in accordance with it, you are not blameworthy.

However, I think this is not a helpful characterisation of 'Uncertaintism'; at least, it is not an adequate characterisation of my view. The clause 'given your beliefs and credences' does not unequivocally pick out a precise sense of 'ought'. Consider again Example 2 on page 4: we assumed that Red is uncertain between a speciesist and non-speciesist axiology, that according to the speciesist axiology it would be better to benefit a person, and according to the non-speciesist axiology it would be better to benefit an animal, and that the speciesist view is right. Then what should Red do, given his beliefs and credences? Again, I think there is no unequivocal answer to this question. In one (objective) sense, his credences are of course irrelevant, and so even given his beliefs and credences, he should benefit the person. So in one sense, I agree, it is false that if you are certain of a false moral theory then, given

your beliefs and credences, you should maximise the expected value of your actions.

So it is not true that principle (D) commits me to the permissive view about blameworthiness. If Harman's rigorist view of blameworthiness is correct, the 'should' that figures in this principle is not relative to your axiological credences. And that is something that I could perfectly well agree with. Again, I am simply not making any claim about blameworthiness.

Perhaps Harman would claim that if I believe your axiological credences are irrelevant for whether you are blameworthy or not, I am not a true 'Uncertaintist'. In fact, in one passage, she seems to define Uncertaintism as including the permissive view about blameworthiness.[29] So if 'Uncertaintism' is *defined* in this way, and if I were to accept her rigorist view of blameworthiness (which I do not), then I would simply not be an 'Uncertaintist'. But that would by no means mean that I had to accept (with her title) the 'irrelevance of moral uncertainty', or that I would have to deny that our 'moral uncertainty (and specific moral credences) are crucially relevant' (as one apparent definition of Uncertaintism says). Harman's argument seems to manifest a very sad view of the role of value and u-value, making an extraordinary fetish of blameworthiness. It suggests that the most important or 'relevant' thing is that we avoid being blameworthy. But this is surely not so. In light of our momentous decisions concerning climate change, human enhancements, global poverty and so on, our main goal should not be to deserve praise, or to avoid being blameworthy. It should quite simply be to do, or try to do, *good*. And it is important to have a theory of u-value since

[29] Cf. the last line of the 'Uncertaintist reasoning' she outlines on page 1.

we are not certain about what is good. It is a concern for doing good that motivates the theory of axiological uncertainty. At least as I understand it, the core relevance of u-value has nothing to do with questions of praise and blame.

In conclusion, I think Harman's rigorist view of blamewohrtiness is false; and even if it were true, it can by no means establish the 'irrelevance of axiological uncertainty'. The importance of u-value does not hinge on questions of praise and blame at all.[30]

Regress and Action-Guidance

I have not yet said why exactly it is important to reflect about u-value. The most prominent motivation for subjective forms of value is the thought that, while we are rarely in a position to know the objective values of our options, we are generally in a position to know their subjective values (e.g., their u-values). In this sense, it is often said, theories of subjective value can be action-guiding whereas theories of objective value cannot.[31] However, as I already hinted at on page 15, there is a regress problem with this claim. Like being uncertain about the true axiology, we might be uncertain about the true theory of axiological uncertainty. So if due to uncertainty about theories of value, we are not in a position to know which options are best

[30]One might argue that the concept of value is somehow conceptually tied to blameworthiness, and that if I thus separate u-value from blameworthiness I can no longer claim that u-value really is a form of value (or that it might have implications for an 'ought'). But for one thing, I doubt that the concept of value is thus tied to blameworthiness. For another thing, I do not care very much about the conceptual question whether u-value is ultimately truly a form of value. What matters is that it is important to reflect about when an option is 'u-better' than another, however exactly we call it. And considerations of blameworthiness do not challenge that.

[31]Cf. e.g. Hudson (1989).

and our axiologies thus allegedly fail to be action-guiding, then it seems that due to uncertainty about theories of u-value, we are neither in a position to know which options are u-best, and our theories of u-value will also fail to be action-guiding. But if our interest in u-value was motivated by the failure of standard axiologies to be action-guiding, and theories of u-value (or any u^n-value) are not action-guiding either, then why should we be interested in u-value (or any u^n-value) in the first place?

I shall not object to the claim that we are not in a position to know the u-value relation. I think we can justify reflection about u-value even if we accept this claim. Unfortunately, however, I cannot give a fully worked out justification. So what I shall do is simply to outline the direction where I think the most plausible justification for exploring theories of u-value lies.

First of all, it is worth saying that axiological uncertainty need not imply that axiologies cannot be action-guiding. At least in a standard sense of that term, we can be 'guided' by norms even if we are not certain about them. Consider the norm that it is best to get 0.8 grams of protein per kilogram of body weight each day. Presumably, we cannot be entirely certain about that norm, nor can we be certain about what it implies in our everyday practice. But the norm seems perfectly action-guiding, in any standard sense of that term. It's not that, without some extra algorithm about how to deal with our uncertainty, we would be completely paralysed, or that any action we do would be a complete stab in the dark. Presumably, if we act in the face of normative uncertainty but without explicitly considering theories about normative uncertainty, we do so an on implicit and maybe unquestioned acceptance of some such theory. We can do that. By the same token, even if

we are uncertain about axiologies, that does not mean that they cannot be action-guiding in any sense.

Why, then, do we need to explore theories of u-value? One answer one might give is that, even though it is in principle possible to be guided by norms about which one is uncertain, there are moments when our axiologies are unable to guide us. One might argue that this is so when our uncertainty becomes *conscious*. While our axiologies might guide us as long as we are not conscious of our uncertainty, at least once we acknowledge our axiological uncertainty, we may feel that our actions would indeed be mere stabs in the dark. And at least in these moments, one might argue, it might be that only a theory of u-value could guide us (even though we may not be altogether certain about that either).

However, as it stands, this response is at least insufficient. It might be a psychological fact that when we are conscious of our axiological uncertainty, we need a theory of u-value to *feel* guided; and it might even be a fact that only our acting on considerations of u-value will then *count* as 'guided'. But there still remains the *normative* question of why our acting on the basis of considerations of u-value is normatively important. The question is not whether an action counts as, or feels like, being guided. As Andrew Sepielli (2014) has pointed out: if we really do feel paralysed by nagging uncertainty, a glass of whisky might make us more confident, and thus prompt committed and 'guided' action more quickly than a complex normative theory. So the mere fact that theories of u-value will sometimes guide us when our axiological uncertainty would otherwise make us feel at loss, *in itself*, does not show why u-value is important. The real question is why acting on considerations

of u-value is normatively important, or any better than acting while unconscious of one's uncertainty, or by (what is and feels like) a stab in the dark, or after a glass of whisky.

Sepielli (2014) gives an argument for why it is. If I understand him correctly, his main point is that it is 'less risky' (2014, 92) to reflect about u-value and act after serious reflection – even if that reflection does not lead to certainty – than to act without any further reflection simply on the implicit assumption of some theory. In the passage that comes closest to an argument for this claim, he says: 'there is no relevant difference between moral reasoning and evidence-gathering, and there is almost always some value from the agent's perspective to gathering additional evidence [and acting upon that new evidence], as I.J. Good demonstrated in his classic 1967 paper "On the Principle of Total Evidence"' (2014, 93).[32]

I think that this is on the right track. But it is important to be clear about what this suggested line of reasoning would show, and what would be needed to turn it into a more full-fledged argument. What Good (1967) proved was that the *expected value* of the action that has the highest expected value after taking into account more evidence is always greater than that of the action with the highest expected value before we take this evidence into

[32]Sepielli also says: 'From this perspective [once you have reflected about theories of u-value, or 'meta-theories'], it must surely seem better to implicitly accept and act on a meta-meta-rule that, as it were, aggregates the opinions of the various meta-rules you find plausible, than to simply act on one of those meta-rules as though there were no alternatives. [...] we might say that while acting on this meta-meta-rule is unguided relative to acting on some competing meta-meta-rule [...], it is not unguided relative to acting simply on a meta-rule' (2014, 93). However, I think these points are red herrings. What matters is not whether it '*seems* better' from one perspective, or whether one action is 'guided' relative to more alternative actions. What matters is whether reflected action *is better* in some sense.

account (if the cost of taking it into account is neglected, and unless the action with the highest expected value would be the same for whatever piece of evidence we gather). So *if* the analogy with Good's argument is sound, then the expected value of acting after reflection about u-value will be greater than that of acting without that reflection (if the cost of moral reasoning is neglected). The expected value of reflecting about one level of value will be characterisable in the next meta-level of value. So in other words, if the analogy is sound, then if expected value maximisation is the true theory of u^2-value, acting after reflection about *u-value* is u^2-better than acting while unconscious of one's uncertainty, or through a stab in the dark, or after a glass of whisky.

It would take a lot of argument to turn this into firm claim. The analogy is not unproblematic. For example, Good assumed that whatever evidence we receive, we will do what maximises expected value, given that evidence. But if we receive good evidence for a theory of u-value that says we should *not* maximise expected value, that is arguably a dubious assumption. Moreover, one would clearly have to characterise formally what acting 'without further reflection' would amount to; one would have to argue more fully for the parallel between evidence-gathering and moral reflection; one would ultimately have to take into account the cost of moral reflection too, and thus say something about the efficacy of moral reasoning. And of course, we would also have to argue that EVM is the correct theory of u^2-value. We are a long way from having a full-fledged argument for the claim that acting after serious reflection about u-value will be u^2-better than acting unreflectedly. But it does seem very plausible, just like it seems plausible that acting after

reflection about value will be u-better than acting unreflectedly.[33]

There is a next worry that the proponent of the regress objection might raise. To show that reflecting about u-value is good, we have to assume some form of EVM – or perhaps some other normative theory with the same implication. More generally, to show that more reflected action will be less risky (normatively better, rather than just feeling better or being more guided), we have to assume some normative criterion about what is risky. The proponent of the regress objection might claim that we are begging the question against her. She might argue that we are not entitled simply to assume EVM, or any other theory, but instead would have to take into account uncertainty about that again, and so on. In other words, she might demand that we show not only that reflection about u-value will plausibly be u^2-better, but also that it is u^3-better, and perhaps u^n-better all the way up (or overall-best in some sense).

But note what the objector is now demanding. She is in effect asking us to prove beyond any doubt that our project is *ex ante* worth pursuing. I do not see how we could do that. But it seems to be an exceptionally high demand. In particular, that demand would question not only why we should reflect about what is u-valuable; it would also question why we should reflect about what is *valuable* – or indeed why we should reflect on anything, or do anything else that seems worthwhile, rather than just drink whisky and do whatever we feel like. The objector raises a most radical normative scepticism. That may be an interesting philosophical problem. But it is not a problem that I can address in this thesis. And I hope it is not one that I

[33]This last claim is defended in MacAskill (2014, ch.7).

have to address to motivate the project of this thesis.

I shall proceed on the – plausible seeming, though not much further substantiated – assumption that, given the prevalence and importance of axiological uncertainty, reflection about u-value, just like much other reflection, is important. So the main motivation for examining and developing theories of u-value, I think, is not that we are always in a position to know the u-value relation. We may not be. The main motivation, I think, is that it is plausibly *ex ante* valuable to reflect about how we should respond to axiological uncertainty and act on that reflection rather than to act in the face of it without further reflection.

Chapter 2

The Expected Value Theorem

Introduction

As I said in section 1.2., the most important goal of this thesis is to outline what one may call the *foundations* of EVM under axiological uncertainty: what EVM means and presupposes, and what axiomatic basis we can give for it. In this chapter, I begin to pursue this goal by introducing the most basic formal result of this thesis – the Expected Value Theorem. This theorem says, roughly, that if u-value satisfies the von Neumann-Morgenstern axioms and a Pareto condition, EVM is true. This result is interesting in itself, but it will also prove helpful for introducing some general features and problems of my approach, and provide a fruitful starting point for the more specific extensions pursued in subsequent chapters.

In section 2.1, I outline in general terms what role axiomatic theorems can play in the theory of axiological uncertainty, and why such theorems are important.

In section 2.2, I first introduce the specific decision-theoretic framework that I shall be using in this thesis – that of state-dependent expected utility theory – and show why it is particularly useful for our purposes. I then reproduce the specific representation theorem on which the main result of this chapter is based.

In section 2.3, I apply this theorem to the context of axiological uncertainty. This application will require some conceptual assumptions, as well as an additional substantial condition. I shall introduce these claims in some detail, state the actual theorem they support – the Expected Value Theorem – and indicate some implications of that theorem.

In section 2.4, I discuss the conditions of the Expected Value Theorem in more depth. Within the limits of this thesis, I cannot provide a full-blown defence of these conditions. So what I shall do is to compare their plausibility in the context of axiological uncertainty with their status in other contexts. The main worry that these conditions raise, I shall argue, is the problem of intertheoretic comparisons. I shall suggest that, given certain assumptions, the theorem is sound if all axiologies are fully comparable and do not compare in a lexical way. This will motivate a thorough discussion of the problem of intertheoretic comparisons in chapter 3.

In section 2.5, I explore briefly whether something like the Expected Value Theorem could be used to axiomatise EVM as a theory of u^2-value. I shall conclude that the prospects for this look bleak.

2.1 The Problem

Before introducing any axiomatic theorems, it will be helpful to indicate briefly why I think we need such theorems in the first place. Unsurprisingly, the question about the role and relevance of axiomatic theorems in decision theory is itself disputed.[1] I shall in different ways come back to it throughout my thesis, and shall try to defend the importance of such theorems at various points. But very roughly, I think representation theorems can serve two important purposes. Firstly, they can help to clarify what EVM *means*; and secondly, they can help to vindicate that EVM is *true*.

Let me briefly elaborate on this. Explaining what EVM means is important because EVM features a number of concepts that have no use in ordinary language, and that I think are in need of explanation. One such concept is the quantitative notion of value. To understand EVM, it does not suffice to understand qualitative, or *ordinal* statements like

(A) according to axiology T_i, outcome x is *better* than outcome y.

Rather, we have to understand statements like

(B) according to axiology T_i, the value difference between the outcomes x and y is n times as great as the value difference between the outcomes z and t.

I shall call such statements *cardinal intratheoretic comparisons (of value)*. They state intratheoretic value difference ratios – the ratios between certain

[1] Cf. e.g. Meacham and Weisberg (2011, 644ff.) for a brief overview over different possible interpretations of representation theorems, and for important challenges to them (some of which are addressed in chapter 4).

value differences, according to one specific axiology. It depends on such value difference ratios whether one option has a higher expected value than another. However, I think that unless more is said we do not understand quantitative statements like (B). As an example, suppose you can either save the life of a human being, or of a non-human animal. And suppose you claim that according to your favourite axiology, saving the person is five, or ten times as good as saving the animal. Unless you give me some account of what you mean by that, I would not understand what you meant – only that you chose a slightly swaggering way of expressing, say, that according to your favourite axiology, saving the person is *considerably* better than saving the animal, and I would have thought the same if your numbers had been seven, or fifteen. This is not to say that you *cannot* give me an explanation of what you meant. For example, with this particular statement of yours, you may mean that saving the person would be equally good as saving *five* animals like the one we considered, and that may count as a proper explanation. What I claim is that you *have to* give me an explanation, or else I do not understand your original statement. So more generally, I think we have to give an account of cardinal intratheoretic comparisons to make clear what we mean by 'EVM'.

But cardinal intratheoretic comparisons are only one kind of statement that EVM presupposes, and that are in need of explanation. Another, and even more problematic kind are claims about intertheoretic value difference ratios – claims like

(C) the value difference between outcomes x and y, according to axiology T_i, is n times as great as the value difference between outcomes z

and t, according to axiology T_j.

I shall call such claims *cardinal intertheoretic comparisons (of value)*. Whether statements like (C) can be meaningful and true is part of the problem of intertheoretic comparisons, which I shall discuss in some depth in chapter 3. What is relevant for now is that, again, I think we do not understand statements like (C) unless we are given some account of them.

Finally, there is a third problematic notion that EVM presupposes. To understand EVM, we have to understand a quantitative notion of the probability of an axiology – statements like

(D) the probability of axiology T_i is p_i,

for some $p_i \in [0, 1]$. And again, such statements have no use in ordinary language and I think we have to provide an account of them to make clear what we mean by 'EVM'.

In all three cases, I think representation theorems can help us define and explain the relevant concepts. This is one of the purposes of these theorems. Only once we've answered these conceptual questions can we ask whether EVM is true. And this, I think, is the other purpose of these theorems. Represensentation theorems can help us argue that EVM is true.

All of these claims are controversial. But I shall argue for them in this and the following chapters. So let me now introduce a pertinent result.

2.2 State-Dependent Utility Theory

There are a number of different versions and formal frameworks of decision theory. I shall not compare these different versions in terms of how well they can be applied to axiological uncertainty. This would be an interesting, but extensive task, and it is not one of the aims of this thesis. Instead, I shall simply use one kind of decision theory that serves my purposes. The framework I employ is so-called state-dependent utility theory. Section 2.2.1 briefly explains what state-dependent utility theory is and why it is suitable for the theory of axiological uncertainty. Section 2.2.2 then reproduces its most basic result.

2.2.1 History and Motivation

State-dependent utility theory is best illustrated by how it departed from Leonard Savage's *Foundations of Statistics* (1954). On Savage's framework, there is a set of *states of nature* (or just 'states'), and a set of *outcomes* (or 'consequences'). An *act* is a mapping from states to outcomes: to each state, it associates an outcome, which – as Savage interprets it – that act brings about if the respective state is the actual state of the world. An agent is uncertain about which state is actual, and thus about the ultimate outcomes of her acts. Savage provided necessary and sufficient conditions for a preference relation over acts to be representable by a utility function over outcomes and a probability distribution over states.

To clarify, suppose you must choose between going on a hike and reading at home on an afternoon, but you are uncertain whether it will be rainy or

sunny. We can then distinguish two states – the state of rain and that of sunshine – and four possible outcomes – your reading or hiking, both either while it is rainy or while it is sunny. Your choice is between an act which leads either to 'hiking in the rain' or 'hiking in the sun', and one that leads either to 'reading while it rains' or 'reading while the sun shines'.

However, note that having distinguished states and outcomes in this way, we can now define an 'act' that leads to the outcome 'hiking in the rain' under *both* states of nature, whether it is rainy *or not*. Indeed, in order for your choices to be representable by a utility function, on Savage's framework, you must have a preference between other acts and this one – a hike that is certainly rain-swept, even if the weather is sunny. More generally, Savage's result required that you have preferences about acts leading to *any* arbitrary outcome in *any* state (from the sets of states and outcomes under consideration).

I am not suggesting that this is a severe problem for Savage's framework.[2] What is important in our context is that it led decision theorists to revise one of his main assumptions – viz., that utility is a function of *outcomes* only. Note that the apparently inconsistent combinations of outcomes and states could be avoided by individuating outcomes more coarsely – in our example, distinguishing not four outcomes, but only 'hiking' and 'reading'. *These* outcomes can arise under any state of the weather. However, such an individuation seems impossible for Savage. Savage takes utility to be a function of outcomes alone, but the pleasantness of your act does not

[2]Cf. e.g. Joyce (1999, 107ff.) for a discussion of this problem. For Savage's own unapologetic stance on it, cf. Drèze (1987, 78).

only depend on whether you're 'hiking' or 'reading'. It also depends on the weather. So what examples like this motivated is the theory of *state-dependent utility*: a formal framework in which the utility of an outcome can depend on the state in which it comes about. In such a framework, the role of Savage's utility functions is played by *state-dependent utility functions*: two-place mappings $u(\cdot, \cdot)$, which assign a utility to every *state-outcome pair*. Given our more coarse-grained individuation of outcomes, such a theory may distinguish u(rain, hiking) from u(sunshine, hiking), as seems plausible. So state-dependent utility theory promises to give a more natural interpretation of your stance, without requiring you to contemplate a rainy hike in a sunny state of nature.

This development in decision-theory is interesting for the theory of normative uncertainty. On a natural application of Savage's framework to axiological uncertainty, the 'states of nature' among which we are uncertain are axiologies, and 'utility' corresponds to (u-)value. But different axiologies generally assign different values to empirical outcomes – i.e., to non-normative states of affairs like a pleasurable hike, or the suffering of an animal, or whatever. So the 'utility' of such an outcome generally depends on the 'state' in which it arises; its value depends on which axiology is true. Accordingly, state-dependent utility theory promises to be a useful tool for the theory of axiological uncertainty. On a natural interpretation, the problem of axiological uncertainty *is* one of state-dependent utility.

Again, this does not mean that state-dependent utility theory is the only formal framework allowing for a decision-theoretic approach to our problem. But the structural equivalence perhaps makes it the obvious starting point.

So it is the framework I shall use in this thesis.

2.2.2 Karni and Schmeidler's Theorem

Let me now introduce the specific result I shall employ in this chapter. The result is due to Edi Karni and David Schmeidler (1980), based on a theorem in Fishburn (1970), and is a fairly simple extension of the standard von Neumann-Morgenstern (1944) result.

To state it, let X be a finite set of outcomes and S a finite set of states. Let Y be the set of state-outcome pairs, $Y = \{(s, x) \mid s \in S, x \in X\}$. Define the set $\mathcal{A}$ as $\mathcal{A} = \{\boldsymbol{a} : Y \to \mathbb{R}_+ \mid \sum_{(s,x)\in Y} \boldsymbol{a}(s, x) = 1\}$. Karni and Schmeidler interpret members of $\mathcal{A}$ as acts of a specific kind; so an act in $\mathcal{A}$ assigns a probability to every state-outcome pair. Their original understanding of this does not need to concern us now;[3] what is important for now is the mathematical fact they present. We shall later see that this fact has a very natural interpretation in the context of axiological uncertainty. For some $p \in [0, 1]$, define the act $p\boldsymbol{a}+(1-p)\boldsymbol{b}$ in $\mathcal{A}$ that leads to $\boldsymbol{a}$ with probability p, and to $\boldsymbol{b}$ with probability $(1-p)$, as $(p\boldsymbol{a}+(1-p)\boldsymbol{b})(s, x) = p\boldsymbol{a}(s, x)+(1-p)\boldsymbol{b}(s, x)$ for all (s, x) in Y. Finally, let $\succeq$ be a reflexive binary relation on $\mathcal{A}$, and let its irreflexive part $\succ$ be defined as usual: $\boldsymbol{a} \succ \boldsymbol{b}$ if $\boldsymbol{a} \succeq \boldsymbol{b}$ but not $\boldsymbol{b} \succeq \boldsymbol{a}$.

We can then define the following four conditions as applying to $\succeq$:

Transitivity$_{\mathcal{A}}$: if $\boldsymbol{a} \succeq \boldsymbol{b}$ and $\boldsymbol{b} \succeq \boldsymbol{c}$, then $\boldsymbol{a} \succeq \boldsymbol{c}$;

[3]They ultimately use the framework of Anscombe and Aumann (1963), who interpret an act as a sequence of two lotteries – a lottery among the states, and then a lottery among outcomes within each state. The result I present in this chapter is a subsidiary result to their main theorem, which I introduce in chapter 4, and which remains closer to the framework of Anscombe and Aumann (distinguishing a subjective 'horse lottery' among states from an objective 'roulette lottery' among outcomes within states).

Completeness$_{\mathcal{A}}$: for any $\boldsymbol{a}$ and $\boldsymbol{b} \in \mathcal{A}$, $\boldsymbol{a} \succeq \boldsymbol{b}$ or $\boldsymbol{b} \succeq \boldsymbol{a}$;

Independence$_{\mathcal{A}}$: if $\boldsymbol{a} \succ \boldsymbol{b}$ and $p \in]0,1[$ then $p\boldsymbol{a}+(1-p)\boldsymbol{c} \succ p\boldsymbol{b}+(1-p)\boldsymbol{c}$ for any $\boldsymbol{c} \in \mathcal{A}$;

Continuity$_{\mathcal{A}}$: if $\boldsymbol{a} \succ \boldsymbol{b}$ and $\boldsymbol{b} \succ \boldsymbol{c}$ then there exist p and $q \in]0,1[$, s.t. $p\boldsymbol{a} + (1-p)\boldsymbol{c} \succ \boldsymbol{b}$ and $\boldsymbol{b} \succ q\boldsymbol{a} + (1-q)\boldsymbol{c}$.

These conditions are the standard axioms of von Neumann-Morgenstern expected utility theory. So let me use the term '*vNM-conformable*' for all binary relations on $\mathcal{A}$ that satisfy these four conditions. Karni and Schmeidler (1980, 8) then state[4]

Karni and Schmeidler's Theorem: If a reflexive binary relation $\succeq$ on $\mathcal{A}$ is vNM-conformable, then there is a function $u : Y \to \mathbb{R}$, unique up to positive affine transformation,[5] such that for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{A}$,

$$\boldsymbol{a} \succeq \boldsymbol{b} \quad \text{iff} \quad \sum_{(s,x)\in Y} \boldsymbol{a}(s,x)u(s,x) \geq \sum_{(s,x)\in Y} \boldsymbol{b}(s,x)u(s,x). \tag{2.1}$$

To apply this theorem to the context of axiological uncertainty, we have to interpret the relevant binary relation ($\succeq$) as the u-value relation, the states (in S) as axiologies, and the state-dependent utility functions ($u(s,\cdot)$) as the value-functions of our axiologies. Let me turn to this now.

[4]The result is restated in Karni (1985, 14). Basically, the theorem is Fishburn's (1970) Theorem 8.2, applied to state-outcome pairs rather than outcomes.

[5]A function $g : Y \to \mathbb{R}$ is a *positive affine transformation* of another function $f : Y \to \mathbb{R}$ if there are $s, t \in \mathbb{R}, s > 0$ such that $g(z) = sf(z) + t$, for all $z \in Y$.

2.3 Applying State-Dependent Utility Theory

In order to apply Karni and Schmeidler's Theorem to our context, I first have to put my core question in a more formal terminology. I shall do that in section 2.3.1. Section 2.3.2 then introduces some additional assumptions and conditions, and applies Karni and Schmeidler's Theorem in a slightly extended form to axiologies. Section 2.3.3 discusses the significance of the resulting theorem.

2.3.1 The Terminological Framework

My focus is on axiological uncertainty, so – except for section 5.2 – I shall ignore non-axiological normative uncertainty, such as uncertainty between various deontological theories. I will assume that all uncertainty is either axiological or non-normative.

To represent *non*-normative uncertainty, let X again be a finite set of outcomes – where an outcome is a non-normative state of affairs whose description does not involve any further probabilities or uncertainty. An *option* is a prospect over X. Each of these prospects leads to particular outcomes in X with particular given probabilities – thus reflecting non-normative uncertainty (or risk[6]). I shall denote these prospects by lower-case letters $a, b, c...$, and the set of these options by $\mathcal{O}$. Formally, an option a in $\mathcal{O}$ is thus a function from X to $\mathbb{R}$, which assigns a probability $a(x)$ to each outcome x in

[6]In line with much philosophical literature, I use 'uncertainty' for what economists and formal decision theorists call 'risk'. My formal definitions should make clear what I mean.

X. Hence, $\mathcal{O} = \{a : X \to \mathbb{R}_+ | \sum_{x \in X} a(x) = 1\}$. As an example, an option in $\mathcal{O}$ may assign probability 0.5 to benefiting an animal, and probability 0.5 to benefiting a human being. Moreover, for any a and b in $\mathcal{O}$, and any $p \in [0, 1]$, define $pa + (1 - p)b$ in $\mathcal{O}$ by $(pa + (1 - p)b)(x) = pa(x) + (1 - p)b(x)$ for all x in X. That is, '$pa + (1 - p)b$' denotes the option that leads to a with probability p, and to b with probability $(1 - p)$. For some outcome x in X, let a_x be the prospect that certainly leads to x: $a_x(x) = 1$.

I shall assume, at least for now, that an *axiology* T_i is a transitive binary relation on $\mathcal{O}$, whose reflexive part is the 'at least as good as' relation. – I say 'for now', because in chapter 3 I shall consider whether axiologies are individuated more finely than by the ordering they imply. But this need not concern us now. – For two options a and b, I shall write '$a \succeq_i b$' to denote that a is at least as good as b according to T_i. The relations of strict betterness and equality in goodness are induced as usual: a is better than b on T_i ('$a \succ_i b$') if $a \succeq_i b$ but not $b \succeq_i a$, and a is equally as good as b on T_i ('$a \sim_i b$') if $a \succeq_i b$ and $b \succeq_i a$. I shall say that an axiology is *non-uniform* if there are a and b in $\mathcal{O}$ such that $a \succ_i b$. For simplicity, I shall assume throughout the thesis that the set $\mathcal{T} = \{T_1, T_2, ...T_n\}$ of axiologies under consideration is finite, and I shall denote its index set by $I = \{1, 2, ...n\}$.

If we are both non-normatively *and* axiologically uncertain, we face more complex prospects: prospects that lead to certain theory-outcome *pairs* with particular probabilities – now corresponding to Karni and Schmeidler's 'acts', which lead to state-outcome pairs. In our context, such prospects represent acts that have certain probabilities of arising under different axiologies, and certain probabilities of yielding different outcomes. For example, such an

option may assign equal probability to a utilitarian and a speciesist axiology, and (regardless of the axiology) equal probability to resulting in a benefit for a human being and a benefit for an animal; it will thus assign probability 0.25 to all four theory-outcome pairs. For simplicity, I shall also call these more complex prospects *options*. The resulting ambiguity of the term 'option' will not produce confusion: I will always make clear which type of options I am referring to. To distinguish them formally, I shall use bold letters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$... to refer to such more complex options, write $\boldsymbol{a}(i, x)$ for the probability with which option $\boldsymbol{a}$ leads to outcome x while axiology T_i is true, and denote the set of all such options by $\mathcal{Q}$. Defining the set of theory-outcome pairs by $Z = \{(i, x) \mid i \in I, x \in X\}$, we thus have $\mathcal{Q} = \{\boldsymbol{a} : Z \to \mathbb{R}_+ \mid \sum_{(i,x)\in Z} \boldsymbol{a}(i, x) = 1\}$. The probability assigned to an axiology T_i under some option $\boldsymbol{a}$ is thus $\sum_{x\in X} \boldsymbol{a}(i, x)$. And we can again define $p\boldsymbol{a} + (1 - p)\boldsymbol{b}$ in $\mathcal{Q}$ as the option that leads to $\boldsymbol{a}$ with probability p, and to $\boldsymbol{b}$ with probability $(1 - p)$, hence $(p\boldsymbol{a} + (1 - p)\boldsymbol{b})(i, x) = p\boldsymbol{a}(i, x) + (1 - p)\boldsymbol{b}(i, x)$ for all (i, x) in Z.

Note that this means that I am *presupposing* a quantitative notion of probabilities. My very definition of options in $\mathcal{Q}$ presupposes that we understand what it means that an axiology has a particular probability of being true (as well as that an outcome has a particular probability of arising). I assume this only for simplicity, and to focus on some other problems for now. I shall provide an explication for this quantitative notion of probabilities in chapter 4.

In line with this definition of options, I assume that the u-value relation is a transitive binary relation on $\mathcal{Q}$. I shall denote its reflexive part – the 'at least as u-good' relation – by '$\succeq_U$'. The strict u-betterness relation ($\succ_U$)

and the 'equally as u-good' relation ($\sim_U$) are induced equivalently to the respective value relations.

Even though options like $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ (in $\mathcal{Q}$) are formally distinct from options like a, b and c (in $\mathcal{O}$), and I said that axiologies order the latter, I shall sometimes say that $\boldsymbol{a}$ *is at least as good as* $\boldsymbol{b}$ according to an axiology T_i. By this I shall mean, intuitively, that the prospect represented by $\boldsymbol{a}$, given T_i, is at least as good according to T_i as the prospect represented by $\boldsymbol{b}$, given T_i. This will only make sense if the probability assigned to T_i under $\boldsymbol{a}$ and $\boldsymbol{b}$ is strictly positive. So to define this concept formally, let $\mathcal{Q}^i \subset \mathcal{Q}$ be the set of options in which T_i has a strictly positive probability, i.e. $\mathcal{Q}^i = \{\boldsymbol{a} \in \mathcal{Q} \mid \sum_{x \in X} \boldsymbol{a}(i,x) > 0\}$. Define for each axiology T_i a function $H_i : \mathcal{Q}^i \to \mathcal{O}$; $\boldsymbol{a} \mapsto H_i(\boldsymbol{a})$, such that

$$H_i(\boldsymbol{a})(x) = \boldsymbol{a}(i,x) / \sum_{y \in X} \boldsymbol{a}(i,y). \tag{2.2}$$

The mapping H_i thus turns an option $\boldsymbol{a}$ into the prospect that $\boldsymbol{a}$ represents, given T_i (if there is such a prospect). So for some $\boldsymbol{a}$ and $\boldsymbol{b}$, with $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^i$, I shall say that $\boldsymbol{a}$ is at least as good as $\boldsymbol{b}$ according to T_i if $H_i(\boldsymbol{a}) \succeq_i H_i(\boldsymbol{b})$ – and similarly for 'better' and 'equally good'.

Since this framework will underlie many of my formal results, let me make two brief remarks about it at this point. First, note that my definition of the set of options $\mathcal{Q}$ has an implication similar to the one we encountered with Savage. For *any* probability distribution over theory-outcome pairs, there is an 'option' in $\mathcal{Q}$ that is defined by it. In particular, $\mathcal{Q}$ includes options that lead to different actual outcomes, or different probability distributions over

outcomes, depending on which axiology is true. That is, for some $\boldsymbol{a}$ in $\mathcal{Q}$, with $\boldsymbol{a}$ in $\mathcal{Q}^i$ and $\mathcal{Q}^j$ for some T_i and T_j, $H_i(\boldsymbol{a}) \neq H_j(\boldsymbol{a})$. These options will be very unnatural. For example, there is no natural option which leads to benefiting an animal if a utilitarian axiology is true, and benefiting a human being if a speciesist axiology is true. Such an option will be definable in the framework I adopt. And indeed, my results will require that the u-value relation ranges over *all* options in $\mathcal{Q}$ – such unnatural options included.

However, unlike (arguably) the ones in Savage's framework, such options are not conceptually or metaphysically impossible. Consider again the option I just mentioned. Suppose a demon constructed a button for you. He tells you that if you push it, then, if the speciesist view is true, a human being will be benefited, and if the utilitarian view is true, an animal will be benefited. On the adequate decision-theoretic representation of pushing the button, the probability with which an animal or a person will be benefited should depend on the true axiology. This is clearly somewhat unnatural. But since this story seems (conceptually and metaphysically) possible, I do not think this is a fatal problem for our framework. However unnatural and extraordinary such options are, it does not seem implausible that u-value relations hold even among them.

Secondly, there is another potentially unnatural feature of the framework. It allows the u-value relation to range over options that involve different probability distributions over axiologies. That is, for some options $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$, and some theory T_i, $\sum_{x \in X} \boldsymbol{a}(i, x) \neq \sum_{x \in X} \boldsymbol{b}(i, x)$. Indeed, again, my results will *require* that the u-value relation ranges over such options. And although all these options may be perfectly natural considered by themselves, we can-

not face *choices* between them. For example, we cannot face a choice between benefiting an animal if utilitarianism is true, and doing so if a speciesist view is true.

But again, it does not seem implausible that u-value relations hold even among such options. If benefiting an animal is better on the utilitarian view than on the speciesist view, then it seems adequate to claim that it has more u-value if the former is certain than if the latter is. So again, this perhaps unnatural, or unpractical aspect of the present framework does not seem to be a problem. In fact, I shall later argue that these somewhat unnatural aspects of my framework may actually be an advantage, since they make it easier to think about certain questions. But I can only show this after much further argument. It will have to wait until page 192.

2.3.2 The Expected Value Theorem

With the formal terminology I have now introduced, Karni and Schmeidler's Theorem straightforwardly applies to the relation $\succeq_U$ and the set of options $\mathcal{Q}$ that represent both our axiological and non-normative uncertainty. Nonetheless, we cannot yet use it to derive EVM. What Karni and Schmeidler's Theorem now shows is this: if $\succeq_U$ is vNM-conformable, there is a function $u : Z \to \mathbb{R}$, unique up to positive affine transformation, such that for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$,

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \boldsymbol{a}(i,x)u(i,x) \geq \sum_{(i,x)\in Z} \boldsymbol{b}(i,x)u(i,x). \tag{2.3}$$

However, (2.3) is not EVM. For all we know, the function u is a purely mathematical function with no non-mathematical significance. To turn (2.3) into EVM, we have to interpret the theory-dependent utility functions $u(i, \cdot)$ as the *value-functions* of our axiologies, and thus give them an extra-mathematical significance. I will do that by introducing three additional assumptions and an extra condition. I shall introduce these claims rather succinctly, and then state the theorem; section 2.4 will elaborate more on these claims.

The Form of Axiologies

In defining axiologies, on page 43, I assumed that each axiology is as a matter of definition transitive. My first additional assumption is that all axiologies that I consider satisfy the other von Neumann-Morgenstern axioms as well. That is, I shall assume that all axiologies under consideration satisfy the following conditions:

Transitivity$_{\mathcal{O}}$: if $a \succeq b$ and $b \succeq c$, then $a \succeq c$;

Completeness$_{\mathcal{O}}$: for any a and $b \in \mathcal{O}$, $a \succeq b$ or $b \succeq a$;

Independence$_{\mathcal{O}}$: if $a \succ b$ and $p \in]0, 1[$ then $pa + (1-p)c \succ pb + (1-p)c$ for any $c \in \mathcal{O}$;

Continuity$_{\mathcal{O}}$: if $a \succ b$ and $b \succ c$ then there exist p and $q \in]0, 1[$, s.t. $pa + (1-p)c \succ b$ and $b \succ qa + (1-q)c$.

Since no confusion will arise, I will often leave out the subscript '$\mathcal{O}$' when referring to these conditions; and as with $\mathcal{Q}$, I shall use the term '*vNM-conformable*' for all binary relations on $\mathcal{O}$ that satisfy these conditions. So

I assume – at least for now – that all axiologies under consideration are vNM-conformable.

This is a substantial restriction. Not all possible axiologies, nor even all the plausible ones, are vNM-conformable. I shall give some examples of non-vNM-conformable axiologies in section 2.4.1. But it is worth emphasising now that this first assumption substantially restricts the range of my result. I will actually not explore axiological uncertainty generally. My results – at least until chapter 5 – concern uncertainty about vNM-conformable axiologies only.

Intratheoretic Comparisons

My second assumption concerns the meaning of cardinal intratheoretic comparisons. As I mentioned, we do not use such statements in ordinary language. And I do not think that we have a pre-theoretic concept of value that has a cardinally significant meaning. That is, I think that we cannot *find out* what cardinal intra- and intertheoretic comparisons of value *really* mean. Instead, when we come to precise philosophical theorising and to formulate a theory like EVM that presupposes cardinal comparisons, we have to choose an *explication* of our pre-theoretic concept of value, and decide what we shall mean by them. This claim may be controversial. I shall go some way towards defending the need of an explication in section 2.4.3, and again (in similar contexts) in chapters 3 and 4. But let me first introduce the explication that I shall use.

To state this explication for intratheoretic comparisons, suppose that for

some utility function $u : X \to \mathbb{R}$ on outcomes and some axiology T_i, and for all options a and b in $\mathcal{O}$,

$$a \succeq_i b \quad \text{iff} \quad \sum_{x \in X} a(x)u(x) \geq \sum_{x \in X} b(x)u(x). \tag{2.4}$$

I shall then say that *the expectation of u represents $\succeq_i$ ordinally* – or for convenience, simply that *u represents $\succeq_i$ ordinally*. Note that if there is a utility function that represents $\succeq_i$ ordinally, then there are infinitely many such utility functions. If u is one of them, any positive affine transformation of u is another such utility function; and any function that represents $\succeq_i$ ordinally will be a positive affine transformation of u. Now suppose that for some utility function u on outcomes, and some axiology T_i, the cardinal intratheoretic comparisons between all outcomes, according to T_i, are the same as the ratios among the utility differences between these outcomes; that is, for all x, y, z, t in X and $n \in \mathbb{R}$, the value difference between outcomes x and y is n times as great as the value difference between z and t, according to T_i, if and only if $(u(x) - u(y))/(u(z) - u(t)) = n$. I shall then say that u *represents T_i cardinally*.

According to my explication, if a utility function u represents an axiology ordinally, then that utility function represents it cardinally. So in that case, the ratios among the value differences between outcomes, according to this axiology, are the same as the ratios among the utility differences between these outcomes. I shall call this the *decision-theoretic explication* of intratheoretic comparisons.[7]

[7]Cf. Broome (2004, 89ff.) for a similar assumption about 'personal goodness', and a helpful discussion of the nature of this assumption. Other illuminating discussions of

If a utility function represents an axiology T_i cardinally, we can represent that axiology with a value-function G_i that determines the goodness of any outcome, according to T_i, quantitatively. For our purposes, we can simply pick one among the family of utility functions that represent T_i cardinally, and take it to be T_i's value-function G_i. The reason is that, for our purposes, the absolute heights of value levels or sizes of value differences do not matter. All that matters are the ratios among the value differences, and these are the same for all functions that represent T_i cardinally. So out of that family, we can pick any function we like. Consider temperature as an analogy. Temperature can be measured equally well in Fahrenheit or in Celsius; and it could be measured by any other positive affine transformation of these scales. In principle, we could have picked any of these scales to measure temperature.[8] We can do the same for goodness. If a utility function u represents an axiology T_i cardinally, we can simply pick one of the functions that represent T_i, and take that to be its value-function G_i. For example, for any x and y, where y has a greater utility than x, we could pick the utility function on which $u(x) = 0$ and $u(y) = 1$, and suppose that $G_i = u$. This would amount to picking a particular *scale*, just as Fahrenheit or Celsius picked a scale for temperature. So in what follows, if an axiology can be represented ordinally by a utility function, I shall often represent it with a value-function. And since the choice of a scale would be arbitrary, I shall do that without specifying a specific scale.

Note what the decision-theoretic explication does. Suppose someone says

the issue of cardinalising goodness are provided in Broome (1991, 142ff., ch.10/11), and Greaves (forthcoming; ms).

[8]Cf. Broome (1991, 145f.) for this analogy.

that according to axiology T_i, the best option is the one that maximises the expected *square root* of value:

$$a \succeq_i b \quad \text{iff} \quad \sum_{x \in X} a(x)\sqrt{G_i(x)} \geq \sum_{x \in X} b(x)\sqrt{G_i(x)}. \tag{2.5}$$

The square root function is concave. So a given increase in the argument of that function will not always produce the same increase in the value of the function. It will produce a greater increase the lower the argument is. This will be manifested in our axiology. Consider the following three options, which all lead to two different outcomes with equal probability (0.5); and suppose the numbers in the the table refer to the values of these outcomes, according to T_i.

	a	b	c
0.5	1	1	4
0.5	0	1	0

Table 2.1

Both b and c differ from a in that one of their possible outcomes is better than in a. In b, the relevant outcome is 1 unit better than that in a; in c, the relevant outcome is 3 units better than that in a. However, since

$$0.5 \cdot \sqrt{1} + 0.5 \cdot \sqrt{1} = 0.5 \cdot \sqrt{4}, \tag{2.6}$$

our theory T_i implies that $b \sim_i c$. Hence according to this axiology, the relevant increases in the value of outcomes (by 1 and 3 respectively) *count* the same in determining the value of prospects, even though they *are* not

the same.

This is ruled out by the decision-theoretic explication. If a relation $\succeq$ on $\mathcal{O}$ is vNM-conformable, then utility is *defined* as that quantity of which $\succeq$ maximises the expectation, in the sense of (2.4). So by definition, increases in the utility of outcomes always *count* the same in determining the utility of the prospect. According to the decision-theoretic explication, we can use utility to represent goodness. So the explication assumes that, if two differences in value count the same in determining the goodness of prospects, then they necessarily are the same. The explication rules out views like the one I just sketched. I shall say it assumes that goodness is *expectational*, like utility.

We might say that according to the decision-theoretic explication, quantities of goodness acquire their cardinal significance in the context of weighing goods under uncertainty. As I emphasised on page 49, I treat this as an *explication*, not as a substantial assumption or faithful analysis of our pretheoretic concept of goodness. Accordingly, I do not think that this is the only possible explication of cardinal intratheoretic comparisons. For example, the context of weighing goods over *time* could also provide a cardinal concept of value. We could assume that if two differences in value coming at different times count the same in determining the goodness of the history of the world over time, then they necessarily *are* the same.[9] And there may be still other possibilities. When we come to precise theorising about goodness, we have to choose some explication. We can do that in different ways; the assumption that goodness is expectational is one way of doing so. It is the one we have to choose to apply Karni and Schmeidler's Theorem to our context. I shall

[9]Cf. Broome (2004, ch.15) for a cardinalization of personal goodness by time.

consider objections to it in section 2.4.3. But let me use it for now.

Intertheoretic Comparisons

I said that, given this explication, if a utility function represents an axiology T_i cardinally, we can simply pick one among the family of utility functions that represent it and take it to be T_i's value-function G_i. But actually, this is somewhat imprecise. That we can arbitrarily pick *any* of these functions is only true if we consider axiologies in isolation. It is not so when we consider multiple axiologies jointly. Suppose we assume that T_i's value-function is G_i. If we then *simultaneously* represent another axiology T_j by some value-function G_j, this implies claims about how values compare intertheoretically among these theories. For example, if $G_i(x) > G_j(x)$, then the value of x is greater according to T_i than according to T_j. At least, this is how I shall understand it.[10] If we consider axiologies jointly, then once we picked a *global scale* for representing axiologies – or once we picked a scale for some T_i – it will not be true that we can arbitrarily pick a scale for some other axiology T_j.

My third assumption concerns what it would mean, more precisely, for two axiologies to compare in a specific way. I have just assumed that, *if* an axiology can be represented ordinally by a utility function, then on that axiology, intratheoretic comparisons acquire their cardinal significance in the context of weighing goods under uncertainty. My third assumption will be, similarly, and very roughly, that *if* the u-value relation could be represented

[10]Of course, one might use the same notation (i.e., represent T_i and T_j simultaneously by G_i and G_j) but assume that the implied intertheoretic comparisons have no significance. I assume that they have.

ordinally by a utility function, then intertheoretic comparisons would acquire their cardinal significance in the context of weighing axiologies under axiological uncertainty. Let me state this more precisely.

For practical purposes, it is only intertheoretic *unit* comparisons that matter. What is relevant are the intertheoretic value difference ratios – the truth of statements like 'the value difference between x and y, according to axiology T_i, is n times as great as the value difference between z and t, according to axiology T_j'. *Level* comparisons do not matter. It is irrelevant whether statements like 'the value of x, according to axiology T_i, is greater than the value of y, according to axiology T_j' are true. The reason is that in practice we do not face choices between options with different underlying probability distributions over axiologies (as I said on page 47). And if the probability distribution over axiologies is the same for all options, then which of these options has the highest expected value does not depend on the heights of value levels. Formally, if

$$\sum_{x \in X} \boldsymbol{a}(i,x) = \sum_{x \in X} \boldsymbol{b}(i,x) \quad \forall i \in I, \tag{2.7}$$

then

$$\sum_{(i,x) \in Z} \boldsymbol{a}(i,x)[u(i,x) + t_i] \geq \sum_{(i,x) \in Z} \boldsymbol{b}(i,x)[u(i,x) + t_i] \;\Leftrightarrow$$
$$\sum_{(i,x) \in Z} \boldsymbol{a}(i,x)u(i,x) \geq \sum_{(i,x) \in Z} \boldsymbol{b}(i,x)u(i,x). \tag{2.8}$$

So for practical purposes, we would not have to explicate intertheoretic comparisons of value levels. However, since Karni and Schmeidler's Theorem

ranges over options with different underlying probability distributions over states, it implies a utility function that is unique up to positive affine transformation – not only unique up to the multiplication with a joint scalar and state-wise addition of a constant ($su(i,x)+t_i$). So we can actually use this theorem to explicate intertheoretic comparisons of value levels as well. That is, we can explicate not only what I called 'cardinal intertheoretic comparisons' on page 36. We can also explicate more complex statements of the form

(E) the difference between the value of x, according to T_i, and the value of y, according to T_j, is n times as great as the difference between the value of z, according to T_h, and the value of t, according to T_k.

I shall call statement like (E) *crosscutting cardinal intertheoretic comparisons (of value)*. If we can explicate such statements, we can explicate value level comparisons – as well as cardinal intra- or intertheoretic comparisons, which are limiting cases of (E).

How exactly can we do so? As I suggested, we can think of axiological uncertainty as a case of state- or theory-dependent utility – where the relevant utility depends on theory-outcome *pairs*, and the utility functions $u(\cdot,\cdot)$ are thus mappings from $I \times X$ to $\mathbb{R}$. So suppose that for some theory-dependent utility function u, the crosscutting cardinal intertheoretic comparisons, according to our axiologies, are the same as the respective intertheoretic utility difference ratios. I shall then say that u *jointly represents all axiologies cardinally*. If that is so, we can represent our theories jointly by u, or one of its positive affine transformations. We can assume, say, that $G_i(\cdot) = u(i,\cdot)$ for

all T_i. If a theory-dependent utility function satisfies (2.3) for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$, I shall say it *represents the u-value relation ordinally.*

Now the most straightforward explication of intertheoretic comparisons would be this: if a theory-dependent utility function u represents the u-value relation ordinally, then it jointly represents all axiologies cardinally. However, this explication would be unfortunate. For one thing, it would make our explication of cardinal intratheoretic comparisons redundant: we would be explicating *both* inter- and intratheoretic comparisons via the u-value relation. For another thing, it would rule out an unnecessarily great number of theories of axiological uncertainty by conceptual *fiat*. For example, consider Expected Disvalue Maximisation, or EDM – the view that one option is u-better than another if and only if it has the greater expected *disvalue*. On the present explication, that view would be ruled out by definition. The reason is that, if EDM is true, then there is a theory-dependent utility function u that represents the u-value relation ordinally. So our explication would imply, by definition, that we can pick a global scale such that $G_i(\cdot) = u(i, \cdot)$ for all T_i. Intuitively, however, EDM should imply that $-G_i(\cdot) = u(i, \cdot)$. And although it is true that every explication will rule out some views by such conceptual *fiat*, it is desirable to rule out as few theories as possible with one's definitions. So I shall make our explication *weaker* than the one we have now considered, and instead introduce another substantial condition into the theorem.

More precisely, I shall use the following definition. Consider again some theory-dependendent utility function $u(\cdot, \cdot)$, from $I \times X$ to $\mathbb{R}$. And suppose that this function is such that, for each axiology T_i, the utility function

$u(i, \cdot)$ represents that axiology cardinally. I shall then say that *u represents each axiology cardinally.* Stated precisely, my explication will be that if a theory-dependent utility function u represents the u-value relation ordinally *and* represents each axiology cardinally, it jointly represents all axiologies cardinally. That is, if there is such a utility function u, then the cross-cutting cardinal intertheoretic comparisons, according to our axiologies, are the same as the respective intertheoretic utility difference ratios according to u. We can thus assume, say, that $G_i(\cdot) = u(i, \cdot)$ for all T_i. I shall call *this* the *decision-theoretic explication* of interheoretic comparisons, and shall sometimes express it by saying that intertheoretic comparisons acquire their cardinal significance in the context of weighing axiologies under axiological uncertainty.

Again, I treat this as an explication, not as a substantial claim. I shall be relying on this explication in the entire thesis. But instead of discussing it in more detail at this point, let me now state the actual theorem it supports.

The Pareto Condition, and the Theorem

To apply Karni and Schmeidler's Theorem to the context of axiological uncertainty, we need a final substantial condition that guarantees that our utility function will represent each axiology cardinally. To that end, we can introduce a Pareto Condition concerning options with the same underlying probability distribution over axiologies, to the effect that if two such options are equally good on all theories with nonzero probability, they are equally u-good, and if one of them is at least as good as another on all theories with

nonzero probability and strictly better on some, then it is strictly u-better.

To state this condition formally, define for any probability distribution P on I the set $\mathcal{Q}^P \subset \mathcal{Q}$ of options in which P is the underlying probability distribution over axiologies, $\mathcal{Q}^P = \{\boldsymbol{a} \in \mathcal{Q} \mid \sum_{x \in X} \boldsymbol{a}(i,x) = P(i) \ \ \forall\, i \in I\}$. Employing the function H (from page 45), we can then state the

> **Pareto Condition**: For any probability distribution P on I, and for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^P$, if $H_i(\boldsymbol{a}) \sim_i H_i(\boldsymbol{b})$ for all T_i with $P(i) > 0$, then $\boldsymbol{a} \sim_U \boldsymbol{b}$; and if $H_i(\boldsymbol{a}) \succeq_i H_i(\boldsymbol{b})$ for all T_i with $P(i) > 0$ and $H_j(\boldsymbol{a}) \succ_j H_j(\boldsymbol{b})$ for some T_j with $P(j) > 0$, then $\boldsymbol{a} \succ_U \boldsymbol{b}$.

Given my previous assumptions, this condition together with the condition that $\succeq_U$ is vNM-conformable suffices to imply EVM. That is, given the decision-theoretic explications, and the assumption that all axiologies under consideration are vNM-conformable, the following theorem holds:

> **Expected Value Theorem**: If the u-value relation $\succeq_U$ is vNM-conformable and satisfies the Pareto Condition, then for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$,

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x) \in Z} \boldsymbol{a}(i,x) G_i(x) \geq \sum_{(i,x) \in Z} \boldsymbol{b}(i,x) G_i(x). \tag{2.9}$$

The sums on the right-hand side of this biconditional denote the expected value of $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively. So according to (2.9), an option is at least as u-good as another if and only if it has at least as great an expected value. This is EVM. So the Expected Value Theorem says that if u-value satisfies the von Neumann-Morgenstern axioms and the Pareto Condition, EVM must be true. This is the main result of this chapter.

This theorem is a very simple consequence of Karni and Schmeidler's Theorem. In the appendix to this chapter, I carry out the derivation in all detail. But the implication is straightforward: given Karni and Schmeidler's Theorem and a very simple application of the Pareto condition, the conditions of the Expected Value Theorem imply that there is a theory-dependent utility function that represents $\succeq_U$ and satisfies the criteria of our explications. Given our explications, this utility function can thus simply be interpreted as a set of value-functions, and instead of the biconditional (2.3) from page 47, we get EVM.

2.3.3 The Significance of the Theorem

The philosophical significance of the Expected Value Theorem depends, on the one hand, on whether its conditions are plausible. That is a question I shall start to consider in the next section, and that I can only answer more fully at the end of chapter 3. Ultimately, I shall conclude that the conditions of the theorem do not hold in general – i.e., as conditions for a theory of axiological uncertainty about *any* vNM-conformable set of axiologies. However, I do think that they hold if we consider only a restricted class of axiologies (the axiologies that are fully comparable and do not compare in a lexical way). So at least concerning uncertainty about this restricted class of axiologies, I think the Expected Value Theorem is indeed sound.

On the other hand, even a sound theorem is philosophically significant only if it establishes an interesting proposition or rules out some interesting rivals to it. And since it will take a lot of space to examine all the conditions,

it may be helpful to highlight at this juncture why the Expected Value Theorem would be interesting if it were sound – or rather, why it is interesting with regards to the restricted class of axiologies within which I think it *is* sound. So suppose for now that the conditions of the theorem are true, and that my additional assumptions are acceptable. And let me highlight briefly how the Expected Value Theorem can then serve the two purposes I assigned to representation theorems in section 2.1.

The Meaning of EVM

Firstly and most fundamentally, the theorem allows us to explain – or, as I understood it, to *explicate* – what EVM means. More precisely, it guarantees that there are conditions under which our explications can be used. These explications were conditionals ('*If* there is a utility function ...'); Karni and Schmeidler's Theorem shows that if $\succeq_U$ is vNM-conformable and satisfies the Pareto Condition, then the antecedents of these conditionals hold, and we can actually employ these explications to define a cardinal concept of value.

This is particularly significant with respect to cardinal intertheoretic comparisons. As far as I see, no one has provided an alternative explication for such statements. And indeed, many standard explications of value seem inadequate for them. Consider the explication via time. As I said on page 53, time is a possible context for explicating *intra*theoretic comparisons: we could assume that if two differences in value coming at different times count the same in determining the goodness of the history of the world over time, then they necessarily are the same. In the context of intertheoretic compar-

isons, however, such an explication seems to be impossible, or at best very unfortunate. We would have to imagine that different axiologies are true at different periods of time; and we would then consider how valuable the overall empirical-cum-axiological history of the world is, when different events occur during these periods of time. Such an explication may be feasible on a technical level. But since the truth of axiologies is (arguably) a timeless matter, each option in which different axiologies are true at different times is a metaphysical impossibility. So it is dubious whether there are any facts about the relevant overall value-relation. And even if there are, we do not seem to have a very good grasp of the (presumed) value relation that orders empirical-cum-axiological histories. Hence it would at best be unfortunate if we explained our theory in terms of that relation.

Similar considerations apply to many other candidate explications. Consider space. Space might again be a context for explicating *intra*theoretic comparisons: we could assume that if two differences in value arising at different *places* count the same in determining the goodness of the world, then they necessarily are the same. Within certain assumptions and restrictions, such an explication may be technically feasible, and sounds fairly natural. But again, the related explication of intertheoretic comparisons seems inadequate. We would have to imagine that different axiologies are true at different places in space; and we would then consider how valuable the world overall is, when different events occur at these places. And though that might again be technically feasible, it again involves a metaphysical impossibility, since the truth of axiologies is (arguably) a spatially universal matter. So it is again dubious whether there are facts about the relevant value-relation,

and if there are, we would not have a very good grasp of them.

In the next chapter, I shall consider whether it is possible to provide a cardinal concept of intertheoretic comparisons via a fitting attitude account of value. I think that this is the most promising alternative proposal, and I cannot rule it out entirely. But I shall raise various doubts about it, and remain very sceptical. So I think that the decision-theoretic explication is ultimately the best explication of cardinal intertheoretic comparisons that we have. This makes the Expected Value Theorem a very important result.

The Truth of EVM

A second and equally significant purpose of the Expected Value Theorem is that – if its conditions are plausible, or within the restricted context in which they are plausible – it allows us to *defend* EVM. So let me elaborate briefly on how the theorem rules out the most prominent alternatives to EVM.

Consider first the view that under axiological uncertainty, we ought to be risk-averse about value. To express this view, let $p_i^{\boldsymbol{a}}$ be the probability of T_i under $\boldsymbol{a}$, and let $V_i^{\boldsymbol{a}}$ be the value of $\boldsymbol{a}$ according to T_i; that is,

$$p_i^{\boldsymbol{a}} = \sum_{x \in X} \boldsymbol{a}(i, x), \tag{2.10}$$

and

$$V_i^{\boldsymbol{a}} = \begin{cases} \sum_{x \in X} \boldsymbol{a}(i, x) G_i(x) / p_i^{\boldsymbol{a}} & \text{if } \; p_i^{\boldsymbol{a}} > 0 \\ 0 & \text{if } \; p_i^{\boldsymbol{a}} = 0. \end{cases} \tag{2.11}$$

Say that *u-value is risk-averse* if there is an increasing strictly concave function ρ, such that for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$,

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{i \in I} p_i^{\boldsymbol{a}} \rho(V_i^{\boldsymbol{a}}) \geq \sum_{i \in I} p_i^{\boldsymbol{b}} \rho(V_i^{\boldsymbol{b}}). \tag{2.12}$$

The inputs of ρ are the values of the prospects that an option represents for given axiologies. Since the function is concave, increases in these values have more weight in determining u-value the lower these values are. This is a standard interpretation of risk-aversion.[11] And *prima facie*, this view is perfectly reasonable. If we want to rule it out, we need some argument.

The Expected Value Theorem may provide one. The view expressed in (2.12) is inconsistent with the assumption that $\succeq_U$ is vNM-conformable. In particular – and as we may have expected from standard decision theory – it is inconsistent with the Independence axiom for $\succeq_U$. To see this, let ρ be the square root function, $\rho(x) = \sqrt{x}$, and consider the following example. Suppose there are two outcomes x and y, and two theories T_1 and T_2 with $p_1 = p_2 = 0.5$, and $G_1(x) = G_2(y) = 0$, $G_1(y) = 30$ and $G_2(x) = 31$. Suppose option $\boldsymbol{a}$ leads to either x or y with a probability of 0.5 each, while $\boldsymbol{b}$ and $\boldsymbol{c}$ lead to x and y respectively (with certainty):

[11] Cf. e.g. Buchak (2013) for a slightly different one.

	a		***b***		***c***	
	T_1	T_2	T_1	T_2	T_1	T_2
	$p_1 = 0.5$	$p_2 = 0.5$	$p_1 = 0.5$	$p_2 = 0.5$	$p_1 = 0.5$	$p_2 = 0.5$
0.5	0	31	0	31	30	0
0.5	30	0	0	31	30	0

Table 2.2

According to (2.12), $\boldsymbol{b} \succ_U \boldsymbol{c}$. So Independence would require that $\boldsymbol{b}$ (or $^1/_2\,\boldsymbol{b} + {}^1/_2\,\boldsymbol{b}$) is u-better than $^1/_2\,\boldsymbol{c} + {}^1/_2\,\boldsymbol{b}$ – which is equivalent to $\boldsymbol{a}$. However, since

$$3.9 \approx 0.5 \cdot \sqrt{0.5 \cdot 30} + 0.5 \cdot \sqrt{0.5 \cdot 31} > 0.5 \cdot \sqrt{31} \approx 2.8, \tag{2.13}$$

(2.12) implies that $^1/_2\,\boldsymbol{c} + {}^1/_2\,\boldsymbol{b} \succeq_U \boldsymbol{b}$. Similar examples could be given for any other increasing strictly concave ρ. If the conditions of the Expected Value Theorem hold, then the utility functions that enter the u-value relation are those that represent our axiologies ordinally. If we assume that goodness is expectational at the level of first-order value, the axioms – and Independence in particular – guarantee that the utility functions that determine u-value represent our axiologies cardinally. In this sense, Independence may provide an argument for risk-neutrality.

Consider next the view that under axiological uncertainty, you simply ought to evaluate options in accordance with the theory you find most plausible: an option is u-better than another if and only if it is better according to the axiology with the highest probability (if there is such an axiology; if more than one axiology has maximal probability, there is some rule for break-

ing ties). This view is generally called *My Favourite Theory.* Some people have endorsed it with regards to moral uncertainty generally.[12] It blatantly contradicts the Pareto Condition. To see this, suppose the probabilities of T_1 and T_2 are 0.6 and 0.4 respectively, so that T_1 determines $\succeq_U$. And suppose two options $\boldsymbol{d}$ and $\boldsymbol{e}$ are equally good according to T_1, but $\boldsymbol{d}$ is strictly better according to T_2. Then, according to My Favourite Theory, $\boldsymbol{d}$ is equally as u-good as $\boldsymbol{e}$, which violates the Pareto Condition. I shall argue presently that this condition is very plausible. So I take it that the Pareto Condition provides a strong – indeed presumably the strongest – argument against My Favourite Theory.

More interestingly perhaps, the Expected Value Theorem also shows that there is no straightforward way for My Favourite Theory to remedy this flaw (if it is a flaw). In light of the above example, one may be tempted to say something like this: under axiological uncertainty, you ought to evaluate options in accordance with the most plausible theory, except if two options are equally good on that theory, and one is at least as good as the other on all axiologies with nonzero probability and strictly better on some, in which case the former is u-better than the latter. Views along these lines have also been suggested.[13] However, the Expected Value Theorem shows that if $\succeq_U$ is vNM-conformable the Pareto Condition implies that the u-value of

[12] cf. Gracely (1996).

[13] With regards to moral uncertainty generally, Gustafsson and Torpman (2014, 169ff.) are at least steering towards such a position. They revise My Favourite Theory so as to make it compatible with a Pareto (or 'Dominance') condition. Unfortunately, they do not discuss the von Neumann-Morgenstern axioms; and they explore uncertainty about theories that 'require' or 'permit' certain options, which may involve disanalogies to our case. But at least within the theory of axiological uncertainty, making My Favourite Theory satisfy the Pareto Condition forces one to deny that $\succeq_U$ is vNM-conformable.

an option is the weighted sum of the values it is assigned by the axiologies. There is simply no consistent intermediate position, on which u-value is vNM-conformable and Paretian, and yet does not reduce to a weighted sum of the axiologies, thus retaining the spirit of My Favourite Theory. If $\succeq_U$ is vNM-conformable, the project of making My Favourite Theory satisfy the Pareto Condition is doomed to fail. This is an important result.

Finally, when defending EVM in conversation, I have often heard the objection that it is not ecumenical enough as a view about normative uncertainty. If EVM is true, the u-best option is determined rather mathematically. But – the objection goes – people who are at heart sympathetic to virtue ethics, say, may deny that the appropriate choice is a matter of strict computation, even insofar as goodness is concerned. Rather, such people may endorse something like a meta-virtue ethical view about axiological uncertainty. On such a view, under axiological uncertainty, one ought to make a virtuous choice: circumspective but not overcautious, neither reckless nor overly anxious, and so on. Beyond that, there are no rigorous rules that determine the u-value relation. Similarly, some people have suggested meta-deontological principles for how to determine one's u-best option under axiological uncertainty, which potentially differ from EVM quite radically.[14] Such less formal views are hardly ever taken seriously or even mentioned in decision theory. But *prima facie*, they are just as plausible in the theory of u-value as they are in moral theory or the theory of ordinary value. If we

[14]E.g., Guerrero (2007, 94) endorses the following principle: 'Don't Know, Don't Destroy: If one knows that one doesn't know whether some entity has moral value, then it is morally blameworthy to destroy that entity, unless one believes that something of substantial moral significance compels one to do so.'

take seriously the challenge of defending a view of normative uncertainty, we need an argument against these views. And again, the Expected Value Theorem may provide one. It shows that these meta-virtue ethical or meta-deontological views must deny that u-value satisfies the conditions of the theorem. If u-value satisfies these conditions, EVM simply follows.

In sum, various views that have been suggested contradict one of the conditions of the Expected Value Theorem. If the theorem is sound – or within the restrictions in which it is sound – it is a very significant result.

2.4 Evaluating the Theorem

Let me now examine the conditions of the theorem in more detail. It is beyond the scope of this thesis to defend all the conditions of the Expected Value Theorem, or even one of them conclusively. So what I shall do instead is to compare their plausibility in the context of axiological uncertainty with their plausibility in other contexts – specifically, that of decision theory or first-order axiology, and (in the case of the Pareto Condition) that of social choice theory. Section 2.4.1 discusses the von Neumann-Morgenstern axioms, and section 2.4.2 examines the Pareto Condition. Section 2.4.3 will then elaborate again on the decision-theoretic explications.

2.4.1 The von Neumann-Morgenstern Axioms

Non-vNM-conformable Axiologies

Is it plausible that the u-value relation is vNM-conformable? It is worth emphasising that this is in fact very implausible unless we assume that all underlying axiologies are vNM-conformable. Note that, as I defined it, EVM is a theory of both axiological *and* non-normative uncertainty. So in the limiting case, $\succeq_U$ ranges over non-normative prospects in which some theory T_i is certain (i.e., options in the set $\{\boldsymbol{a} \in \mathcal{Q} \mid \sum_{x \in X} \boldsymbol{a}(i,x) = 1\}$). And if T_i itself is not vNM-conformable with respect to non-normative prospects, then $\succeq_U$ does arguably not satisfy the axioms with respect to these prospects if T_i is certain.[15] Or more generally, if T_i is one of the axiologies under consideration, then $\succeq_U$ does arguably not satisfy the von Neumann-Morgenstern axioms on $\mathcal{Q}$. And as I mentioned on page 49, not all reasonable axiologies are vNM-conformable.

It has even been disputed whether the betterness relation is transitive.[16] I cannot enter this debate here, but I think it is an analytic fact that 'better than' and 'at least as good as' are transitive. 'Better' is the comparative of 'good', and all comparatives are transitive.[17] At any rate, this is what I am assuming.

The most problematic axiom, I think, is Completeness. Many axiologies

[15] In fact, given a suitable Pareto condition, $\succeq_U$ then cannot satisfy these axioms. The Pareto Condition on page 59 is slightly too weak to guarantee this. It does not rule out that the u-value relation may represent a *sharpening* of an underlying axiology that is incomplete. But a slightly stronger condition (like the one on page 236) would do.

[16] Cf. most prominently Temkin (2012); also Rachels (1998).

[17] This view is defended in Broome (2004, ch.4); cf. e.g. Binmore and Voorhoeve (2003) and Voorhoeve (2013) for further defences of the transitivity of betterness.

are incomplete by allowing for incommensurability – i.e., the existence of two options a and b, such that it is false that a is at least as good as b, and false that b is at least as good as a.[18] The Expected Value Theorem cannot range over such axiologies. More precisely, the u-value relation is arguably not vNM-conformable (i.e., not complete) if it ranges over them. This is a very significant restriction. Many incomplete axiologies are plausible, and have been prominently defended by philosophers.[19] We should clearly not be certain that all such views are false. So this is one major reason for exploring axiomatisations without the Completeness condition. A representation theorem without the Completeness axiom would allow the theory of axiological uncertainty to cover incomplete axiologies. I shall turn to that in chapter 5.

Something similar is true for Continuity. There are axiologies that violate that condition. If a theory assigns infinite value to some outcome and finite values to others it will violate Continuity. If we do not put any restrictions on the set of outcomes X, then many plausible axiologies are of this sort. An example is standard total utilitarianism, according to which a world that contains beings whose lives are worth living, and only such beings, for an infinite stretch of time, will be infinitely valuable. But there are other ways in which axiologies violate Continuity. Some of them do so, for example, by allowing that some kinds of values dominate others lexically. Consider an axiology on which the value of wellbeing dominates the value of beauty lexically – in that whenever an option expectably leads to more wellbeing than another, that option is better, but *ceteris paribus* an option is better than

[18]Note that I shall use 'incommensurable' and 'incomparable' interchangeably.
[19]Cf. e.g. Raz (1986, ch.13).

another if it expectably leads to more beauty. This axiology does not satisfy Continuity.[20] Finally and most straightforwardly, an axiology may simply violate Continuity by implying a discontinuous behaviour of value with respect to probabilities. For example, consider an axiology on which, for some $p_0 \in]0,1[$, outcomes that arise with a probability of less than p_0 are irrelevant for the evaluation of prospects. This axiology will not be continuous. Again, the u-value relation is arguably not vNM-conformable if it ranges over such axiologies, and this is a practically significant restriction. My sense is that – in contrast to Completeness – most people find views that violate Continuity comparatively implausible (at least if discontinuities arise through lexical dominance between different values, or through probability thresholds). But presumably, we cannot be certain that such views are false. So in principle, even this presents a restriction of the Expected Value Theorem.

Similarly, there are axiologies that violate Independence. For example, some views are 'risk-averse' in a sense that is inconsistent with Independence.[21] And again, the u-value relation is arguably not vNM-conformable if it ranges over such axiologies. Here too, my sense is that most people find views that violate Independence comparatively implausible. But again, we can arguably not be certain that all such views are false.[22] So this too presents a restriction of the Expected Value Theorem.

In sum, unless we consider only vNM-conformable axiologies, there is a systematic reason why the von Neumann-Morgenstern axioms are less plausible, or indeed implausible, if we take into account *both* non-normative and

[20]Cf. e.g. Vallentyne (1993) for a view with this structure.
[21]Cf. e.g. the example in Allais (1953).
[22]Cf. e.g. Buchak (2013) for a defence of risk-sensitivity in decision theory.

normative uncertainty. *Even if* we think that according to the most plausible axiology, the standard betterness relation satisfies these axioms, we should arguably not be certain that this is so. And this uncertainty is enough to imply that u-value is not generally vNM-conformable.[23]

Fortunately, this is not so if we assume that all underlying axiologies are vNM-conformable. This is why I made this assumption in section 2.3.2. So suppose again that all axiologies are vNM-conformable, and let me examine whether the von Neumann-Morgenstern axioms are plausible within this limited scope.

The Problem of Intertheoretic Comparisons

Unfortunately, there is an additional reason for why the u-value relation may not be vNM-conformable – viz., the problem of intertheoretic comparisons. This problem affects the Completeness and Continuity axioms in particular.

To assume that $\succeq_U$ is complete is to assume that all axiologies are *fully comparable*, in the sense that they are all jointly representable by single value-functions. This is a strong assumption. Some radical sceptics about intertheoretic comparisons may hold that axiologies are not comparable at all, and thus that the u-value relation should be *radically* incomplete – that is, that an option $\boldsymbol{a}$ is at least as u-good as an option $\boldsymbol{b}$ only when $\boldsymbol{a}$ is at

[23]This has a very important implication for decision theory, understood as a theory of rational preferences. If u-value is not vNM-conformable, then our preferences arguably need not be vNM-conformable either, because we can justifiably care about u-value in our preferences. More generally, once we admit that people may justifiably be *uncertain* about the axioms of decision theory, and that they may take this normative uncertainty into account in their preferences, we should admit that their preferences need not satisfy these axioms. This may in fact be the most straightforward and convincing way to criticise the axioms of decision theory as general axioms of rational preference.

least as good as $\boldsymbol{b}$ on all axiologies with nonzero probability. Less radically, one might hold that some axiologies are at least not *fully* comparable. For example, suppose that T_i is a view on which only beauty has value, and T_j is a form of total utilitarianism. Even if one is not a radical sceptic, one might hold that the value of beauty, according to T_i, may not be fully comparable to the value of wellbeing on T_j, even if both T_i and T_j are themselves complete with respect to $\mathcal{O}$.

So to assess whether the Completeness condition is plausible, we have to address the large problem of intertheoretic comparisons. The next chapter is devoted to that problem, and I shall thus formulate my ultimate stance on Completeness only at the end of that chapter. To anticipate: I do not think that $\succeq_U$ is radically incomplete, but I do not think that it is fully complete either, even if all underlying axiologies are complete. On the view about intertheoretic comparisons that I find most plausible (a view I shall call 'Absolutism'), there are axiologies that are less than fully comparable. And I think some such less than fully comparable pairs of axiologies are very plausible. So this will present another major reason for exploring axiomatisations without the Completeness condition. A representation theorem without the Completeness axiom would allow the theory of axiological uncertainty to cover axiologies that are less than fully comparable.

The Continuity condition also raises the problem of intertheoretic comparisons. One might hold that there are axiologies that compare in a lexical way – axiologies T_i and T_j such that any positive value difference between options according to T_i is greater than any positive value difference between options according to T_j. This may be so, one might hold, even if T_i and T_j

themselves satisfy Continuity with respect to $\mathcal{O}$. And if two theories compare in a lexical way, then the u-value relation will not satisfy Continuity with respect to them.

Again, I shall come back to this question in the next chapter. On the view about intertheoretic comparisons that I find most plausible, I think there are indeed axiologies that compare in a lexical way. However, even if such theories are in principle *possible*, I think that they present a very extreme case, and are comparatively implausible. So unlike with the case of Completeness, I think it is not an all too significant restriction if we disregard such theories. But I have to defer this to the next chapter.

Conclusion

As far as I see, apart from the problem of non-vNM-conformable axiologies and the problem of intertheoretic comparisons, the question about whether the von Neumann-Morgenstern axioms are plausible constraints on $\succeq_U$ resembles the question about whether they are plausible constraints on first-order betterness. And I think that apart from these two caveats, they are indeed plausible.

Perhaps some people would question whether $\succeq_U$ is even transitive. But since I take u-betterness to be a form of betterness, I again think it is an analytic fact that $\succeq_U$ satisfies Transitivity.

As with ordinary value, u-value would fail to satisfy Continuity if it displayed a discontinuous behaviour with respect to probabilities – say, if there were probability thresholds $p_0 \in]0, 1[$ below which an axiology suddenly be-

came irrelevant for determining u-value. This could be true even if all axiologies were continuos. But such thresholds or discontinuities seem comparatively implausible, in the context of non-normative as in the context of normative uncertainty. Perhaps there are still other reasons why one might doubt Continuity.[24] I shall not enter that debate. I take Continuity to be a plausible axiom in the context of first-order value (at least for finite worlds), and so I think that *modulo* our two caveats it is a plausible condition about u-value too. I have to content myself with that.

U-value may also violate the Independence axiom even if all underlying axiologies are vNM-conformable. In particular, it is possible that (u-)value is risk-neutral with respect to non-normative uncertainty, but risk-sensitive with respect to axiological uncertainty, in ways that are inconsistent with Independence. But as far as I see, there is no reason why that should be the case – at least none that is somehow specific to the problem of axiological uncertainty. There is a long debate about the Independence axiom in the context of decision theory or first-order value.[25] Again, I shall not enter this debate. I take it to be a plausible axiom in these contexts, so I think it is a plausible axiom about u-value too, and I content myself with that.

So to conclude: as soon as we assign some nonzero probability to a non-vNM-conformable axiology, our u-value relation will not generally be vNM-conformable. If we assume that all our axiologies are vNM-conformable, the problem of intertheoretic comparisons still raises doubts about Completeness

[24]E.g., Temkin (2012, 245ff.) mentions that some outcomes may be 'good enough', or very significantly better than others, so that Continuity fails. I am not sure whether he would regard those as instances of the kind of 'lexical betterness' I mentioned.

[25]For recent contributions to this debate, cf. e.g. McClennen (2009), Temkin (2012, 237ff.), Buchak (2013, 157ff.).

and Continuity. But assuming that all axiologies are vNM-conformable, and *modulo* the problem of intertheoretic comparisons, I think it is plausible that $\succeq_U$ satisfies the von Neumann-Morgenstern axioms.

2.4.2 The Pareto Condition

Let me now turn to the Pareto Condition. Recall what this condition says: for any options with the same underlying probability distribution over axiologies, if two such options are equally good on all theories with nonzero probability, they are equally u-good, and if one of them is at least as good as another on all theories with nonzero probability and strictly better on some, then it is strictly u-better. Since options are prospects, this is a kind of *ex ante* Pareto requirement: it is concerned not with outcomes, but with what is expectably better.

One may have doubts about this condition too. Pareto conditions are rarely discussed in decision theory. However, there is also a close formal analogy between the theory of axiological uncertainty and social choice theory. Where we are concerned with an overall u-value ordering that depends on the value-orderings of axiologies, social choice theory (on one interpretation at least[26]) deals with a 'general betterness' ordering which depends on the 'individual betterness' orderings of people.[27] And *ex ante* Pareto conditions are controversial in social choice theory, with regards to people. That is, it is controversial whether, if one prospect is at least as good for all individuals as

[26]Cf. particularly Broome (1991).

[27]Indeed, note that the Expected Value Theorem is similar to John Harsanyi's (1955) famous 'utilitarian cardinal welfare theorem'. I have chosen to work with Karni and Schmeidler's rather than Harsanyi's theorem mainly because of the extensions that have been proved to the former (such as Karni and Schmeidler's Theorem 2, on page 185).

another and strictly better for some, it is *ceteris paribus* better (as far as general goodness is concerned). Given the close formal analogy between the two contexts, we should thus be careful to assume an *ex ante* Pareto condition in the theory of axiological uncertainty. So let me elaborate on this condition.

The Egalitarian Objection

The most pertinent argument against the *ex ante* Pareto condition for individual goodness is due to Marc Fleurbaey and Alex Voorhoeve (2013). Their argument shows that we have reason to reject this condition if egalitarianism is true (where egalitarianism is the view that equality in people's wellbeing is intrinsically valuable). To see their argument, suppose that we can either let both Blue and Red end up with 20 units of good (option f) or let them each face a lottery yielding either 10 or 31 with probability 0.5 (option g). Suppose further that, in this lottery, if Blue gets a benefit of 31, Red gets a benefit of 10, and vice versa:

	f		g	
	Blue	*Red*	*Blue*	*Red*
0.5	20	20	10	31
0.5	20	20	31	10

Table 2.3

Supposing that individual goodness is expectational, g is better than f for both Blue and Red. So if general goodness is *ex ante* Paretian with respect to personal goodness, g is better than f. However, according to (a suitable form of) egalitarianism, a state in which both Blue and Red have 20 units of

good is better than a state in which one has 31 and the other 10. Suppose this is true. Then if it were certain that g would lead to 31 benefits for Blue and 10 for Red, f would be better. Similarly, f would be better than g if it were certain that the latter would lead to 10 benefits for Blue and 31 for Red. So whatever the outcome of g will be: if it was certain, g would be worse than f. Therefore, Fleurbaey and Voorhoeve believe, g is then worse than f even if its outcome is uncertain. This is because we should 'decide as [we] would with full information' (2013, 113), they argue. More precisely, they invoke what they call the 'Principle of Full Information' (and what is basically Savage's (1954, 21ff.) 'Sure-Thing Principle'): 'When one knows that, in every state of the world with positive probability, one would rightly rank two alternatives in a particular way, then one should so rank them' (2013, 120).

For egalitarians like Fleurbaey and Voorhoeve, this may be a good reason to reject the *ex ante* Pareto condition concerning individual goodness. I shall not explore this. However, there does not seem to be a plausible parallel reasoning with regards to axiologies. Fleurbaey and Voorhoeve's argument crucially depends on the premise that f would be better than g if the outcome of g were certain. They can justify this claim as egalitarian. But it is not clear how an analogous premise could be justified in the context of axiological uncertainty. This premise would say that in the following choice $\boldsymbol{f}$ is u-better than $\boldsymbol{g}$:

$\boldsymbol{f}$		$\boldsymbol{g}$	
T_1	T_2	T_1	T_2
$p_1 = 0.5$	$p_2 = 0.5$	$p_1 = 0.5$	$p_2 = 0.5$
20	20	31	10

Table 2.4

But why might $\boldsymbol{f}$ be u-better? What immediately comes to mind is risk-aversion – the view I expressed as

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{i \in I} p_i^{\boldsymbol{a}} \rho(V_i^{\boldsymbol{a}}) \geq \sum_{i \in I} p_i^{\boldsymbol{b}} \rho(V_i^{\boldsymbol{b}}). \tag{2.14}$$

For some suitable increasing, strictly concave function ρ (e.g., $\rho(x) = \sqrt{x}$), this view indeed implies that $\boldsymbol{f} \succeq_U \boldsymbol{g}$. However, we cannot use it to argue against the Pareto Condition. As is easily verified, (2.14) actually *satisfies* this condition.

To argue against the Pareto Condition, we would thus have to endorse a different form of risk-aversion. Let me say that u-value is *ex post risk averse* if there is an increasing strictly concave function ρ, such that

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x) \in Z} \boldsymbol{a}(i,x) \rho\big(G_i(x)\big) \geq \sum_{(i,x) \in Z} \boldsymbol{b}(i,x) \rho\big(G_i(x)\big). \tag{2.15}$$

In (2.15), the function ρ is applied not to the *prospects* that our options represent on given axiologies, but instead to *outcomes*. This form of risk-aversion does not violate Independence in the sense in which the theory defined in (2.14) does. On the other hand, as is again easily verified, it does indeed violate the Pareto Condition, and imply that $\boldsymbol{f} \succeq_U \boldsymbol{g}$ (for some

relevant ρ). So it might be used in an argument against the Pareto Condition.

However, (2.15) is a very problematic view. Note what it implies if you are certain of one theory. Suppose you are certain of T_1, and face options $\boldsymbol{k}$ and $\boldsymbol{l}$: $\boldsymbol{k}$ leads to an outcome of value 20 (according to T_1), and $\boldsymbol{l}$ with equal probability either to an outcome of value 10, or to one of value 31 (according to T_1). Since you are certain of T_1, if goodness is expectational at the level of axiologies, you are certain that $\boldsymbol{l}$ is better than $\boldsymbol{k}$. Yet, (for the relevant ρ) (2.15) implies that the latter is u-better. This is very implausible. Surely, if we are certain that one axiology is true, we ought to rank options in accordance with it.

Perhaps there are other reasons, apart from these forms of risk-aversion, for believing that $\boldsymbol{f}$ is u-better than $\boldsymbol{g}$, and for denying the Pareto Condition on that basis. But I cannot think of any plausible candidate. So let me tentatively conclude that there is no parallel to Fleurbaey and Voorhoeve's objection in the context of axiological uncertainty. As far as their objection is concerned, the condition seems plausible with respect to u-value.

Pareto Improvements and False Beliefs

It may be worth mentioning that in social choice theory there is a second standard objection against the *ex ante* Pareto condition, as understood in a particular way.[28] Some authors find *ex ante* Pareto improvements problematic when they depend on differences in people's *beliefs*. To see this worry, suppose Red and Blue have an apple and an orange each, but Red does not like apples at all, and Blue does not like oranges at all. In this case, there is

[28]Cf. e.g. Gilboa et al. (2014), and Mongin and D'Aspremont (1998, 442).

a Pareto superior state in which Red gets both oranges and Blue gets both apples. This state is possible because the two have different tastes, and it seems clearly preferable from the point of view of general goodness. But now suppose White is certain that it will rain tomorrow, and Brown is certain that it will snow, and suppose they both own £100 (which they both cherish very much). Relative to their own beliefs, it is *ex ante* better for both of them to agree to the bet in which White will receive Brown's £100 if it rains, and Brown will receive White's £100 if it snows. However, since one of them clearly has false beliefs, there seems to be something problematic about their agreement, even if it is an *ex ante* Pareto-improvement. This may be particularly so, for example, if one of them was intentionally ill-informed about the weather.

However, this objection too does not carry over to our context. I do allow that the probability distribution over outcomes in an option in $\mathcal{Q}$ may differ from theory to theory. But that is not because axiologies themselves somehow assign different probabilities to outcomes – let alone because they do that in some objectionable way. Rather, I am simply stipulating that these probabilities are objectively correct. So whatever exactly it is that we find problematic about belief-relative Pareto improvements, that will not apply to our Pareto Condition.

Conclusion

Perhaps there are other kinds of objections to our Pareto Condition, other than reasons that parallel the egalitarian objection, or the objection from

false beliefs. But I cannot see any that I find remotely convincing. So as far as I see, in our context, the *ex ante* Pareto Condition is indeed very plausible. It seems to be a major drawback if a theory violates it.

2.4.3 The Decision-Theoretic Explications

Let me finally discuss the decision-theoretic explications of intra- and intertheoretic comparisons. These do not appear as substantial conditions of the Expected Value Theorem, but they were nonetheless necessary to derive that theorem from Karni and Schmeidler's result. And one might have worries about these definitions.

Some people worry that the decision-theoretic explication of intertheoretic comparisons makes the argument from Karni and Schmeidler's Theorem to EVM *circular*. Indeed, it may seem that the explication simply *defines* quantities of value in a way that renders EVM true – that in assuming that intertheoretic comparisons acquire their cardinal significance in the context of weighing axiologies under uncertainty I must already *assume* that EVM is correct. And this would obviously be a vicious circle.[29]

This objection is misguided. The argument from Karni and Schmeidler's Theorem to EVM is not circular. To see this, it may be helpful to distinguish two assumptions that are involved in the explication. The first assumption is that, if a theory-dependent utility function represents the u-value rela-

[29]This objection seems to be raised (or suggested) by Andrew Sepielli (2009, 27): 'The main problem with [the decision-theoretic explication of intertheoretic comparisons] is that it simply assumes the rationality of maximizing [expected value] under normative uncertainty. But this is a position that should be argued for independently of one's solution to the [problem of intertheoretic comparisons], not merely assumed as a means to solving the problem.'

tion ordinally, and represents each axiology cardinally – and in particular, if $\succeq_U$ thus satisfies *Completeness* – then all axiologies are *somehow* fully comparable, and can *somehow* jointly be represented by single value-functions. This claim does not say anything about which specific value-functions jointly represent our theories, or how that is determined. It merely says that axiologies are somehow fully comparable. The second assumption involved in our explication is that the *specific* value-functions that can figure in a joint representation of our theories are then determined, in a certain way, by the u-value relation in which those theories are involved. In my explication, I have joined those two assumptions together. I could also have stated them separately, and perhaps that would have aroused less suspicion of circularity.

Consider the former assumption first. This is simply a claim about what (perhaps among else) it means that two theories are comparable. And though the truth of this assumption is necessary for EVM, it does by no means *presuppose* EVM. Moreover, the assumption simply does not seem very problematic. As I shall elaborate in the next chapter, there might be stronger and weaker senses in which axiologies can be 'comparable'. But at least in the sense that is relevant for our purposes – i.e., in the sense that is relevant for the u-value relation – the assumption seems innocuous. If there is a theory-dependent utility function that represents the u-value relation ordinally, and represents each axiology cardinally, there is a unique way in which axiologies weigh against each other to determine a complete u-value relation. As far as the problem of axiological uncertainty is concerned, that simply means that all axiologies are 'comparable'. I cannot see anything problematic about this. And in any case, again, it clearly does not presuppose EVM.

So suppose this first assumption is true: if a theory-dependent utility function represents the u-value relation ordinally, and represents each axiology cardinally, then all axiologies can *somehow* jointly be represented by value-functions. In that case, if there is such a utility function, then the respective state-wise utility functions must be positive affine transformations of the value-functions of our theories. That is, if G_i are our value-functions, there must be numbers s_i and t_i in $\mathbb{R}$, $s_i > 0$, such that for all options $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \boldsymbol{a}(i,x)[s_i G_i(x)+t_i] \geq \sum_{(i,x)\in Z} \boldsymbol{b}(i,x)[s_i G_i(x)+t_i]. \quad (2.16)$$

Hence the only alternative way in which our value-functions could determine the u-value relation would be if, apart from their probabilities, one theory had systematically and constantly across decisions more weight in determining u-value than another (in that $t_i \neq t_j$ or $s_i \neq s_j$ for some i and j). So the second assumption – that intertheoretic comparisons acquire their *specific* cardinal significance in the context of axiological uncertainty – does not rule out risk-aversion, or the quasi-deontological or virtue-ethical accounts sketched in section 2.3.3. It only rules out such constant unequal weighing. And although EVM presupposes that u-value is not determined by such unequal weighing, to turn this final step into a matter of definition is again by no means to assume EVM.

What *may* be questioned is whether there *is* a utility function that represents $\succeq_U$ in the relevant way – that is, whether $\succeq_U$ satisfies the conditions of our theorem, and particularly Completeness. *These* assumptions are doing

the important work in guaranteeing intertheoretic comparability. Once these assumptions are in place, the explication of interhteoretic comparisons itself does not do very much further work, and certainly not enough to warrant a charge of circularity.

Let me discuss another objection against the decision-theoretic accounts of intra- and intertheoretic comparisons. I have suggested in section 2.3.2 that we do not use cardinally significant concepts of value in ordinary language, that we therefore cannot *find out* what cardinal intra- and intertheoretic comparisons *really* mean, and have to choose some explication to decide what we shall mean by them. One might object to this. One might argue that we do have a pretheoretic understanding of comparisons of *value differences*. For example, *prima facie*, we seem to understand statements like 'investing these resources in poverty reduction would do more good than investing them in repainting the town hall'. That is, *prima facie*, we seem to understand statements of the form

(F) according to T_i, the value difference between x and y is *greater than* the value difference between z and t.

Now, it is not true in general that a comparative ranking of value differences among a set X of outcomes will entail a cardinal concept of value. For example, suppose that X has only three members, and that according to some theory, the difference between the first and the second outcome is greater than the difference between the second and the third. Clearly, this is not enough to yield a utility function that is unique up to positive affine transformation.[30]

[30]Cf. e.g. Bossert (1991, 212) for such an example.

However, under *some* conditions – if the relevant set of outcomes is rich enough – difference comparability is enough to yield cardinal measurability. That is, under some conditions, if a theory entails a comparative ranking of value differences, there will be a value-function representing that theory that is unique up to positive affine transformation.[31] So if these conditions hold, an intuitive understanding of value difference comparisons would be enough to provide a cardinally significant concept of value. So one might argue that *this* is where cardinally significant comparisons get their significance, indeed that this is what they must *really* mean, and that the decision-theoretic 'explication' is therefore inadequate.

However, I am not entirely convinced by this proposal. For one thing, I am not convinced that we have a clear understanding of value difference comparisons. We might understand (or think we understand) such comparisons in, as it were, unproblematic contexts where various aspects of our pretheoretic understanding of value differences come together. For example, if investing our resources in poverty reduction would do more good than investing them in repainting the town hall, that probably implies that it would now be better to invest them in poverty reduction than in paint, that it would be better to invest them tomorrow in poverty reduction than today in paint, that a prospect in which they are invested with a 50% chance in poverty reduction is better than an equivalent prospect in which they are invested with a 50% chance in paint, and so on. But as soon as these aspects come apart, it is much less clear whether we have a direct grasp on value difference compar-

[31] Cf. e.g. Basu (1983) and Bossert and Stehling (1994) for statements of sufficient conditions; e.g.Bossert (1991), Bossert and Weymark (2004, 1126ff.) and Bossert et al. (2005, 34f.) for a general discussion.

isons. Consider again the view I introduced on page 52, on which the best option is the one that maximises the expected square root of value:

$$a \succeq_i b \quad \text{iff} \quad \sum_{x \in X} a(x)\sqrt{G_i(x)} \geq \sum_{x \in X} b(x)\sqrt{G_i(x)}. \tag{2.17}$$

As I have shown, on this theory, two value differences can count the same in determining the value of prospects even though they are not the same. I'm inclined to think that unless more is said we do not understand this view. If someone claims to endorse (2.17), I think he owes us an explanation of what he means by it. For example, suppose someone says that for him the *ex post* value of money increases linearly with monetary value, but in his betterness ordering of prospects he is risk averse about monetary value. I think he does need to tell us in what sense he (*ex post*) values money linearly instead of assigning it diminishing marginal value, if that is how it appears in his ordering of prospects. Our grasp of value difference comparisons is as it were not robust enough, or not precise enough. I think we really do need an explication.

But moreover, even if we grant that we have a direct understanding of *intra*theoretic value difference comparisons, that is only half of what we need to formulate EVM. EVM also presupposes *inter*theoretic value difference comparisons – statements of the form

(G) the value difference between x and y, according to T_i, is greater than the value difference between z and t, according to T_j.

And it seems fairly uncontroversial that we have no direct grasp on statements like (G). That is the difficulty of the problem of intertheoretic com-

parisons. So ultimately, our direct grasp of value difference orderings will in any case not be enough.

In the next chapter, as a possible solution to the problem of intertheoretic comparisons, I shall introduce fitting attitude accounts of value. I shall consider whether an ordering of the strength of *attitudes* corresponding to value differences might provide a cardinally significant concept of value. As I mentioned on page 63, I cannot rule this out entirely. But I am very sceptical about it. And added to this, fitting attitude accounts in general face problems. So I shall ultimately proceed without assuming the truth of such an account. And at least without these accounts, again, we do not seem to have the necessary grasp on intertheoretic value difference orderings.

So let me consider a final worry about the decision-theoretic explications. One might think that these explications get the *order of explanation* the wrong way round. Consider the intratheoretic case first. Take two prospects a and b, and suppose that (according to the true axiology), a is better than b. Intuitively, it's not that outcomes compare in a specific way *because* there is the brute and inexplicable fact that a is better than b; rather, it seems, a is better than b precisely *because* outcomes compare in a specific way. So my explication seems to put the cart before the horse: it explains comparisons in terms of betterness relations, whereas actually betterness relations seem to be explained by comparisons. And the same worry arises in the intertheoretic case. Intuitively, it's not that axiologies compare in some specific way because certain u-value relations hold; rather, these u-value relations hold *because* axiologies compare in a specific way.[32]

[32]This objection is made in MacAskill (2014, 146); it is often raised against similar

Let me say two things in response to this objection. For simplicity, I shall only discuss the intratheoretic case; but what I say applies *mutatis mutandis* to intertheoretic comparisons. First, a point of clarification: nothing in the decision-theoretic explication implies that the value relations must be 'brute and inexplicable'. In fact, according to the explication, value relations on prospects might well be explained by facts about outcomes. To illustrate, suppose that according to the true axiology, both natural beauty and wellbeing have value; suppose option a leads to the destruction of some natural beauty with a probability of 0.4, and to some benefits for Blue with a probability of 0.6; option b leads to the status quo:

	a	b
0.4	destruction of natural beauty	status quo
0.6	benefits for Blue	status quo

Table 2.5

And suppose that a and b are equally good. This may well be explained by the fact that a – while having a good probability of benefiting Blue – also has a decent probability of destroying some natural beauty. What the decision-theoretic explication implies is only that it cannot be explained by the *cardinal* fact that the value of benefitting Blue is 2/3 times the disvalue of the destroying the bit of natural beauty. That is, the explication does not imply that *no* facts about outcomes can explain the value relation among prospects. It only implies that no independent *cardinal* facts can do so.

And this reveals the second point I wish to make: claiming that the value

explications in slightly different contexts (cf. e.g. Eriksson and Hájek (2007, 207)).

relation is explained by this *cardinal* fact presupposes an independent cardinal notion of intratheoretic comparisons. So if the present objection is to get off the ground, it has to provide an alternative account of such cardinal facts. And if my response to the previous objection was sound, our direct grasp of value difference comparisons is insufficient to provide one.

Now, as I indicated on page 53, there *are* other possibilities for explicating a cardinal concept of value in contexts where values are weighed against each other. For example, we could assume that values acquire their cardinal significance in the context of weighing goods over time. So one *could* substantiate the present objection against my explication by using a different explication of the relevant cardinal facts. But my guess is that with *all* these explications, it will *prima facie* seem that they get the order of explanation wrong. With this explication concerning time, it will seem that two differences in value coming at different times count the same in determining the goodness of the history of the world over time, *because* they are the same. In terms of this order-of-explanation objection, nothing seems to be gained by the move from one explication to the other.

But this *tu quoque*-response does not imply that the entire project of getting a cardinal concept of value is hopeless. Instead, I take it, it reveals that the intuitive appeal of the order-of-explanation objection is deceptive. We may intuitively feel that we understand the alleged explanation that this objection relies upon. But on closer inspection, I think that that assumption is dubious, and the objection gets beaten at its own game.

2.5 Further Explorations: Expected Value and u^2-Value

Before I conclude this chapter, it may be interesting to note an upshot of my discussion. It concerns the regress problem that I mentioned in section 1.3. As I explained, a regress threatens because we might not be certain about the true theory of axiological uncertainty. If so, we need a theory of uncertainty about theories of axiological uncertainty – a theory of u^2-value, and so on.

Now, one might think that, if Expected Value Maximisation is a plausible theory of u-value, it is also a plausible theory of u^2-value – and in fact, that the argument I have provided for EVM as a theory of u-value can straightforwardly be extended to Expected Value Maximisation as a theory of u^2-value, using theories of axiological uncertainty instead of axiologies, and the u^2-value relation instead of the u-value relation.

Unfortunately, however, this is true only to a very restricted extent. As I said, the Expected Value Theorem presupposes that all first-order theories (i.e. all axiologies) are vNM-conformable; and it shows that, then, EVM is the only theory of u-value that is vNM-conformable and satisfies the Pareto Condition. Accordingly, on the level of u^2-value, a similar argument will presuppose that all theories of u-value are vNM-conformable. But – as the Expected Value Theorem shows – apart from theories that violate the Pareto Condition, EVM is the *only* theory of u-value that is vNM-conformable. And as I argued in section 2.4.2, as far as u-value is concerned, the Pareto Condition is very plausible. So the set of theories of u-value that

are vNM-conformable but not Paretian is not a very interesting set. The more interesting set of theories are probably those that fail to satisfy the von Neumann-Morgenstern axioms. It follows that my argument, applied to u^2-value instead of u-value, is almost completely uninteresting.

More generally, we cannot simply assume that the same kind of argument will in principle be available, and of the same importance, on each level of value. If representation theorems are indeed as important as I will argue throughout this thesis – not only in defending, but in so much as *defining* our views – then the problem of uncertainty about theories of axiological uncertainty is even more serious than it might have seemed. We may not even be able, in any interesting sense, to *define* our views of uncertainty about theories of axiological uncertainty.

Conclusion

In this chapter, I have shown that state-dependent utility theory provides a promising framework for an axiomatic approach to axiological uncertainty. More specifically, I introduced a simple representation theorem from state-dependent utility theory, and derived from it the Expected Value Theorem about axiological uncertainty: given the decision-theoretic explications and the assumption that all axiologies under consideration are vNM-conformable, EVM is true if $\succeq_U$ is vNM-conformable and satisfies the Pareto condition.

I argued that the u-value relation is not complete if some axiologies are not fully comparable, and that it is not continuous if some compare in a lexical

way. So the Expected Value Theorem raises the problem of intertheoretic comparisons. Apart from this problem, and if all underlying axiologies are vNM-conformable, I suggested that the Expected Value Theorem presents a powerful means both to formulate EVM and to defend it against alternative theories of u-value.

Chapter 3

Intertheoretic Comparisons of Value

Introduction

The most important question that the Expected Value Theorem raised is the problem of intertheoretic comparisons – the question whether sizes of value differences or heights of value levels can be compared across axiologies. So in this chapter, I shall address this problem in more depth.

In section 3.1, I shall state more precisely what the problem of intertheortic comparisons is, and what I take to be the main motivation for denying that intertheoretic comparisons are possible.

In section 3.2, I shall outline what I call the 'Minimal Argument': an argument to the effect that at least some intertheoretic comparisons are possible. This argument does not provide a positive account of why these comparisons hold, nor does it show that a large number of comparisons do. Instead, it

claims that absolute scepticism about intertheoretic comparisons has unacceptably implausible implications, and thus that at least *some* intertheoretic comparisons must *somehow* be possible.

Assuming that this is true, I then explore two accounts of *why* it is true. In section 3.3, I discuss Comparativism, the view that axiologies are merely orderings, and that there are independent facts about how they enter the u-value relation, and thus how they compare. I shall defend Comparativism against various objections, and argue that it entails a kind of positive slippery slope: if Comparativism explains why *some* intertheoretic comparisons are possible, a very large number of theories are plausibly comparable. Finally, I shall outline how Comparativism is, or can be made, compatible with the Expected Value Theorem.

In section 3.4, I do the same for Absolutism, the view that axiologies are more than merely orderings, and themselves make claims about the sizes of value differences or heights of value levels on an intertheoretic scale. I show that Absolutism too entails a broad range of intertheoretic comparisons, and outline how it is, or can be made, compatible with the Expected Value Theorem. For reasons I shall outline, I personally find Absolutism more plausible than Comparativism. So for most parts of chapters 4 and 5 I shall assume a form of Absolutism.

In all these sections, I assume that if theories are not comparable, then the u-value relation is incomplete. To end this chapter, in section 3.5 I shall explore the alternative assumption that the u-value relation may be complete *even if* intertheoretic comparisons are impossible. I shall provide axiomatisations for My Favourite Theory and Weighted Value Maximisation

on the basis of this assumption. But ultimately, I will reject this alternative approach.

3.1 The Problem

To clarify what I mean by the problem of intertheoretic comparisons, let me introduce some definitions. Let a *positive fact about intertheoretic comparisons* be a fact of the form

(A) the value difference between x and y, according to T_i, is greater than the value difference between z and t, according to T_j; or

(B) the value of x, according to T_i, is greater than the value of y, according to T_j.

In most parts of this chapter, I shall not be concerned with the distinction between intertheoretic unit-comparisons (like (A)) and intertheoretic level-comparisons (like (B)). I shall say that two vNM-conformable axiologies are *fully incomparable* if no positive facts about intertheoretic comparisons hold between them. And as stated roughly on page 72, I shall say that two vNM-conformable axiologies are *fully comparable* if they can jointly be represented cardinally by a theory-dependent utility function u, such that the crosscutting cardinal intertheoretic comparisons between them are the same as the respective intertheoretic utility difference ratios on u (i.e., if they can jointly be represented cardinally by two utility functions that are unique up to positive affine transformation by the same scalar and constant).

Thus understood, full comparability and full incomparability are not exhaustive alternatives. It might be that no two axiologies in a set of vNM-conformable theories are fully comparable, but that there are facts of the form 'the difference between the value of x, according to T_i, and the value of y, according to T_j, is at least n times and at most m times as great as the difference between the value of z, according to T_h, and the value of t, according to T_k', for some $n, m \in \mathbb{R}$. If such facts hold for all pertinent value differences, we might say that these theories are *roughly comparable*.

By the *problem of intertheoretic comparisons* I mean the question whether there are some positive facts about intertheoretic comparisons – or, as I shall say equivalently, whether intertheoretic comparisons are possible. At least, that is how I understand this problem for now. On page 106, I shall mention another understanding of the problem; but this is what I mean for now, and unless otherwise stated. *Scepticism about intertheoretic comparisons*, or just *scepticism*, is the view that there are no such facts.

The main motivation behind scepticism, I think, is the rough thought that axiologies are simply sets of claims about which options are better than which. According to sceptics, axiologies do not contain any information about how good options are on some global, intertheoretic value-scale or compared to other axiologies, nor are there any independent facts that would determine that. Many philosophers have endorsed scepticism, or at least something close to it.[1]

[1]To be precise, none of the people quoted in the following explicitly say that all axiologies are fully incomparable. So I am not sure whether they endorse what I called 'scepticism', or some less radical claim as that *most* standard axiologies are fully incomparable. It does not matter, as I am not concerned with exegetical questions, and I shall argue against both of these claims.

For example, Edward Gracely considers a form of person-affecting utilitarianism and total utilitarianism. He asks:

> is a small loss of utility as seen by a [person-affecting utilitarian] more or less important under that theory than a large loss of utility (involving lives not created) under total utilitarianism? I don't quite see how this question could be answered. (I'll refrain from saying that it is like comparing apples and oranges, but it is!) [...] There is no abstract scale of "wrongness" outside of the rank provided *within* a theory. (1996, 331)

Similarly, John Broome is concerned with the fact that total and average utilitarianism have different 'units of value' (2012, 185): wellbeing, and well-being per person respectively. He says:

> We cannot take a sensible average of some amount of well-being and some amount of well-being per person. It would be like trying to take an average of a distance, whose unit is kilometres, and a speed, whose unit is kilometres per hour. Most theories of value will be incomparable in this way. (2012, 185)

And in a similar vein, James Hudson imagines a person who has some credence in the view that pleasure-minus-pain is the only good (its units being 'hedons'), and in the view that self-realization is the only good (its units being 'reals'). He argues:

> What is the common measure between hedons and reals? Note that the agent, for all her uncertainty, believes with complete confidence that there is *no* common measure: she is sure that one or the other – pleasure or self-realization – is intrinsically worthless. Under the circumstances, the two units must be incomparable by the agent, and so there can be no way for her uncertainty to be taken into account in a reasonable decision procedure. (1989, 224)

Other people have expressed similar views.[2]

[2]Cf. Gustafsson and Torpman (2014), Hedden (forthcoming).

As I explained in the last chapter, if the conditions of the Expected Value Theorem hold, there are positive facts about intertheoretic comparisons. So if scepticism is correct, the argument based on the Expected Value Theorem – as well as my main arguments in chapters 4 and 5 – are not sound. So let me now argue against this view.

3.2 The Minimal Thesis and Two Explanations

There is a close analogue in social choice theory to the problem of intertheoretic comparisons – viz., the problem of interpersonal comparisons of utility, or wellbeing; the question whether, or how, the wellbeing enjoyed by one person can be compared to that enjoyed by another. It is a difficult question what precisely the basis for such interpersonal comparisons is, or what precisely they mean. But I think it is helpful to begin this discussion in social choice theory with a simple observation. It is a plain fact that in everyday life – say, in cases of minor distributive moral problems – we frequently make (what look like) interpersonal comparisons of wellbeing; and in these contexts, there does not seem to be anything spurious about them.[3] This observation does not *prove* that interpersonal comparisons are possible, nor does it explain what the basis of such comparisons is. But it does strongly suggest that *something* has gone wrong if we deny the possibility of interpersonal comparisons altogether.

[3]This is observed e.g. by List (2003, 229).

I think it is helpful to begin a discussion of intertheoretic comparisons with a dialectically similar observation. So let me begin by giving the *Minimal Argument*. I will introduce this argument in section 3.2.1, and elaborate on it in section 3.2.2.

3.2.1 The Minimal Argument

In the last chapter, I suggested that u-value plausibly satisfies Transitivity, Independence and the Pareto Condition. I mentioned that Continuity is false if some axiologies compare in a lexical way. However, if there are such axiologies, then intertheoretic comparisons are in any case possible. So it is only a concession to scepticism to assume that such axiologies do not exist. And I have suggested that, barring such axiologies, Continuity is plausible too. So for present purposes, let me assume Transitivity, Independence, Continuity and the Pareto Condition. If $\succeq_U$ satisfies these conditions, we can give the following argument.

(C) The u-value relation is not radically incomplete;

(D) if the u-value relation is not radically incomplete, intertheoretic comparisons are possible; therefore

(E) intertheoretic comparisons are possible.

As mentioned, I call this the *Minimal Argument*. It is minimal in two ways. First, if sound, it establishes only the comparatively weak claim that there are *some* positive facts about intertheoretic comparisons. The argument does

not show that all, or most of our standard axiologies are comparable. Second, it does not provide any positive explanation for why these comparisons are possible. If sound, it only implies that intertheoretic comparisons must *somehow* be possible. I shall thus call (E) the *Minimal Thesis*. For now, this Minimal Thesis is all I am concerned with. Let me elaborate on premises (C) and (D) in turn.

Premise (C)

What do I mean by premise (C)? Say that two options $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$ are *in the same possibility-space* if they assign non-zero probability to exactly the same axiologies, i.e., if for all i in I, $\boldsymbol{a} \in \mathcal{Q}^i$ if and only if $\boldsymbol{b} \in \mathcal{Q}^i$. Let a binary relation $\succeq$ on $\mathcal{Q}$ be *radically incomplete* if for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in the same possibility space,[4]

$$\boldsymbol{a} \succeq \boldsymbol{b} \quad \text{only if} \quad H_i(\boldsymbol{a}) \succeq_i H_i(\boldsymbol{b}) \ \text{ for all } i \text{ in } I \text{ with } a, b \text{ in } \mathcal{Q}^i. \tag{3.1}$$

So intuitively, if the u-value relation is radically incomplete, then whenever there is one theory with nonzero probability according to which $\boldsymbol{b}$ is better than $\boldsymbol{a}$, $\boldsymbol{a}$ cannot be at least as u-good as $\boldsymbol{b}$. (C) claims that this is not so. It says that, *sometimes*, an option $\boldsymbol{a}$ is at least as u-good as another option $\boldsymbol{b}$ even if, according to some axiologies with nonzero probability, $\boldsymbol{a}$ is worse.

This is very plausible, I think. To see how extreme the denial of (C) is, consider a toy example. Take some version of standard total utilitarianism, and a form of 'anthropocentric total utilitarianism', according to which one

[4]As a reminder, $H_i : \mathcal{Q}^i \to \mathcal{O}$ was defined as $H_i(\boldsymbol{a})(x) = \boldsymbol{a}(i,x) / \sum_{y \in X} \boldsymbol{a}(i,y)$.

option is better than another if and only if it leads to more total *human* wellbeing (while the wellbeing of non-human animals is irrelevant).[5] Suppose you can control the fate of some enormous number of animals and of one human being. You can choose among the outcomes

- x all the animals live very long, happy and painless lives but the person suffers from a pinprick; and
- y all the animals live very long lives full of torture and agony, but the person does not suffer from the pinprick (and lives otherwise as in x).

According to the anthropocentric utilitarian theory, y is better than x; according to the standard theory, x is better than y. Suppose you find the anthropocentrism of the former very repulsive, and thus assign it a 0.1% probability – as opposed to the 99.9% you assign to total utilitarianism. It is very implausible that you necessarily have to judge your options incomparable in u-value. Surely, you can truly claim that x is u-better than y.

If so, this is enough to guarantee that (C) is true. In fact, we could consider two theories that are even more similar to each other than standard and anthropocentric utilitarianism, and construct an even more radical example. It is enough if there is one single case where the u-value relation is not radically incomplete. So denying (C) really is a very extreme view. Perhaps it is implausible that $\succeq_U$ is *fully* complete. I shall consider that in more detail below. But it is definitely implausible that $\succeq_U$ is radically incomplete. I think that this is something we can say even before we give any positive account of *why* any intertheoretic comparisons hold.

[5] I thank William MacAskill for suggesting this example to me.

Premise (D)

Premise (D) says that if the u-value relation is not radically incomplete, intertheoretic comparisons must be possible; or, if intertheoretic comparisons are not possible, the u-value relation must be radically incomplete. Given our other conditions, I think we can treat this as a matter of definition. I suggested in the last chapter that, as a matter of definition, if the u-value relation is *complete* and satisfies Transitivity, Independence, Continuity and the Pareto Condition, then all axiologies are *fully* comparable. Similarly, I take it, *if* the u-value relation satisfies Transitivity, Independence, Continuity and the Pareto Condition, we can assume as a matter of definition that if the u-value relation is not radically incomplete with respect to two underlying axiologies, these axiologies are not fully incomparable. At least in a sense that is relevant for us – that of how they determine the u-value relation – they will then be (at least) roughly comparable. They will weigh in a particular way against each other, and we can thus say that they compare in this way.

Unfortunately, the technical aspects of this definition will be slightly more complex. Roughly speaking, given all our conditions and premise (C), we can represent axiologies by *sets* of utility functions. And we can say, for example, that the value difference between x and y, according to T_i, is greater than the value difference between z and t, according to T_j, if that is true on all utility functions in the sets that represent these axiologies. But I shall explore that on a technical level only in chapter 5. I think it is intuitively clear, and for

now I will content myself with that. What I said on a non-technical level in defence of the decision-theoretic explication of full comparability will carry over to this present context, to the idea of roughly comparable axiologies. Assuming Transitivity, Independence, Continuity, the Pareto Condition and premise (C), taking (D) to be a matter of definition is not circular, or overly stipulative, or wrong in some respect.

So I believe that the Minimal Argument is sound. Very plausibly, there are at least some positive facts about intertheoretic comparisons.

But before moving on, let me elaborate again on My Favourite Theory. As I have argued, My Favourite Theory violates the Pareto Condition, and this is a major drawback. But there is another, and in a sense more fundamental problem with My Favourite Theory. Note that according to My Favourite Theory, (C) is true: $\succeq_U$ is not radically incomplete. Instead, $\succeq_U$ is complete (or as complete as the most plausible axiology is). However, people who endorse My Favourite Theory generally think that it is compatible with scepticism – indeed, they take scepticism to *motivate* My Favourite Theory.[6] In principle, this does not contradict anything I have said. I have only claimed that scepticism is incompatible with (C) if $\succeq_U$ among else satisfies the Pareto Condition. But it nonetheless contradicts, as it were, the spirit of premise (D), which suggests that scepticism should give rise to incompleteness. So let me say a word about the inference from scepticism to My Favourite Theory.

I think that this inference is very dubious. Take the analogy with intratheoretic comparisons again. Suppose art and philosophy are radically

[6]Cf. e.g. Gracely (1996) and Gustafsson and Torpman (2014).

incomparable in the sense that, for *any* pair of lives, if one contains more philosophy and one contains more art, then they are neither exactly equally good, nor is one of them better. Now consider the following two options:

	a	b
0.6	Comparatively good philosophy; no art	Comparatively bad philosophy; no art
0.4	Comparatively bad art; no philosophy	Comparatively good art; no philosophy

Table 3.1

With a probability of 0.6, both a and b lead to lives containing only philosophy, and a contains better philosophy than b; with a probability of 0.4, they both lead to lives containing only art, and b contains better art than a. If these values are radically incomparable in the way I assumed, no positive value relation should hold between a and b. The betterness relation should be incomplete with regards to these options. There is no reason for just focusing on the probabilities, and ranking a strictly better than b, just because philosophy-lives are the most likely outcome and a promises better philosophy than b. This would simply be to *disregard* the incomparability between philosophy and art, and pretend that the difference between good and bad art is as great as that between good and bad philosophy.

More generally: when we determine the value of a prospect, *absent* any incomparability, it is extremely implausible that we can focus on probabilities only and ignore the values of outcomes. Both the probabilities of outcomes *and* their values are relevant to determine the goodness of prospects. One might say that this is the basic tenet of decision theory, in its simplest form.

So if there *is* incomparability in values, there is no reason why we could ignore *that*, and just focus on probabilities.

Similarly, there seems to be no positive reason why scepticism about intertheoretic comparisons should lead to a unique focus on axiological probabilities, and thus to My Favourite Theory. Accepting My Favourite Theory seems like an *ad hoc* solution for sceptics to accommodate (C). That is another drawback of this view.

But in any case, My Favourite Theory also violates the other conditions I have outlined. And since I take these to be plausible, I think that My Favourite Theory is false, and the Minimal Argument is sound. Very plausibly, not all axiologies are fully incomparable.

One might perhaps say that whether or not the Minimal Thesis (E) is true is not the most important question. The *real* problem of intertheoretic comparisons, one might say, is not whether *all* axiologies are fully incomparable, but whether most, or most standard axiologies are. And from a practical point of view, that is certainly true. But I think *philosophically*, the truth of (E) is actually very significant. The reason is that once we assume that there are *some* positive facts about inthertheoretic comparisons, we open the door for a vast number of them. At least, that is what I shall argue in the remainder of this chapter.

3.2.2 Two Explanations

If the Minimal Thesis (E) is true, there must be some explanation for *why* it is true, or why intertheoretic comparisons are possible. Various classify-

ing schemes have been proposed for distinguishing accounts of intertheoretic comparisons.[7] I think the most fundamental distinction is between two types of accounts that differ in terms of their conception of what axiologies are.

On the first conception of axiologies, axiologies are only betterness orderings of prospects. These orderings may be cardinal orderings – that is, they may give rise to facts about the ratios of value differences. But they do not themselves make any claims about the heights of value levels or sizes of value differences on some *inter*theoretic scale. If there is nonetheless a fact about how these orderings compare intertheoretically in a relevant sense, then that is some independent principle or fact about how axiological orderings enter the u-value relation; axiologies themselves are silent on it. I shall call this view of what axiologies are *Comparativism*; and I shall refer to accounts of intertheoretic comparisons that presuppose Comparativism as *comparativist accounts (of intertheoretic comparisons).* At least as far as the problem of intertheoretic comparisons is concerned, we might say that Comparativism is a conservative view about axiologies: it does not assume that axiologies themselves make claims about the heights of value levels or sizes of value differences on some intertheoretic scale. So for comparativist accounts of intertheoretic comparisons, I take it, the challenge is not to defend Comparativism itself. Instead, the challenge is to argue that – even though axiologies do not make these relevant claims – there are nonetheless positive facts about intertheoretic comparisons.

On a second conception of what axiologies are, axiologies themselves make claims about the heights of value levels or sizes of value differences on some

[7] Cf. e.g. MacAskill (2014, 129ff.).

intertheoretic scale. That is, an axiology is not only an ordering of prospects, but also in itself makes claims about how good certain options are, or how much better some are than others, on a global, intertheoretic scale. I shall call this view of what axiologies are *Absolutism*; and I shall refer to accounts of intertheoretic comparisons that presuppose Absolutism as *absolutist accounts (of intertheoretic comparisons)*. If Absolutism is true, we have a straightforward explanation for why intertheoretic comparisons are possible: axiologies themselves simply imply the relevant facts. So the challenge for absolutist accounts of intertheoretic comparisons, I take it, is to defend Absolutism itself – to explain why, or in what way axiologies imply these intertheoretic facts.

I am not aware that anyone has made this distinction between Absolutism and Comparativism.[8] It will become clearer when I elaborate on the two views in the next sections. But one helpful way to see the distinction is via the question of how many axiologies there are. If Comparativism is true, there is only one axiology corresponding to any ordering of prospects. If Absolutism is true, there are infinitely many axiologies corresponding to any ordering. The reason is that, if Absolutism is true, a comprehensive theory of goodness must be more than a theory of betterness: it has to include claims about the sizes of value differences or heights of value levels on an intertheoretic scale. And if such claims make sense, then one can

[8]MacAskill (2014, ch.4) also makes a distinction between 'Absolutism' and 'Comparativism' in the context of the problem of intertheoretic comparisons. However, his distinction concerns the question whether value-relations ground value-properties, or vice versa. As far as I understand it (and as he tells me), my distinction is different from his: a view that is absolutist in my sense could be either absolutist or comparativist in MacAskill's sense.

consistently and meaningfully combine one and the same claim about the betterness ordering with different claims about this intertheoretic scale. The intuitive reason for this is that an ordering can be represented by infinitely many utility functions: there is simply nothing in an *ordering* that would make it inconsistent to combine that ordering with certain claims about this global scale. And that it is consistent for a comprehensive theory of goodness to combine one and the same betterness ordering with different claims about the sizes of value differences or heights of value levels on a global scale, simply means that there *are* theories that imply the same betterness ordering but different claims about that global scale. That different theories can be meaningfully formulated is what it means for there to *be* different theories. So if Absolutism is true, there are infinitely many axiologies corresponding to any ordering.[9]

Comparativism and Absolutism are mutually exclusive and (at least on one level of generality) jointly exhaustive. Axiologies either do or do not make claims about the heights of value levels or sizes of value differences on some intertheoretic scale; so if (E) is true, any explanation for this must either be absolutist or comparativist. So let me examine both of these views in turn.

[9]I should note that I do not take 'Absolutism' to mean that axiologies in themselves make claims like 'the value of x is 5', or 'the value of x is 10'. Even on Absolutism, axiologies will imply such claims only once we picked a scale (as explained on page 51). But once we *have* picked a scale, on Absolutism, axiologies themselves imply value-facts on that scale; on Comparativism, axiologies themselves do not imply such facts, and if there nonetheless are such facts, they are implied by independent principles.

3.3 Comparativism

Again, if Comparativism is true, and axiologies themselves do not make any claims about the heights of value levels or sizes of value differences on some intertheoretic scale, then if the u-value relation is not radically incomplete, there are *independent* facts that determine how axiologies compare, or weigh against each other to determine the u-value relation.

Considerable work has been devoted to finding principles that integrate axiological orderings into the u-value relation in a plausible seeming way. For example, one principle that has been suggested is what we might call *Best/Worst Normalisation*. According to this principle, under uncertainty, the best and worst conceivable outcomes should be considered equally good and bad according to all theories. To formulate this idea in terms of value-functions, let x_i^w and x_i^b be the worst and best conceivable outcomes respectively, according to T_i. Best/Worst Normalisation then says that $G_i(x_i^w) = G_j(x_j^w)$ and $G_i(x_i^b) = G_j(x_j^b)$ for all T_i and T_j.[10] Another principle that has been suggested is what we might call *Variance Normalisation*. The variance of an axiology is a measure of how, according to that axiology, moral value is spread out over different outcomes – viz., the average of the squared differences in the value of outcomes from their mean value. According to Variance Normalisation, under uncertainty, this quantity should be identical for all theories.[11] A third principle that has been proposed is what

[10]This principle is a slight variation of the 'Principle of Equity Among Moral Theories', as stated in Lockhart (2000, 84); it is considered in Sepielli (2013b, 588), and MacAskill (2014, 103ff.).

[11]A particular interpretation of this idea is endorsed in MacAskill (2014, ch.3); cf. also Cotton-Barratt et al. (ms).

we might call *Paradigm Normalisation*. On this proposal, there are certain paradigmatically heinous outcomes (say, that people are tortured), and certain paradigmatically value-neutral outcomes (say, an empty world with no sentient beings), and the value of these outcomes is the same according to all theories.[12] Labelling these outcomes as x^h and x^n, this proposal says that $G_i(x^h) = G_j(x^h)$ and $G_i(x^n) = G_j(x^n)$ for all T_i and T_j.

In theory, there are infinitely many other principles that could be suggested. And one might also hold a kind of particularist view, on which there are facts about how any two orderings compare, but no very general principle about how all theories do. All these proposals face certain problems. An obvious problem of Best/Worst Normalisation is that there might not be best and worst conceivable outcomes; a problem for Variance Normalisation is that it needs to define a *measure* over outcomes in order to be well-defined;[13] and Paradigm Normalisation faces the problem that there seem to be theories on which these 'paradigmatically' neutral or heinous outcomes are not neutral or heinous. I shall not say which specific comparitivist account I find most plausible, although I shall offer a general consideration on page 122. But before I come to that, let me consider some general objections against Comparativism. That will be the main task of section 3.3.1. In section 3.3.2, I shall then consider how Comparativism can be incorporated into the Expected Value Theorem.

[12] This is suggested in Sepielli (2010, 186f.).

[13] Cf. MacAskill (2014, ch.2) for a suggestion about how to solve this problem.

3.3.1 Possibility, Arbitrariness, and Swamping

The Possibility of a Comparativist Account

As I mentioned, considerable work has been devoted to finding a plausible seeming principle to integrate axiological orderings into the u-value relation. However, one might think that if Comparativism holds, there is a general and fundamental reason why *all* such principles must be false. I think that this is not so, but it is worth getting clear about it.

The thought is this. Consider an axiology T_{AP} on which both artist- and philosophy-lives are valuable. Suppose that T_{AP} implies a complete betterness ordering of the artist-lives, and a complete ordering of the philosophy lives, but on T_{AP}, there are no positive facts of the form 'the value difference between artist-lives L_A^1 and L_A^2 is at least (or at most) as great as the value difference between philosophy-lives L_P^1 and L_P^2'. So according to this axiology, we might say, the ordering of artist-lives and the ordering of philosophy-lives are fully incomparable. As I argued in section 3.2.1, the overall value relation of this axiology should be radically incomplete with respect to art and philosophy.

Now consider two axiologies, an axiology T_A on which only artist-lives are valuable, and an axiology T_P on which only philosophy-lives are valuable. It may seem that – if axiologies are merely orderings and do not make claims about value on an intertheoretic scale – then according to our axiologies, there are again no facts of the form 'the value difference between artist-lives L_A^1 and L_A^2, according to T_A, is at least (or at most) as great as the value difference between philosophy-lives L_P^1 and L_P^2, according to T_P'. So

if on our single axiology T_{AP} the ordering of artist-lives and the ordering of philosophy-lives were radically incomparable, it seems that the axiologies T_A and T_P must be fully incomparable too. Hence, in claiming that axiologies are merely orderings, we seem to be denying that they are comparable, and seem forced to conclude that the u-value relation is radically incomplete. It is no good, it seems, just to outline a *prima facie* plausible principle about how to combine axiological orderings into a complete u-value relation. If Comparativism is true, we have to provide a more fundamental reason for why *any* such principle could be true in the first place.

However, this argument is invalid. Note that our axiology T_{AP} must imply a verdict on whether, or how, artist- and philosophy-lives compare. If it simply remains *silent* on that, it is arguably not a fully specified axiology. So if it is a fully specified axiology, then it positively claims that there are *no* facts of the form 'the value difference between artist-lives L_A^1 and L_A^2 is at least (or at most) as great as the value difference between philosophy-lives L_P^1 and L_P^2'. However, if theories are merely orderings, then they are simply *silent* on the heights of value levels or sizes of value differences on an intertheoretic scale – just like, say, Darwin's theory of evolution is silent on Kepler's laws of planetary motion. By themselves, they simply do *not* make any claims about value-facts on an intertheoretic scale; so neither do they claim that there are *no* such facts. If they did, they would after all imply (negative) facts about an intertheoretic scale; and in that case, there could arguably be theories that imply positive facts about that scale, and so our picture of axiologies would collapse into Absolutism. In slogan form, we might say that the absence of a verdict (about intertheoretic comparisons)

is different from the verdict of absence (of such comparisons).

For this reason, the above analogy fails. For all we know, if Comparativism is true, there may be independent facts – facts that are not part of any axiology itself – that determine how axiologies weigh against each other to form a u-value relation that is not radically incomplete. We might perhaps say that the sense in which theories are then 'comparable' is somehow weaker than under Absolutism. Indeed, under Comparativism, it is slightly misleading to say 'the value difference between x and y, *according to* T_i, is greater than the value difference between z and t, *according to* T_j', since these theories themselves do not *make* claims about the sizes of value differences. They might merely *imply* them, given certain other truths. But that should not mislead us into thinking that the u-value relation would therefore necessarily have to be incomplete. It need not be. If it is not, axiologies would still be 'comparable' in the sense that is relevant for our purposes – i.e., that of featuring in a u-value relation that is not radically incomplete. And given that, I think it is more plausible to accept (E), and the possibility of at least *some* such facts about how axiologies weigh against each other to form a u-value relation, than to deny (E).

Arbitrariness

However, even if (E) could be true under Comparativism: in arguing for it, I considered standard total utilitarianism and a form of anthropocentric total utilitarianism, according to which only human wellbeing matters. These two theories are very similar. So this still leaves open whether only very few

intertheoretic comparisons are possible, or whether in fact an interestingly large number of comparisons are. So let me address some further sceptical arguments that are particularly salient under Comparativism.

One prominent motivation for being sceptical about comparing certain more distinct axiologies is that for some of these theories, any intertheoretic comparisons seem *arbitrary*. For example, Brian Hedden claimed that any comparisons between total and average utilitarianism would be 'arbitrary and unmotivated' (forthcoming, 9). Similarly, Johan Gustafsson and Tom Torpman consider certain allegedly general principles for making intertheoretic comparisons, and conclude from their discussion that 'there does not seem to be any way of making non-arbitrary intertheoretic comparisons of value' (2014, 160). And a similar concern with the arbitrariness of any alleged 'decision procedure' also seems to have motivated James Hudson's scepticism (1989, 224).

However, I think that this arbitrariness-objection can be answered. One criterion that many authors find plausible is the rough idea that apart from their probabilities, all theories should somehow get 'equal say' in the u-value relation. I shall call this the *Idea of Equal Say*. This is certainly the motivating intuition behind Best/Worst Normalisation and Variance Normalisation.[14] I shall consider on page 122 whether this idea is plausible. But suppose for now that it is. As MacAskill (2014, 100ff.) argues, it then offers a response to the arbitrariness objection. The Idea of Equal Say provides a general, non-arbitrary criterion for making intertheoretic comparisons, and that idea can arguably itself be cashed out non-arbitrarily. For example, it

[14]Cf. e.g. Lockhart (2000, 86), MacAskill (2014, ch.3).

can be proved that Variance Normalisation is the only normalisation that satisfies certain *prima facie* appealing conditions (such as that, in a specific sense, all value-functions have the same 'distance' from a uniform value-function that assigns all outcomes the same value).[15] So insofar as these general principles and the Idea of Equal Say are concerned, it is simply not true that any normalisation would be arbitrary.

Perhaps the arbitrariness concern is more pertinent if we assume that the Idea of Equal Say is flawed and that there is no general principle about how axiologies enter the u-value relation – hence that, *if any*, a more particularist picture would have to be true. But I think that the arbitrariness-objection can be answered even then.

As an analogy, consider the class of pluralist axiologies on which there are two different sorts of value. Presumably, there is no very general principle about which of these axiologies is the most plausible in each case. It seems unlikely, say, that the most plausible axiology on which both beauty and well-being have value combines these values in a way that is formally somehow equivalent to the way in which the most plausible axiology on which biodiversity and virtue have value combines *these* two values. So an arbitrariness worry might arise in this context too. Consider the pluralist axiologies on which both beauty and wellbeing have value. Different such axiologies weigh these values differently against each other. On one possible axiology, the disvalue of destroying a work of Raphael is the same as the disvalue of someone's suffering from a pinprick; on another, it is the same as the disvalue of the sufferings caused by the First World War; on others, the comparison

[15] Cf. Cotton-Barratt et al. (ms).

is somewhere in between. Presumably, any precise comparison will to some extent be arbitrary. But that does not mean that the most plausible such pluralist axiology is one on which these values are radically incommensurable – on which the disvalue of destroying any child painting would be incomparable to the disvalue brought about by the First World War. Instead, for one thing, it arguably means that we should spread our credences among different axiologies that make different comparisons between the values of beauty and wellbeing. For another thing, it arguably means that we should assign considerable credence to views on which these values are less than perfectly comparable – axiologies on which, say, destroying a Raphael is worse than giving someone a pinprick, not as bad as a war, but not exactly as good or bad as any wellbeing difference in between. The inference from an alleged arbitrariness of any specific comparison to complete *in*commensurability is simply a *non-sequitur.*

Similarly, suppose Comparativism is true, and consider the axiology on which only beauty is valuable, and the utilitarian axiology on which only total wellbeing is valuable. And suppose we think that these axiologies should not get 'equal say' in determining u-value, in the sense of some general principle like Best/Worst Normalisation or Variance Normalisation. Then even if there is no uniquely privileged way of comparing these two orderings, that does not imply that these axiologies are fully incomparable. At least if it is in principle possible for there to be independent facts about how axiologies determine the u-value relation, it arguably means that we should spread our credences among different comparisons, and particularly, that we should have credence in the view that these axiologies are less than *fully* comparable.

Again the inference from an alleged arbitrariness to complete incomparability is unmotivated.

Swamping

Let me consider a final argument to the effect that certain specific axiologies are incomparable. The argument is given in MacAskill (2014), and concerns the case of total and average utilitarianism – although it could be raised in a very similar way against other pairs of axiologies. Suppose that we have some adequate cardinal representation of wellbeing, and consider the following three options in $\mathcal{O}$ (n being some natural number):

- a leads to n people at wellbeing 100;
- b leads to $10n$ people at wellbeing 99;
- c leads to $1000n$ people at wellbeing 1.

According to the total utilitarian ordering, c is the best option, but b is 99% as valuable as c; according to the average utilitarian ordering, a is the best option, but b is 99% as valuable as a.[16] This is so for any number n.

Now suppose you assign both of these theories probability 0.5, and let $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$ be the corresponding options in $\mathcal{Q}$, relative to this credence distribution.[17] MacAskill claims that, since b is *almost* as good as the best option for both theories, for any number n, $\boldsymbol{b}$ should be your safest and thus u-best option: '[$\boldsymbol{b}$] seems to represent the best hedge between the two views' (2014, 94), for any number n. As he points out, however, there is no pair of value-

[16]Note that these statements make sense: as mentioned on page 107, even on Comparativism, axiologies may be (intratheoretically) cardinal orderings.

[17]That is, $\boldsymbol{a}$(total utilitarianism, n people at wellbeing 100)=0.5, and so on.

functions G_T and G_A for the total and average utilitarian orderings for which this will be the case. To see this, let z be the outcome in which there is one person, at wellbeing 1, and let $G_T(z) = x$ and $G_A(z) = mx$. We then have

$$\begin{aligned} \boldsymbol{b} \succ_u \boldsymbol{a} \quad & \text{iff} \quad 1/2 \cdot 990nx + 1/2 \cdot 99mx > 1/2 \cdot 100nx + 1/2 \cdot 100mx \\ & \text{iff} \quad 990n + 99m > 100n + 100m; \end{aligned}$$

$$\begin{aligned} \boldsymbol{b} \succ_u \boldsymbol{c} \quad & \text{iff} \quad 1/2 \cdot 990nx + 1/2 \cdot 99mx > 1/2 \cdot 1000nx + 1/2 \cdot mx \\ & \text{iff} \quad 990n + 99m > 1000n + m. \end{aligned}$$

For $n = 1$, these inequalities are jointly satisfied only if $890 > m > 10/98$; for $n = 10'000$, they are jointly satisfied only if $8'900'000 > m > 100'000/98 \approx 1020$. So whatever value we choose for m (and x), it cannot be the case that $\boldsymbol{b}$ is your u-best option for any number n.

MacAskill concludes from this argument that 'it's unclear how any way of comparing [average] and [total utilitarianism] could be correct' (2014, 95), and hence that this is a case 'where choice-worthiness differences [i.e. value differences] seem to be incomparable between different theories' (2014, 97). Other people have drawn similar conclusions from this case.[18] If MacAskill and these people are right, total and average utilitarianism cannot be comparable.

However, I think that this is a dubious objection, and there is a debunking argument against the intuition it relies upon. It is easy to see why the alleged problem arises. Intuitively, we may start by considering options involving comparatively few people. We can choose our value-functions in such

[18] Cf. e.g. Hedden (forthcoming).

a way that both total and average utilitarianism matter significantly in determining the u-value of these options. If we now increase the differences in the number of people that our options affect while leaving their average wellbeing the same, this will not affect the value of our options according to average utilitarianism, but it will increase the stakes of our decision according to total utilitarianism. So if we increase these numbers sufficiently, total utilitarianism will swamp average utilitarianism: the u-best option will simply be that which is best according to total utilitarianism. If instead we consider options involving comparatively many people, and choose our value-functions in such a way that both theories matter significantly in determining the u-value of *these* options, then there will not only be options with *still more* people, for which total utilitarianism swamps average utilitarianism. But also, if we now consider options where the difference in the number of people affected is comparatively small, average utilitarianism will swamp total utilitarianism.

But this is exactly what we should expect, and what should happen if these theories are comparable. It is *clear* that one theory comes to dominate u-value if we increase the stakes for that theory while leaving them exactly the same for the other. *Hopefully* it does that. It is simply not true that $\boldsymbol{b}$ should represent the safest option for any number n. We may think that it does, because b is so close to the best option on both these *orderings*. But this fact about the orderings is insignificant once we think of these two axiologies as comparing somehow on an intertheoretic scale. What matters is not how close *in the ordering* b comes to being the best option ('99%'). What matters is the *absolute* difference between b and the best option on

the intertheoretic scale – as it were, the *size* of the 1%. And naturally, for total utilitarianism this will increase for increasing numbers of people. So we should not at all be tempted by the intuition that $\boldsymbol{b}$ should be your safest and thus u-best option for any number n. This intuition is fundamentally flawed.

Conclusion

Let me end this discussion here. I argued that even if theories themselves are merely orderings, there may be independent facts that determine how they enter the u-value relation, and I claimed that it is more plausible to assume (E) and the existence of at least *some* such facts, than to reject (E). I then argued that neither the arbitrariness nor the swamping worry are convincing objections against the assumption that in fact all theories are comparable.

3.3.2 Comparativism and EVM

Completeness and Contuinity

At the end of chapter 2 we were left with the question whether in light of the problem of intertheoretic comparisons it is plausible that $\succeq_U$ satisfies Completeness and Continuity. So it remains to ask about the status of these axioms under Comparativism. This status depends on which facts about comparisons would be most plausible. So let me finally say something about that question. My remarks shall be brief, and somewhat speculative. The discussion about the problem of intertheoretic comparisons is in the early stages, and no one has addressed the problem explicitly in light of the dis-

tinction between Absolutism and Comparativism. And to some extent, my formal results are open to very different views about what the most plausible comparison-facts are. But let me nonetheless suggest a rough view.

As far as I see, the Idea of Equal Say does indeed have at least a *prima facie* plausibility under Comparativism. To begin with, there may be reasons of simplicity and parsimony that favour a theory with some general principle over a more particularist theory of u-value. But moreover, if we really do think of comprehensive theories of value as being merely orderings, and if intertheoretic comparisons are given by independent facts about how these orderings enter the u-value relation, it seems plausible that they should in some sense (apart from their probability) weigh equally in the u-value relation. If a theory of the good says that the best world is the world with most beauty, say, and that is all it *can* claim, there does not seem to be any reason why (apart from its probability) it should weigh heavier or less heavily in the u-value relation than some other ordering. In other words, there does not seem to be any ground for a particularist view that treats theories differently. Or more positively, in treating axiologies differently, our very theory of axiological uncertainty would arguably be biased against some axiologies. It would not be neutral among them, as a theory of uncertainty arguably should be. So if there is some independent fact about how orderings enter the u-value relation, it seems *prima facie* plausible that this fact is some general principle, and that on this principle all orderings somehow get equal weight (apart from their probabilities). I think that the case is in some sense different under Absolutism, and I shall come back to this distinction in section 3.4.3. But at least under Comparativism, the rough Idea of Equal

Say seems plausible.

Suppose that this is true. Then, if all underlying axiologies are complete, it does in fact seem plausible under Comparativism that $\succeq_U$ is complete as well. If there is a general equal-say-principle, this principle plausibly determines a complete u-value relation – as indeed all the principles that have been suggested in the literature do. Of course, there could be an equal-say-principle that induces some equally balanced incommensurability in the theories, and yields a less than fully complete u-value ordering. And perhaps we cannot be altogether certain that all such incommensurability-inducing equal-say-principles are false. But it is hard to see why such a principle should hold. If only for reasons of elegance and simplicity, it seems that equal-say-principles that render $\succeq_U$ complete will be more plausible.

The Continuity condition is all the more plausible, at least as far as considerations of intertheoretic comparisons are concerned. As I said, this condition fails if some axiologies compare in a lexical way – if there are axiologies T_i and T_j such that any positive value difference between options according to T_i is greater than any positive value difference between options according to T_j. But given the Idea of Equal Say, that will not be the case. If one theory dominates another lexically, they have *anything but* equal weight. We might say that the former has infinitely more weight. Perhaps we can again not assign zero credence to the possibility of such intertheoretic lexical dominance. But under Comparativism it does seem very implausible.

Suppose that all of this is correct. Then, if Comparativism explains why *some* intertheoretic comparisons hold (thesis (E)), it is plausible that a large number of theories are comparable. Indeed, it seems plausible that all vNM-

conformable axiologies are comparable – that is, that all vNM-conformable axiologies combine in some way into a complete and continuous u-value relation. Thus I take it that, under Comparativism, and within the restricted scope of vNM-conformable axiologies, all the conditions of the Expected Value Theorem are plausible.

Axiomatising Specific Comparativist Accounts

Nonetheless, we can extend the Expected Value Theorem in some interesting ways. As it stands, and given the decision-theoretic explication of intertheoretic comparisons, the Expected Value Theorem does entail EVM; but it does not entail a *specific* form of EVM, a form of EVM with certain specific comparisons. So if we find a particular comparison principle most plausible, we might say that the theorem is not as informative as it could be. But we can of course add conditions to the theorem to make it imply a specific form of EVM.

For example, consider Best/Worst Normalisation again, the view that the best and worst conceivable outcomes are equally good and bad according to all theories. And let $\boldsymbol{a}_{(i,x)}$ in $\mathcal{Q}$ be the option that certainly leads to x while T_i is true: $\boldsymbol{a}_{(i,x)}(i,x)=1$. If we want to axiomatise EVM with Best/Worst Normalisation, we could assume the

> **Best/Worst Comparison Principle**: For all T_i and T_j, $\boldsymbol{a}_{(i,x_i^w)} \sim_U \boldsymbol{a}_{(j,x_j^w)}$ and $\boldsymbol{a}_{(i,x_i^b)} \sim_U \boldsymbol{a}_{(j,x_j^b)}$.

It is easy to see that if we add this principle to the Expected Value Theorem, given the decision-theoretic explications, the only value-functions that satisfy

our assumptions are such that $G_i(x_i^w) = G_j(x_j^w)$ and $G_i(x_i^b) = G_j(x_j^b)$ for all T_i and T_j.

Similarly, consider Paradigm Normalisation again, the view on which there are certain paradigmatically heinous and value-neutral outcomes, x^h and x^n, whose value is the same according to all theories. If we want to axiomatise EVM with Paradigm Normalisation, we could assume the

> **Paradigm Comparison Principle**: For all T_i and T_j, $\boldsymbol{a}_{(i,x^h)} \sim_U \boldsymbol{a}_{(j,x^h)}$ and $\boldsymbol{a}_{(i,x^n)} \sim_U \boldsymbol{a}_{(j,x^n)}$.

Similar principles will be formalisable for other specific comparativist accounts. If we add these principles to the Expected Value Theorem, the theorem will imply specific comparativist forms of EVM.

3.4 Absolutism

Let me now turn to Absolutism. Again, according to Absolutism, an axiology is not only an ordering of prospects, but also in itself makes claims about how good certain options are, or how much better some are than others, on an intertheoretic scale. I think that Absolutism can come in two variants. On the one hand, it might be that axiologies make claims about the heights of value levels and sizes of value differences on an intertheoretic scale in a way that is entirely independent of the u-value relation. I shall call this *Substantial Absolutism*. However, at least *prima facie*, it might also be that an axiology itself makes claims about the heights of value levels and sizes of

value differences on an intertheoretic scale, but in a way that is not ultimately independent of the u-value relation. I shall call this idea *Non-Substantial Absolutism*.

What I mean by this distinction will become clearer when I elaborate more on both of these versions. In section 3.4.1, I shall outline what I take to be the most promising version of Substantial Absolutism. In section 3.4.2, I shall elaborate on how that view is compatible with the Expected Value Theorem. And in section 3.4.3, I shall explore the idea of Non-Substantial Absolutism.

3.4.1 Fitting Strengths of Attitudes

Fitting Attitude Accounts

I think that the most plausible form of Substantial Absolutism is based on so-called fitting attitude accounts of value, or *FA-accounts*. FA-accounts are analyses of what it means that something has value, or that one thing has more value than another. According to these accounts, there is an attitude or set of attitudes such that the fact that x is better than y means that it is fitting to have these attitudes towards x and y; or similarly, there is an attitude or set of attitudes such that the fact that x has value means that it is fitting to have these attitudes towards x. So FA-accounts define value in terms of attitudes that it is fitting to have. Versions of the account can differ in the set of attitudes they consider relevant in defining value. If we take 'preference' to be the relevant (two-place) attitude, then the account

says: that x is better than y means that it is fitting to prefer x to y.[19] More plausibly perhaps, there may be a variety of relevant attitudes – including, say, some form of *pleasure* about the fact that x rather than y was brought about (if that is so), *hope* that x rather than y will be or was brought about (if you do not know it), *regret* that y rather than x was brought about (if that is so), and so on.[20]

At least some versions of FA-accounts can be understood as implying a form of Absolutism. Suppose that the attitudes should come in degrees, and that the intensity of the attitudes it is fitting to have correlates with the sizes of the relevant value differences or the heights of the relevant value levels. For example, suppose that if according to some theory, the value difference between x and y is larger than the value difference between z and t, it would, according to that theory, be fitting to feel more regret about the fact that y rather than x was brought about than about the fact that t rather than z was brought about (if these are facts). If that is so, we can take the strength of these attitudes to be the 'common scale' between different axiologies. And axiologies will imply facts about the sizes of value differences or heights of value levels on an intertheoretic scale because they imply facts about the strengths of the relevant attitudes. Consequently, the value difference between x and y, according to T_i, would be greater than the value difference between z and t, according to T_j, if the regret that it would be fitting to feel, if T_i was true and y rather than x was brought about, is stronger than the regret it would be fitting to feel if T_j was true and t rather

[19]Cf. Rabinowicz (2008) for such an account.

[20]For defences of FA-accounts, cf. e.g. Brentano (1889, 18), Broad (1930, 283), Ewing (1947, 152), Wiggins (1987, 206), Gibbard (1990, 51) and Scanlon (1998, ch.2).

than z was brought about. A similar account could be given for heights of value levels, involving some monadic attitude. For example, suppose the greater the value of x, the greater the pleasure it is fitting to feel about the fact that x was brought about (if that is a fact). Then the value of x according to T_i would be greater than the value of y according to T_j if the pleasure it would be fitting to feel if T_i was true and x was brought about is greater than the pleasure it would be fitting to feel if T_j was true and y was brought about. So a *strength-sensitive FA-account* would vindicate a form of Absolutism.[21] I say that it vindicates *Substantial* Absolutism, because if a strength-sensitive FA-account is true, there is something substantial, independent of the u-value relation, that determines value-facts on an intertheoretic scale – viz., facts about fitting attitudes.

It is also clear that the truth of such a strength-sensitive FA-account will give rise to the large number of axiologies that I suggested on page 108. One and the same axiological ordering can consistently be combined with different claims about the strength of attitudes that it is fitting to have with respect to various outcomes. For example, if x is better than y, then if y rather than x was brought about, it may be fitting to feel just a mild form of regret, or it may be fitting to feel a very strong and intense form of regret, or something in between. In principle, there are infinitely many strengths of regret that it may be fitting to feel. And the same is true for any other attitude. So if an axiology is a consistent combination of a betterness ordering and a claim about what attitudes it is fitting to have with what strength, then there are

[21] Such an account is endorsed by Ross (2006b); something similar is suggested in Sepielli (2010, 181ff.); MacAskill (2014) discusses but rejects it.

infinitely many axiologies corresponding to any betterness ordering. If you want to assign a particular probability to an axiology, intuitively, you not only have to decide what values there are, but also how important these values are – that is, what attitudes would be fitting.

Interestingly, if this fitting attitude account is true, we have an error theory for why people thought that any comparisons between two orderings like the total and the average utilitarian one are arbitrary. If a strength-sensitive FA-account is true, this is precisely what we should expect. There *is* no uniquely privileged or non-arbitrary way in which the total and the average utilitarian orderings should be compared, because there are infinitely many ways in which theories implying these orderings compare. Different versions of these orderings will compare in different ways, and no two versions may stand out as most plausible. Moreover, MacAskill's argument against the comparison between total and average utilitarianism will again be flawed: we should again not at all be tempted by the principle that he invokes for his sceptical conclusion. And of course, the fundamental objection against Comparativism that I discussed on page 112 does not even arise for Substantial Absolutism. So as far as the objections against Comparativism are concerned, a fitting attitude account promises to provide a plausible response to the problem of intertheoretic comparisons.

Objections

However, this solution to the problem of intertheoretic comparisons will also face objections. Some of these objections apply to FA-accounts in general.

One of them is the *Wrong Kind of Reasons Problem.* For any candidate attitude, there are cases in which you ought to have that attitude towards an object (or pair of objects) even though that object is not valuable (or the value relation does not hold). For example, if some disaster occurs unless you are pleased about a cup of mud, you arguably ought to be pleased about that cup of mud – but that does not make it valuable.[22] So proponents of an FA-account need some story about when a reason to have the relevant attitudes is of the right kind to define value. A second prominent objection to FA-accounts is the *Circularity Objection.* The objection is that it is not possible to find a relevant set of attitudes that are indeed fitting to have towards what is valuable, and that are not itself evaluative judgments (which would render the account circular).[23] A third worry is that normative ethics and value theory should primarily be concerned with what we ought to *do*, or perhaps what kind of people we ought to be, but not with what *attitudes* it would be fitting to have – when it would be fitting to feel guilty or pleased, saddened, disappointed or regretful. If those are matters of ethics, the argument goes, they are surely marginal. And it seems that on FA-accounts, axiologies are ultimately theories about *that.*[24]

Moreover, there will be problems that arise specifically – or at least most urgently – for *strength-sensitive* FA-accounts. One such problem is to motivate the idea that there indeed *are* truths about which strength of attitudes are fitting.[25] We might call this the *Truth Problem.* One might be sceptical

[22]This example is due to Crisp (2000, 459); cf. also D'Arms and Jacobson (2000).
[23]Cf. Bykvist (2009a) for this objection.
[24]John Broome raised this objection in conversation.
[25]I thank Bastian Stern for pointing this out to me.

about that, even if one generally accepts the existence of moral facts (in some sense). For example, one might believe that it is a fact that pain is bad, and perhaps even that a fitting response to pain is some negative attitude, but doubt whether there are facts about how strong that negative attitude would fittingly be. True, we may have intuitions about what strength of attitudes are fitting: we may find some people oversensitive, and others rather too apathetic. But perhaps there is a debunking argument against these intuitions: if for some reason, all of mankind had always been more sensitive or more apathetic than we are, we would presumably have different intuitions. Perhaps the same debunking argument could not be given against the claim that pain is bad, and perhaps not even against the claim that a fitting response to the existence of pain is some negative attitude. So even if one finds FA-accounts plausible, one might be sceptical about the relevant versions required for our purposes.

I do not think that all these objections are sound, or devastating.[26] But this is not the place to address them. FA-accounts are popular enough, and they remain a promising candidate for grounding Substantial Absolutism. So let me investigate how such an account could be used to vindicate EVM.

3.4.2 Substantial Absolutism and EVM

Completeness and Continuity

Suppose a strength-sensitive FA-account is true, and explains the truth of

[26]I think that Schroeder (2010) has provided a convincing response to the Wrong Kind of Reasons Problem. Moreover, as far as I see, the Circularity Objection raised in Bykvist (2009a) simply reduces to an instance of the Wrong Kind of Reasons problem, and is also answerable with Schroeder's account.

the Minimal Thesis (E). Obviously, in that case too positive facts about intertheoretic comparisons hold among a large number of axiologies. But what exactly is the status of Completeness and Continuity? Unfortunately, strength-sensitive FA-accounts are very underexplored.[27] And the general picture that emerges from such an account might depend on its details – e.g., the precise way in which strengths of attitudes are cashed out, or the way in which incommensurability is understood. So again, much of the following will have to be somewhat speculative. But let me offer some general considerations that I think should be true on any plausible strength-sensitive FA-account.

To see whether Completeness and Continuity are plausible, we need to ask again how many, or what kind of axiologies there exist under such an account. And I think we can answer that question by considering the set of *pluralist* axiologies on which there are two different sorts of value. If an FA-account is true, these pluralist axiologies must imply facts about what attitudes are fitting towards these two values. And for each such pluralist axiology, there will be two non-pluralist axiologies that each accept only one of the two values that the pluralist axiology accepts, and imply the same fitting attitudes with respect to these values as the pluralist axiology does. If fitting attitudes reflect value comparisons, then these axiologies should compare in the same way in which the two separate value-orderings compare according to the pluralist axiology. So we can learn about how axiologies can compare under this form of Substantial Absolutism by looking at what

[27]Though cf. e.g. Rabinowicz (2008; 2012).

possible pluralist axiologies there are.[28]

So what pluralist axiologies are there? Presumably, for any two values, there should be *complete* pluralist axiologies according to which these two values are fully comparable. That is, a strength-sensitive FA-account should allow for the existence of such theories. If it did not, our very account of value would have the substantial implication that there *must* be value incommensurability (or that there cannot be pluralist axiologies), which would be very unfortunate. Consequently, it seems that for any two orderings there will be axiologies implying these orderings that are fully comparable. That will be axiologies that imply the same attitudes with respect to these values as the respective complete pluralist axiologies do. With respect to these axiologies, the u-value relation is indeed complete.

However, there should also be non-complete pluralist axiologies on which the two values are less than fully comparable. By the same reasoning as above, for any two orderings, there will thus be versions of these orderings that are less than fully comparable. With respect to these axiologies, the u-value relation is not complete. In fact, presumably, there should also be non-complete pluralist axiologies on which the two values are radically incom-

[28]I do not mean to suggest that there is a principled distinction between pluralist and non-pluralist axiologies. Perhaps almost any non-uniform axiology could be understood as pluralist, accepting the separate values of x, of y, of z, and so on. This only makes the above argument more general.

Nonetheless, there are axiologies that may not naturally be combinable into a pluralist axiology. For example, if ordering $\succeq_i$ is a sort of extension of $\succeq_j$ that accepts *more* valuable outcomes (in the sense in which standard utilitarianism is an extension of anthropocentric utilitarianism), there may be no natural *third* axiology which accepts 'both' values. However, in that case, we can simply consider bridging-orderings $\succeq_k$ with some third value (like beauty, say) and apply the above reasoning. We can look at how pluralist axiologies can combine the orderings $\succeq_i$ and $\succeq_k$, as well as the orderings $\succeq_k$ and $\succeq_j$, and thus learn how theories implying the orderings $\succeq_i$ and $\succeq_j$ can compare. For convenience, I consider only the simple case above.

parable (like the art/philosophy theory I considered in section 3.2.1). Those axiologies may be less plausible, but it should be possible to state them consistently, and thus they should exist. Consequently, for any two orderings, there will also be versions of these orderings that are fully incomparable. With respect to these axiologies, the u-value relation is radically incomplete. So if all of this is true, then under a strength-sensitive FA-account, it is false that $\succeq_U$ satisfies Completeness. It satisfies Completeness only with respect to the restricted class of fully comparable axiologies.

This would not be a practically significant restriction for the Expected Value Theorem if it was plausible that we should have credence only in axiologies that are fully comparable. So this raises the question which *versions* of each ordering we should have credence in. One answer would be that there is a general principle about what the most plausible versions of each ordering are. In fact, we could again invoke something like the Idea of Equal Say, or something like Best/Worst Normalisation. For example, we *could* claim that if you find T_1 the most plausible version of total utilitarianism, you should have credence only in versions of the other orderings on which the best and worst conceivable outcomes are equally as good and bad as they are on T_1.

But I think that this is implausible. The reasons I gave for the Idea of Equal Say under Comparativism do not support an equivalent principle about what the most plausible axiologies are corresponding to each ordering. The present question is a matter of first-order axiology. And if we give a particular axiology only little *credence*, that is not (necessarily) to be biased against that axiology or against its ordering, let alone does it make our theory of u-value biased. And neither does our theory of u-value become less

simple or parsimonious if we suppose that there is no such principle. Under Absolutism, our theory of u-value is simply EVM, and does not include an additional, independent weighing-principle like under Comparativism.

So if a strength-sensitive FA-account is true, I take it, it is implausible that there is such a general equal-say-principle about what our credences should be. Instead, a more particularist view will be plausible. For example, it will seem reasonable to believe that compared to the most plausible version of anthropocentric utilitarianism, the most plausible version of total utilitarianism simply posits *more* value in the world. That is, it will seem reasonable to believe that if animal wellbeing in *addition* to human wellbeing had value, that would not change the importance of human welfare (i.e., the attitudes that would be fitting towards it). And it will seem less reasonable to believe that if animal wellbeing had value, that would lessen the value of human welfare precisely to the extent so that the best and worst conceivable outcomes become equally good and bad as on anthropocentric utilitarianism. The question about which versions of each ordering we should have credence in seems to become similar to the question about how these orderings would most plausibly combine into a pluralist axiology.

Suppose that this is true. Then it does seem that we should have considerable credence in axiologies that are less than fully comparable – just as non-complete pluralist axiologies are plausible. As I mentioned on page 117, if both beauty and wellbeing have value, it seems plausible that these values are less than fully comparable. Similarly, I take it, the most plausible versions of the orderings on which only beauty and only wellbeing have value need not be fully comparable either. So the Completeness assumption of the

Expected Value Theorem is arguably a very significant restriction. At least under Absolutism, is important to explore axiomatisations of the u-value relation without the Completeness axiom. That will be the task for chapter 5.

Unfortunately, the theorem I shall introduce in chapter 5 does not cover theories that are *fully* incomparable. In a very restricted sense, it will still feature something like a Completeness condition. To that extent, even the scope of the theorem in chapter 5 will be restricted. It will not apply to theories that are fully incomparable, even though under Absolutism, such theories seem to exist. However, I take it that even if fully incomparable axiologies *exist*, they are an extreme case, and comparatively very implausible – just like pluralist axiologies that posit *radical* incommensurability. In this sense, the restriction of the theorem in chapter 5 will at least be less severe than that of the Expected Value Theorem.

How about the Continuity condition? We can again apply the same reasoning as above. Presumably, for any two values, there should be possible pluralist axiologies on which one of the two values dominates the other lexically. Accordingly, for any two orderings, there will be versions of these orderings that compare lexically. With respect to these axiologies, the u-value relation is not continuous.

Unfortunately, as with fully incomparable axiologies, the theorem in chapter 5 will not apply to theories that compare in a lexical way. It will still feature a Continuity condition. So this is another respect in which even that theorem's scope will be restricted. However, axiologies that lexically dominate others or are lexically dominated by them also seem to be an extreme

case, and comparatively very implausible – like pluralist axiologies that posit lexical dominance among values. So in this respect too, the restriction of the theorem in chapter 5 will be less severe than that imposed by Completeness.[29]

Fitting Attitudes and Cardinal Value

If these considerations were sound, then under an FA-account, there *are* axiologies that are fully comparable, and for which the Completeness condition of the Expected Value Theorem will hold. Let me thus elaborate at this point on how EVM, as restricted to these axiologies, could be axiomatised. This will then also suggest how the result from chapter 5 could be applied to an FA-account.

As I emphasised in section 2.1, EVM presupposes a *cardinal* concept of intertheoretic comparisons. In order to explain what EVM means, we have to know what it means that the value difference between x and y, according to T_i, is n times as great as the value difference between z and t, according to T_j. So if something along the lines of an FA-account should explain our concept of value, we would have to ensure that it provides a cardinal struc-

[29]On the other hand, note that implausible and extreme as fully incomparable and lexically comparing theories are, they present interesting philosophical problems. *Prima facie*, it seems that we should not have *zero* credence in such theories. And *prima facie*, it seems that if we have *some* nonzero credence in such theories, they should affect the u-value relation very radically: if a theory is fully incomparable to all our other axiologies, and has nonzero probability, that should lead to massive and widespread incompleteness in the u-value relation; and if a theory lexically dominates all our other axiologies, and has nonzero probability, it should effectively swamp the u-value relation completely. But *prima facie*, it seems implausible that the u-value relation is so very incomplete, or dominated by one theory. Similar problems arguably arise under Comparativism, once we take into account uncertainty about the true facts about comparisons. It is not clear to me what the solutions to these problems are. They resemble other paradoxes in decision theory that arise, roughly speaking, due to infinity or unbounded utility functions. In this thesis, I ignore all these problems; I ignore the extreme axiologies that would give rise to them. (Cf. e.g. MacAskill (2013) for a discussion.)

ture. We arguably have no unmediated understanding of what it means that one attitude is, say, 3, or 3.5, or 5 times as strong as another.

Whether or how it is possible to supply an FA-account with cardinality depends on precisely what kind of FA-account would be true (if any). A first suggestion might be that there is some quantitative *empirical* state that allows for a cardinal concept of the strength of attitudes. For example, perhaps there is a particular region in the brain where neurons generally fire when we feel regret; and perhaps if we feel more regret then more such regret-neurons fire. If so, we could assume that axiologies imply claims about precisely how many regret-neuron firings would be fitting in different cases; and we could explicate intra- and intertheoretic comparisons in terms of that. For example, that the value difference between x and y, according to T_i, is twice as great as the value difference between z and t, according to T_j could mean, say, that the number of regret-neuron firings that would be fitting if T_i was true and y rather than x was brought about is twice as great as the number of regret-neuron firings that would be fitting if T_j was true and t rather than z was brought about.

However, I am sceptical about that. For one thing, emotions like regret are extremely complicated patterns. Being deeply and intensely regretful does not mean having a one-time flush of some bodily feeling. If regret is to be distinguished from a mere sad mood, for example, it should arguably include certain cognitive attitudes – say, beliefs of the form 'this was bad', or something of the sort. Furthermore, regret arguably also includes certain behavioural dispositions, and these can come in extremely diverse forms – e.g., as attempts to rectify things, to apologise, to comfort, and so on. And

even if there is a feeling that accompanies regret, being more deeply regretful about an event need not mean having a more intense episode of that feeling. It might mean revisiting that event more frequently in one's mind, revisiting it for a longer period, having the relevant behavioural dispositions more strongly, and so on. Or more precisely still: different people regret very differently. In some, regret will manifest in their active engagement with the event; in others, it will manifest in their evading any thought and memory of it. And similar points seem to apply to all plausible candidate attitudes. Hence I doubt whether it is possible to find a quantitative measure – such as neuron firings – that captures our intuitive understanding of their intensity even roughly.

For another thing, even if that were possible, the above-mentioned Truth Problem would arguably be all the more serious if we had to assume that axiologies imply facts about (say) the numbers of neuron firings that are fitting. Even if one generally accepts the existence of moral facts, it seems difficult to believe that there could be truths about *that*. So I am sceptical about explicating comparisons by means of some quantitative empirical state.

A second suggestion would be to explicate cardinal intra- and intertheoretic comparisons by means of an intuitive *comparative* notion of the strength of attitudes. While we have no unmediated understanding of what it means that one attitude is 3, or 5 times as strong as another, we arguably do have at least *some* understanding of what it means that one attitude is *stronger* than another. For example, it seems clearly true that if you regret some event by weeping and mourning for ten days then your regret is stronger than if you regret that event by briefly thinking 'what a pity'. So while our under-

standing of attitudes will not directly give us a cardinal concept of value, it might directly give us a comparative concept of value differences – that is, a ranking of value differences according to their size. And as I mentioned in section 2.4.3, under *some* conditions difference comparability among a set X is enough to yield cardinal measurability. So if these conditions hold, then the truth of a relevant FA-account, paired with a relevant comparative notion of the strength of attitudes, will allow us to explicate cardinal intratheoretic comparisons without the help of decision theory. And by the same token, they would allow us to explicate cardinal intertheoretic comparisons, without the help of decision theory.

However, it is a large and difficult question whether that is indeed possible. On the one hand, the formal question about the conditions under which difference comparability is sufficient for cardinal measurability seems very underexplored; as far as I see, the question of when the two are equivalent is still open.[30] On the other hand, and more importantly, it is not clear whether we really do have a comparative understanding of the strength of attitudes that is rich enough to meet the relevant conditions.

To see the difficulties that will arise here, consider a condition introduced by Kaushik Basu. Basu (1983, 197) proved that if an ordering of a set X of outcomes can be represented by a utility function $u : X \to \mathbb{R}$ such that the image $u(X)$ of X under u is dense in a connected subset of $\mathbb{R}$, then the ability to compare all differences of utility is equivalent to cardinality. In our context, this formal condition will only help to construct a cardinal intertheoretic concept of value if our comparative notion of the strength of

[30] As of 2005, this is stated in Bossert et al. (2005, 35).

attitudes is fine enough, and if axiologies indeed imply the relevant truths, to allow for difference comparability in a space of alternatives with utilities that are dense in a connected subset of $\mathbb{R}$. And it is not at all clear whether that is the case. To compare utility differences in $\mathbb{R}$ (or a set that is dense in a connected subset of it), I take it, our notion of the strength of attitudes would itself have to involve some variable that varies on a real-valued scale. For example, we would have to measure the strength of attitudes by the length of *time* for which they are felt. However, time is not the only variable that determines strength of attitudes. For example, a period of deep mourning can intuitively be a stronger form of regret than a slightly longer period of the comparatively unaffected thought 'what a pity'. So we would need a second variable for the intensity of regret. But as I just argued, it is doubtful whether we have one – whether, say, we could measure intensity by the numbers of neuron firings. And again, the richer the required truths about attitudes become, the more dubious it arguably is that there could be truths of that kind.[31]

It might be a valuable research project to explore the possible details of strength-sensitive FA-accounts. But this project is beyond the scope of this thesis. So I shall not rely on the assumption that we can explicate a cardinal intertheoretic concept of value via fitting attitudes alone.

[31]One might argue that this is only a difficulty about *precise* comparisons, that we certainly do have an understanding of rougher differences in attitude strength, and that with respect to them, we should at least get rough cardinal comparisons. This would require a formal theorem providing a cardinal structure from an incomplete ordering of differences, and I am not aware of such a theorem. But more importantly, as I mentioned on page 133, our general account of value should allow for the possibility that all options and outcomes are fully comparable, and not have the substantial implication that there *must* be incommensurability.

However, it is worth mentioning a third proposal. Suppose that an FA-account is true, and that axiologies imply certain facts about fitting strengths of attitudes, but that they do so only in a very rough manner, which does not directly provide a cardinal notion of strength, nor a relevant comparative notion of strength rich enough to imply cardinality. For example, suppose axiologies imply facts of the form 'if y rather than x was brought about, it is fitting to feel a mild form of regret', or a 'strong form of regret', or a 'very strong form of regret'. In that case, it will very often be indeterminate whether a given instance of regret is fitting or not. But now suppose that our axiologies nevertheless enter the u-value relation in a specific way, and that this u-value relation satisfies the conditions of the Expected Value Theorem. As far as I see, if the rough facts about attitudes are consistent with the decision-theoretic explication – in a manner I shall spell out presently – we could then explicate cardinal intertheoretic comparisons by means of the decision-theoretic explication, even if an FA-account is true. We might then say that it is facts about attitudes that *explain* why the u-value relation is not radically incomplete, or that ultimately 'ground' intertheoretic comparisons; but these comparisons nonetheless acquire their *cardinal* significance in the context of weighing axiologies under uncertainty. We could use the decision-theoretic explication as a kind of sharpening of the notion of value implied by our intuitive notion of attitudes alone. Given the above-mentioned difficulties of any richer FA-account, this might ultimately be the most promising proposal.

In any case, if a fitting attitude account is true, and it is fitting attitudes that ultimately ground intertheoretic comparisons, then the conditions of the Expected Value Theorem are not enough to imply EVM. Whatever precisely the structure and richness of the entailed attitudes: for all that the Expected Value Theorem says, the facts about attitudes might not be consistent with EVM. For example, suppose that the u-value relation is such that, according to the decision-theoretic explication, the value difference between x and y, according to T_i, is smaller than the value difference between z and t, according to T_j, but that, if T_i is correct and y rather than x was brought about, it is fitting to weep and mourn for ten days, and if T_j is correct and t rather than z was brought about, it would be fitting to think briefly that that is a pity. If it is fitting attitudes that ultimately explain why theories imply facts about an intertheoretic scale, then that would would mean that EVM is false and some form of *weighted* EVM is true (in which T_j is given more weight than T_i). So if an FA-account is true, and we want to ensure on an axiomatic basis that EVM rather than some form of *weighted* EVM is true, we have to add additional conditions to the Expected Value Theorem.

Whatever precisely the structure and richness of the entailed attitudes are, the following two principles will do. Let $\boldsymbol{a}_{0.5(i,x)+0.5(j,y)}$ in $\mathcal{Q}$ be the option that leads with equal probability to x while T_i is true, or to y while T_j is

true:

$$\boldsymbol{a}_{0.5(i,x)+0.5(j,y)}(i,x) = 0.5,$$
$$\boldsymbol{a}_{0.5(i,x)+0.5(j,y)}(j,y) = 0.5. \tag{3.2}$$

Let *favouring attitudes for x over y* be attitudes that are fitting if x is better than y (such as hoping that x rather than y will be brought about). And let *favouring attitudes towards x* be attitudes that are fitting if x is good (such as taking pleasure in x), and *disfavouring attitudes towards x* be attitudes that are fitting if x is bad (such as being sad about x). To secure consistency with respect to unit comparisons, we could then state the

> **Dyadic Attitude Principle**: Suppose that $a_x \succ_i a_y$ and $a_z \succ_j a_t$. If the favouring attitudes it is fitting to have for x over y if T_i is true are stronger than the favouring attitudes it is fitting to have for z over t if T_j is true, then $\boldsymbol{a}_{0.5(i,x)+0.5(j,t)} \succ_U \boldsymbol{a}_{0.5(i,y)+0.5(j,z)}$; if they are equally strong, then $\boldsymbol{a}_{0.5(i,x)+0.5(j,t)} \sim_U \boldsymbol{a}_{0.5(i,y)+0.5(j,z)}$.

Similarly, to secure consistency with respect to level comparisons, we could state the

> **Monadic Attitude Principle**: If the favouring attitudes it is fitting to have towards x if T_i is true are stronger than the favouring attitudes it is fitting to have towards y if T_j is true, or if the respective disfavouring attitudes are weaker, then $\boldsymbol{a}_{(i,x)} \succ_U \boldsymbol{a}_{(j,y)}$; if they are equally strong, then $\boldsymbol{a}_{(i,x)} \sim_U \boldsymbol{a}_{(j,y)}$.

We could add these two principles to the Expected Value Theorem. If we do, and if the conditions of this extended theorem are true, facts about attitudes cannot contradict EVM, whatever precisely their structure and richness.[32]

3.4.3 Non-Substantial Absolutism

In the previous two sections, I have explored what I called *Substantial* Absolutism. FA-accounts give rise to Substantial Absolutism, because if a strength-sensitive FA-account is true, there is something substantial, independent of the u-value relation that determines value-facts on an intertheoretic scale – viz., facts about fitting attitudes. So as a final general proposal, let me now elaborate on the idea of a non-substantial absolutist account.

The Idea of Non-Substantial Absolutism

Very roughly, the idea of Non-Substantial Absolutism is that everything is like Substantial Absolutism says, except that there is nothing 'substantial',

[32]To see this, suppose we have a comparative notion of the strength of attitudes and a set of outcomes that are rich enough so that our ability to make difference comparisons implies unique crosscutting cardinal intertheoretic comparisons: that is, the relevant facts about attitudes concerning outcomes imply a theory-dependent utility function $u(\cdot, \cdot)$, unique up to positive affine transformation. As we know from the Expected Value Theorem, its conditions will also imply a theory-dependent utility function $v(\cdot, \cdot)$, unique up to positive affine transformation. Moreover, note that given the above principles, v must respect all the unit and level comparisons that our facts about attitudes imply. So in effect, the u-value relation between the options $\boldsymbol{a}_{0.5(i,x)+0.5(j,t)}$ and $\boldsymbol{a}_{0.5(i,y)+0.5(j,z)}$, and $\boldsymbol{a}_{(i,x)}$ and $\boldsymbol{a}_{(j,y)}$, would be enough to determine v up to positive affine transformation, just as our facts about attitudes are enough to determine u. And since the relevant difference and level comparisons will be the same, v must be a positive affine transformation of u.

Suppose instead that we do *not* have a comparative notion of the strength of attitudes rich enough to yield cardinal measurability. Then again, given the Attitude Principles, the utility function v implied by the conditions of Expected Value Theorem must respect all the difference and level comparisons that our facts about attitudes imply. So we could use that function as a kind of sharpening of the notion of value implied by our intuitive notion of attitudes.

independent of the u-value relation, in which facts about the heights of value levels or sizes of value differences are manifested. Intuitively, we replicate all the claims of Substantial Absolutism, but without its extra substantial baggage. So on Non-Substantial Absolutism – as on Substantial Absolutism – an axiology not only says which prospects are better than which, but also how good these prospects are, or how much better they are than others, on an intertheoretic scale. Consequently, there are again infinitely many theories corresponding to each ordering. For example, intuitively, there is the view that only beauty has value, and that this value is important (on an intertheoretic scale); and there is the different view that only beauty has value, but that its value is comparatively insignificant (on an intertheoretic scale). These axiologies weigh differently in the u-value relation. And indeed, on Non-Substantial Absolutism, there are no other facts in which these facts about the heights of value level and sizes of value differences would be manifested, independently of the u-value relation. That the value of beauty is important on one such axiology and unimportant on another, does not imply, say, that certain attitudes are fitting according to one but not the other. It only implies that these axiologies weigh differently in determining u-value.

This raises the question whether, understood non-substantively, Absolutism collapses into Comparativism. As I introduced it, the distinction between these two views is that under Comparativism, axiologies are merely orderings and the facts about their comparisons are independent facts; under Absolutism, axiologies themselves make claims about how they compare, and there are infinitely many axiologies corresponding to each ordering.

However, suppose Comparativism is true. Then, in principle, there are

infinitely many possible ways in which any of our axiological orderings might compare to any other, just as there are infinitely many versions of each axiological ordering if Absolutism is true. And even if one way of comparing two axiologies is in some sense the true one, we might (or indeed should) be uncertain about which that is. So once we take this kind of uncertainty into account, then under Comparativism, there is ultimately not only the question about which ordering is true, but also the question about how our orderings weigh against each other to determine u-value, and possible uncertainty about both. Moreover, note that if Non-Substantial Absolutism is true, then there is ultimately no meaningful question *outside* the context of axiological uncertainty that determines which version of an ordering is true. In that sense, under Non-Substantial Absolutism too there is ultimately only the question about which ordering is true, and the question about how these orderings weigh against each other to determine u-value.

So in principle, once we take into account uncertainty about the true comparisons under Comparativism, everything that can be expressed in non-substantial absolutist terms can also be expressed in comparativist terms, and vice versa. For example, the (comparativist) view that axiological orderings enter the u-value relation by satisfying the Best/Worst Comparison Principle could be expressed as the (non-substantial absolutist) view that if you find T_1 the most plausible version of total utilitarianism, you should have credence only in versions of the other orderings on which the best and worst conceivable outcomes are equally good and bad as they are on T_1. Similarly, the (non-substantial absolutist) view that if you find T_1 the most plausible version of total utilitarianism, you should find that version of anthropocentric

utilitarianism most plausible on which the value of human wellbeing is the same as on T_1, could be expressed as the (comparativist) view that the total and anthropocentric utilitarian orderings enter the u-value relation in such a way that human wellbeing is given the same weight for both orderings.

However, if my previous considerations were sound, there does remain a distinction. The comparativist and the absolutist way of thinking will have different normative implications. As I mentioned in section 3.3.2, if we really do think of comprehensive theories of value as being merely orderings, the Idea of Equal Say seems plausible. If a comprehensive theory of the good only says which options are better than which, and there is some independent fact about how orderings enter the u-value relation, it seems plausible that this fact is some general principle, on which (apart from their probabilities) all orderings somehow get equal weight. But as I suggested in the last section, if we think of a comprehensive theory of value as having to say something about how important that value is, the idea of giving each ordering 'equal say' in our credence distribution according to some general principle will become less plausible. A more particularist view will seem more plausible. It will seem reasonable to believe that if animal wellbeing in *addition* to human wellbeing had value, that would not change the importance of human welfare. And it will seem less reasonable to believe, say, that if animal wellbeing had value, that would lessen the value of human welfare precisely to the extent so that the best and worst conceivable outcomes become equally good and bad as on anthropocentric utilitarianism.

I think that this distinction remains intact even under Non-Substantial Absolutism, and hence that there remains a distinction even between Non-

Substantial Absolutism and Comparativism. Even though, in principle, everything that can be expressed in non-substantial absolutist terms can also be expressed in comparativist terms and vice versa, the two pictures seem to have different normative implications.[33]

By the same token, I take it that the status of Completeness and Continuity under Non-Substantial Absolutism is the same as their status under Substantial Absolutism. Even under Non-Substantial Absolutism, there will be theories that are less than fully comparable – theories that do not enter a complete u-value relation. And indeed, even under Non-Substantial Absolutism, there should be possible axiologies that are fully incomparable, and others that compare in a lexical way. The latter two sorts of axiologies will again be very extreme, and comparatively implausible. But less than fully comparable axiologies will not be implausible. So it remains important to explore axiomatisations of the u-value relation without the Completeness axiom, even under Non-Substantial Absolutism. Fortunately, however, with regards to the theories that are fully comparable, the Expected Value Theorem provides a perfect axiomatisation of Non-Substantial Absolutism. We cannot make the theorem more informative (as with comparativist accounts), nor do we have to insure consistency with independent facts about comparisons (as with substantial absolutist accounts).

[33]Of course, we could also simply introduce a new conceptual distinction, and distinguish the (normative) view that there is a general Equal-Say-principle by definition from the (normative) view that there is no such principle. However, it seems to me that the disagreement between those views is not simply a normative disagreement, perhaps like that between Best/Worst Normalisation and Variance Normaliation; instead, it seems that the Idea of Equal Say and the particularist view stem from fundamentally different ways of thinking about intertheoretic comparisons and axiologies. I have tried to capture this more fundamental distinction, which, I think, gives rise to these two different normative views. So I shall stick to this terminology.

The Plausibility of Non-Substantial Absolutism

Clearly, we can defend Non-Substantial Absolutism against MacAskill's worry and the arbitrariness objection by the same arguments that I gave in section 3.3. To that extent, it does not have a disadvantage over Comparativism.

Personally, I find Absolutism much more plausible than Comparativism. I think the idea that the most plausible version of total utilitarianism posits *more* value in the world than the most plausible version of anthropocentric utilitarianism, with the value of human welfare being the same on both theories, is very compelling. In fact, I think the Idea of Equal Say that seems plausible *conditional* on Comparativism has very unfortunate implications. Take total utilitarianism (T_{TU}) and ethical egoism, the agent-relative view, roughly, that (relative to me) a world is better than another if I am better off in it (T_E).[34] And consider the options of either inflicting a significant pain on me, or inflicting the same pain on 100'000 people. Note that, compared to the total number of sentient beings, the wellbeing of these 100'000 people is *vanishingly* insignificant according to T_{TU}; but according to T_E, my wellbeing is *all* that matters. So under the Idea of Equal Say, my wellbeing (on T_E) matters *enormously*, as compared to the wellbeing of these people (on T_{TU}). Presumably, on any standard interpretation of the Idea of Equal Say, even if the probability of T_{TU} is 100'000, or a million times greater than that of T_E, it would be u-better to torture these people, if we take into account only these two views. I find this implausible. And although we *could* also

[34]I thank William MacAskill for suggesting to use the example of ethical egoism in this argument.

accept Comparativism and a particularist view about the comparison-facts to avoid this conclusion, it seems that this will be less plausible than simply accepting Absolutism.

The obvious advantage of Non-Substantial Absolutism over Substantial Absolutism is that it does not suffer from the problems arising for any specific version of Substantial Absolutism, such as the problems of the relevant FA-accounts. So although I take both Comparativism and Substantial Absolutism to be more plausible than the denial of the Minimal Thesis (E), I think Non-Substantial Absolutism may ultimately be the most convincing solution to the problem of intertheoretic comparisons. For this reason, in the remainder of this thesis, I shall assume Non-Substantial Absolutism unless otherwise indicated.

3.5 Further Explorations: Social Choice Theory

As I mentioned in section 3.2, some people believe that the u-value relation may be complete even if not all axiologies are fully comparable. For example, some people have endorsed My Favourite Theory on the basis of scepticism about intertheoretic comparisons. I have argued that if intertheoretic comparisons are impossible, the u-value relation should be radically incomplete. But the inference from scepticism to My Favourite Theory seems popular; and clearly, I have not provided a knock-down argument against it. So as a

further exploration beyond the theory I have been outlining, it is worth investigating their alternative view in some detail. Doing so will be something of a digression from my main argument. But it will shed light on possible relationships between claims about intertheoretic comparability on the one hand, and claims about the u-value relation on the other. And so I think it is worth pursuing.

This is particularly so because we can use an existing formal framework to do this on an axiomatic basis. In this case, the framework is not from decision theory but from social choice theory. As I mentioned on page 99, there is a close analogue in social choice theory to the problem of intertheoretic comparisons – viz., the question whether, or to what extent, the wellbeing enjoyed by one person can be compared to that enjoyed by another. And there are many results about how the social preference relation depends on the measurability and comparability of wellbeing. So it is these results that I shall use to explore the relationship between different theories of axiological uncertainty and different accounts of intertheoretic comparability. Section 3.5.1 will set up the formal framework. Section 3.5.2 will then provide axiomatisations of My Favourite Theory and of the view that an option is u-better than another if it has the greater weighted value.

3.5.1 The Formal Framework

One standard framework for exploring measurability and comparability constraints in social choice theory is this. There is a set of individuals $\{1, ..., n\}$ and a set of options $\mathcal{X}$. It is assumed that each individual i has a real-valued

welfare function $W_i(\cdot)$, defined on $\mathcal{X}$, representing the individual's preference ordering over $\mathcal{X}$. An n-tuple $\{W_i\}$ of personal welfare functions is called a *profile*. A social welfare functional F is a function from the set of profiles to the complete orderings on $\mathcal{X}$. It specifies exactly one social preference ordering R for a given profile, or n-tuple $\{W_i\}$:

$$R_W = F(\{W_i\}). \tag{3.3}$$

The basic question is about the form of the functional F.

In the obvious analogy with our context, individual welfare functions correspond to value-functions of axiologies, and the social preference ordering corresponds to the u-value ordering. But there is also an important disanalogy. In social choice theory, it is most often assumed that an individual i may have a number of *different* preference orderings. This is to assume that the relevant functional is defined for multiple different profiles; so this general approach is called the *multi-profile approach*. Indeed, the results I shall employ assume that the domain of a social welfare functional is *universal* – that the functional yields an ordering R for any possible profile. Intuitively, this means that the functional is defined for *any* welfare function that any individual i may have.

This interpretation makes conceptual sense because we can individuate people independently of their preference ordering or welfare function. One and the same individual may have different preference orderings, so we can take the indexes i as referring to specific individuals. However, the same is not true in our context. We cannot let an index i refer to a particular

axiology, and assume that u-value relations are defined for *any* value ordering or function that *this* very same axiology may imply. The reason is that one and the same theory cannot imply different value-orderings. So we cannot interpret the mathematical indexes i as referring to specific axiologies.

However, I think there is a different interpretation for this framework. Instead of assuming that an index i refers to a specific theory, I shall assume that it corresponds to a fixed *probability*. Just like each individual can have different preference orderings, each probability can naturally be associated with different value orderings. So this interpretation will make conceptual sense. It means that our results will not concern a fixed, finite set of theories, but instead a fixed, finite set of probabilities. In the first instance, they will thus be results about one given probability distribution. But since the specific underlying probability distribution will be irrelevant in our results, they then generalise to results about any arbitrary probability distribution over any arbitrary (finite) set of axiologies.[35]

To formalise these results, let $\mathcal{X}$ be a finite set of options, and P a finite set of probabilities, labelled $1, ..., n$. I shall assume that $n \geq 2$, and that there are at least three options, $|\mathcal{X}| \geq 3$. For each i, axiologies with the probability

[35]It might be worth mentioning that there is an alternative framework in social choice theory, the so-called *single-profile approach*, where each individual is assumed to have a single fixed preference ordering. If we apply this framework to our context, we can simply replace individuals with axiologies. In that sense, it might be somewhat more natural for our purposes than the multi-profile framework. However, to make up for the loss of formal structure, this single profile approach has to make comparatively strong assumptions about the set of options $\mathcal{X}$. It is not clear to me which framework is ultimately preferable for the theory of axiological uncertainty. Fortunately, there is a general method for converting results from the multi-profile framework into the single-profile framework. (Cf. particularly Roberts (1980b).) As far as I see, that would be possible for all results that I state. But for reasons of space I shall not do that in this thesis. So since I find the multi-profile framework viable, and since the multi-profile results will in any case be necessary to establish their single-profile analogues, I state my results in this framework.

labelled i are represented by real-valued value-functions $V_i(\cdot)$, defined on $\mathcal{X}$. A u-value-functional F is a function from the set of all n-tuples $\{V_i\}$ of value-functions to the complete u-value orderings on $\mathcal{X}$. It specifies exactly one u-value ordering $\succeq_V$ for any given n-tuple $\{V_i\}$ (given the probability distribution at hand). I shall write

$$\succeq_V = F(\{V_i\}), \tag{3.4}$$

and similarly $\succeq_{V'} = F(\{V_i'\})$ for the relation induced by some different n-tuple of value-functions $\{V_i'\}$. The respective strict u-betterness relation, and the equally-as-u-good relation, will be denoted by '$\succ_V$' and '$\sim_V$' respectively.

So put thus, the u-value relations $\succeq_V$ are relative to the underlying n-tuples $\{V_i\}$ of value-functions associated with the probabilities. But we can generalise the u-value relation. To that end, let $\mathcal{L}$ be the set of all possible n-tuples of value-functions, and assume that the domain of our u-value-functional is this entire set:

Unrestricted Domain (U): The domain of F is $\mathcal{L}$.

This is certainly a desirable constraint. It would be unfortunate if there was no plausible theory of axiological uncertainty that was general, in the sense of being defined for any set of value-functions. Next, assume that if two options are equally good according to all axiologies, they are equally u-good:

Pareto Indifference (PI): For any pair x, y and any n-tuple $\{V_i\}$, if $V_i(x) = V_i(y)$ for all i, then $x \sim_V y$.

This too seems to be a very plausible assumption about u-value; in fact, the

same idea was part of the Pareto Condition that I introduced in chapter 2. Finally, suppose the u-value relation between any two options should depend only on the values that *these* options have according to the axiologies under consideration. The values of any other options should be irrelevant. That is:

> **Independence of Irrelevant Alternatives (I)**: For any two n-tuples $\{V_i\}$ and $\{V_i'\}$, and for any pair x, y, if $V_i(x) = V_i'(x)$ and $V_i(y) = V_i'(y)$ for all i, then $x \succeq_V y$ if and only if $x \succeq_{V'} y$.

Though perhaps more controversial, I take it that this assumption too is at least *prima facie* plausible. If all these conditions hold, then all the orderings $\succeq_V$ can be represented by a single u-value relation $\succeq$ on $\mathbb{R}^n$. That is, for $V(x) = (V_1(x), ..., V_n(x))$ we have

> **Theorem 3.1**: If a u-value-functional F satisfies (U), then F satisfies (PI) and (I) if and only if there exists an ordering $\succeq$ on $\mathbb{R}^n$, such that for all x and y in $\mathcal{X}$, and for all n-tuples $\{V_i\}$, $x \succeq_V y \;\Leftrightarrow\; V(x) \succeq V(y)$. Moreover, if these conditions hold, $\succeq$ is unique.[36]

This means that, given (U), (PI) and (I), we need not restrict our attention to the indexed relations $\succeq_V$. Instead, we can focus on the more general relation $\succeq$ on $\mathbb{R}^n$. For the most part of this section, this is what I shall do.

So let me now introduce measurability and comparability assumptions into this framework.[37] The basic idea is simple: the specific extent to which value is measurable or comparable determines a set of n-tuples $\{V_i\}$ that are informationally equivalent. For example, suppose value is cardinally measurable

[36] This theorem is due to D'Aspremont and Gevers (1977, Lemma 3) and Hammond (1979, Theorem 1).

[37] My exposition of these constraints closely follows Blackorby et al. (1984).

on each axiology, and fully intertheoretically comparable. Then two n-tuples $\{V_i\}$ and $\{V_i'\}$ reflect the same probability distribution over the same axiologies if and only if one can be obtained from the other by subjecting each $V_i(\cdot)$ to one and the same positive, affine transformation. If value is less well measurable or comparable, the set of such informationally equivalent n-tuples increases.

Our assumption will be that given cardinality and full comparability, say, we can require that such informationally equivalent $\{V_i\}$ and $\{V_i'\}$ should induce the same u-value ordering on $\mathcal{X}$, i.e. that $\succeq_{V'} = \succeq_V$. Or rather, our assumption will be that saying that value is cardinally measurable and fully intertheoretically comparable *means*, perhaps among else, that such $\{V_i\}$ and $\{V_i'\}$ induce the same u-value ordering. This is at least the underlying assumption on this social choice approach. So we shall define a set Φ of *invariance transformations* $\phi = (\phi_1, ..., \phi_n)$, from $\mathbb{R}^n$ to $\mathbb{R}^n$, which specifies the relationship between all informationally equivalent n-tuples: $\{V_i\}$ and $\{V_i'\}$ are informationally equivalent if and only if they are related by an invariance transformation in Φ – i.e. $V(x) = \phi(V'(x)) = [\phi_1(V_1'(x)), ..., \phi_n(V_n'(x))]$ for all x. And we shall require that all $\{V_i\}$ and $\{V_i'\}$ that are related by an invariance transformation in Φ induce the same u-value relation. Specific informational assumptions can then be expressed by defining the set Φ of invariance transformation in specific ways. Again, the lesser the extent to which value is measurable or comparable, the greater the set Φ, and thus the grater the set of informationally equivalent $\{V_i\}$ and $\{V_i'\}$.

In the presence of (U), (PI) and (I), such informational constraints will impose structure on the general u-value relation $\succeq$. To see this, suppose $\{V'\}$

is obtained by applying the invariance transformation ϕ to $\{V\}$. If $v = V(x)$ and $w = V(y)$, we have $v' = V'(x) = \phi(v)$ and $w' = V'(y) = \phi(w)$. Since $\{V\}$ and $\{V'\}$ are informationally equivalent, $\succeq_V$ and $\succeq_{V'}$ must rank x and y in the same way. Thus the general u-value relation between v and w must be identical with that between v' and w'. We shall thus state the invariance requirement formally as:

> **Invariance Requirement**: For all $v, v', w, w' \in \mathbb{R}^n$, if $v' = \phi(v)$ and $w' = \phi(w)$ for some $\phi \in \Phi$, then $v \succeq w \Leftrightarrow v' \succeq w'$.

The constraint I have just characterised, of cardinally measurable and fully comparable values, can now be specified as follows:

> **Cardinal Full Comparability (CF)**: $\phi \in \Phi$ if and only if there are real numbers a and b, $b > 0$, such that $\phi_i(t) = a + bt$ for all i.

(CF) is an optimistic informational assumption; it renders the set of invariance transformations comparatively small. In what follows, I shall consider two more pessimistic assumptions than (CF). Our more pessimistic informational assumption will be

> **Cardinal Noncomparability (CN)**: $\phi \in \Phi$ if and only if there are real numbers a_i and b_i, $b_i > 0$ for all i, such that $\phi_i(t) = a_i + b_i t$ for all i.

If the set of invariance transformations is characterised by (CN), then value is measurable cardinally within each axiology, and in no way comparable across axiologies. This makes the set of invariance transformations relatively large. Our more optimistic assumption will be that value is measurable cardinally within each axiology, and unit-comparable across axiologies:

Cardinal Unit-Comparability (CU):$\phi \in \Phi$ if and only if there are real numbers a_i and b, $b > 0$, such that $\phi_i(t) = a_i + bt$ for all i.

Many further constraints could be expressed in this manner. But for present purposes this will do.

Before moving on, however, it might be worth emphasising the philosophical background assumptions of this framework. Note that we are assuming that a u-value-functional F maps an n-tuple $\{V_i\}$ to a *complete* binary relation on $\mathcal{X}$. So in this framework, we are from the outset – irrespective of the measurability and probability assumptions that we then impose – considering only *complete* theories of u-value. We are simply assuming that the u-value relation is complete. As I have argued, I find this assumption dubious. I think that scepticism about intertheoretic comparisons should actually give rise to incompleteness. But for the sake of argument, I am accepting this alternative assumption now. So let me now turn to discussing how these constraints vindicate different theories of axiological uncertainty.

3.5.2 My Favourite Theory and Weighted Value Maximisation

My Favourite Theory

To define My Favourite Theory for some credence distribution P, let $M \subset P$ refer to the set of maximal probabilities in P, i.e. the set of probabilities in P that are at least as great as any other. This set may have more than one member, if at least two theories both have maximal probability. For present purposes, we shall define My Favourite Theory formally as

My Favourite Theory: For some $m \in M$, and for all $v, w \in \mathbb{R}^n$, if $v_m > w_m$, then $v \succ w$.[38]

We shall derive this view from Cardinal Noncomparability. This will substantiate the abovementioned rationale behind My Favourite Theory, the argument from scepticism. The equivalent assumption in social choice theory – i.e. the incomparability of wellbeing across people – figured prominently in Kenneth Arrow's (1963) impossibility result. And in fact, My Favourite Theory is simply what Arrow called a 'dictatorship' – in our analogy, the dictatorship of one probability. So to derive My Favourite Theory, we shall use a version of Arrow's (1963) impossibility result and turn it into a possibility result for an Arrowian 'dictatorship'.

As Arrow's result showed, Noncomparability is a surprisingly strong assumption. Only a few rather weak additional conditions are necessary to derive a dictatorship from it. I have already introduced three of them, the Unrestricted Domain condition, Independence of Irrelevant Alternatives, and Pareto Indifference. I now have to add a slightly different Pareto condition. But it is again very weak and plausible:

Weak Pareto (WP): For all $v, w \in \mathbb{R}^n$, if $v_i > w_i$ for all i, then $v \succ w$.

Together with (U), (PI) and (I) and the relevant informational assumption, (WP) suffices to imply the 'dictatorship' of one probability. So if we accept all of these conditions, we *have* to accept that within each probability distribution, there is one probability such that the axiology with that probability

[38]Note that, defined thus, My Favourite Theory is only a one-way implication; so the present view is slightly different from how I understood 'My Favourite Theory' up to now. But since the core idea is preserved, I shall refer to it with the same name.

comes to dominate u-value. So once we are thus far, a fairly weak final assumption suffices to imply My Favourite Theory, viz.:

> **Anti-Improbabilism (AI)**: For all $i \in P \setminus M$, there are $v, w \in \mathbb{R}^n$, such that $w_i > v_i$ and $v \succeq w$.

This condition requires that for each non-maximal probability, at least some theory with that non-maximal probability does not *always* dominate u-value against all other theories. So Anti-Improbabilism too is very plausible. The only way to deny it is to assign to a non-maximal probability the kind of dominance that My Favourite Theory assigns to the maximal probability. This does not seem very attractive, and has never been suggested in the literature. Given (AI), we can state

> **Theorem 3.2**: If the u-value-functional F satisfies (U), (PI), and (I), then the u-value relation $\succeq$ satisfies (WP), (AI) and (CN) if and only if My Favourite Theory is true.[39]

This results spell out the formal relationship between scepticism about intertheoretic comparisons and My Favourite Theory. It shows that, at least given our framework, the transition from the former to the latter is indeed compelling.

Note that, strictly speaking, we have now established this result only for one underlying probability distribution P. But we have not made any assumption about P. So our result naturally generalises to the whole set of

[39]Cf. Blackorby at al. (1984, Corollary 4.1) for the claim that, without (AI), the conditions imply that for *some* i in H, if $v_i > w_i$, then $v \succ w$. (AI) straightforwardly implies that this must be the case for some $m \in M$. In fact, Theorem 3.2 also holds if we replace (CN) with *Ordinal Noncomparability*: $\phi \in \Phi$ if and only if ϕ_i is an increasing transformation for all i. Cf. Blackorby et al. (1984, Theorem 4.1).

probability distributions over n axiologies: if u-value satisfies the axioms of Theorem 3.2, then for any such probability distribution, My Favourite Theory is true. It is thus, then, generally true.

Weighted Value Maximisation

It is interesting to see the radical effects of introducing intertheoretic comparability. Consider now the type of view to which EVM belongs, viz.,

> **Weighted Value Maximisation**: there is a vector $a \in \mathbb{R}^n_+$, with $a_i > 0$ for some i, such that $v \succeq w$ if and only if $\sum_i a_i v_i \geq \sum_i a_i w_i$.

Note that My Favourite Theory is consistent with some forms of Weighted Value Maximisation. More precisely, the borderline case in which only one weight a_i is strictly positive corresponds to My Favourite Theory. But the family of weighted value maximising views is much broader. It also includes views on which more than one theory is given a positive weight, and these are important rivals to My Favourite Theory. In social choice theory, the equivalent views are sometimes called 'weighted', or 'generalised utilitarianism'. They are forms of utilitarianism on which the welfare functions of different individuals potentially have different weights.

If value is cardinally measurable and unit comparable, one additional condition again suffices to axiomatise this family of theories. This condition is a continuity axiom. Intuitively, it requires that, for any u, v and w in $\mathbb{R}^n$, if $u \succ v$, and $v \succ w$, any curve connecting u and w must cross the indifference curve containing v; there are no sudden jumps from being u-better than v to being u-worse than v. A formal way to require this is:

Continuity (C): For all $v \in \mathbb{R}^n$, $\{w \in \mathbb{R}^n | w \succeq v\}$ and $\{w \in \mathbb{R}^n | v \succeq w\}$ are closed.

Given this axiom we can state

Theorem 3.3: If the u-value-functional F satisfies (U), (PI), and (I), then the u-value relation $\succeq$ satisfies (WP), (C) and (CU) if and only if Weighted Value Maximisation is true.[40]

Again, this result generalises over probability distributions, and the axioms imply that, for any probability distribution over axiologies, a form of Weighted Value Maximisation must be true. So once unit comparisons are allowed, these theories become very plausible. Note also that Weighted Value Maximisation is consistent with (U), (PI), (I), (WP) and (AI). These conditions, together with the relevant informational constraint, implied My Favourite Theory. So in this sense, the difference between these views is importantly a matter of our assumptions about the comparability of value.

Conclusion

In a sense, the theorems I have introduced were just examples. Many further theorems could be proved in this way.[41] Using the framework from social choice theory, we can show formally how different theories of axiological uncertainty can be derived from different assumptions about the measurability and comparability of value.

[40]Cf. Blackorby et al. (1984, Theorem 7.1), also Roberts (1980a, Theorem 2).

[41]E.g., Blackorby et al. (1984) also derive a 'lexicographic maximin rule' from level comparisons; this could be used to axiomatise a lexical form of risk aversion in our context.

However, let me emphasise again that this entire section was premised on an assumption that I ultimately find dubious. The present framework simply assumed that the u-value relation is complete, irrespective of facts about measurability and comparability. In contrast, I think that incomparability should give rise to incompleteness. So even though the results in this section explore an interesting space of possible views, in the next chapter, I shall adopt my previous approach and framework again.

Conclusion

In this chapter, I first argued that there are at least some positive facts about intertheoretic comparisons. I then explored two explanations of *why* these facts hold.

According to comparativist accounts, axiologies are merely orderings, and there are independent facts about how these orderings enter the u-value relation. I suggested that under this view, the Idea of Equal Say seems plausible, and that this idea in turn plausibly implies that $\succeq_U$ satisfies Completeness and Continuity. According to substantial or non-substantial absolutist accounts, axiologies themselves make claims about the sizes of value differences or heights of value levels on an intertheoretic scale. I suggested that under this view, there should be theories that are less than fully comparable, as well as theories that are fully incomparable or compare in a lexical way, even though the latter two sorts of theories are extreme and comparatively implausible. So ignoring these extreme theories, $\succeq_U$ will satisfy Continuity,

but not Completeness. I outlined how specific comparativist and substantial absolutist accounts could be axiomatised by extending the Expected Value Theorem. Ultimately, however, I have tentatively adopted Non-Substantial Absolutism.

Chapter 4

Subjective Probabilities Under Axiological Uncertainty

Introduction

Before I go on to explore the possibility of an incomplete u-value relation, there is another limitation of the Expected Value Theorem that I need to address. In the Expected Value Theorem, the concept of a probability distribution over axiologies figured as a primitive. More specifically, the options over which the u-value relation ranges were *defined* as leading to particular outcomes with particular probabilities, while different axiologies have particular, quantitatively specified probabilities of being true. So the concept of a probability distribution over axiologies appeared as a primitive in the very definition of these options: each option $\boldsymbol{a}$ in $\mathcal{Q}$ specifies an underlying probability distribution over axiologies. Taking the concept of probabilities for granted in this way was useful to focus on the other problems I have

been discussing so far. Ultimately, however, it is unsatisfying. The theory of axiological uncertainty should not take these probabilities as undefined primitives. We need an account of what it *means* that an axiology has a particular probability. Or more precisely, we need an account of what it means that an agent has a particular *degree of belief* in an axiology – or so I shall argue. The present chapter addresses this problem. I shall call these degrees of belief *axiological credences*, and shall speak of an *axiological credence distribution* accordingly. So the main question of this chapter is how can we understand axiological credences for the purposes of EVM.

The nature of this question depends on what precisely axiologies are. If an axiology is merely an ordering, the relevant question is how we can understand credences in orderings; if an axiology is an ordering together with a claim about the heights of value levels or sizes of value differences on an intertheoretic scale, the question is how we can understand credences in these more complex theories. The answers to these questions might differ. It is beyond the scope of this thesis to discuss the problem of credences for each account of axiologies. As I mentioned on page 150, I find the idea of Non-Substantial Absolutism most plausible. Moreover, in a sense I shall explain at the end of this chapter, the theory of Non-Substantial Absolutism is the most *general* one, lending itself most readily to different interpretations and specifications. So for most parts of this chapter, I shall ask this question about credences specifically for Non-Substantial Absolutism.

To my knowledge, no author writing on normative uncertainty has addressed the problem of how to understand the relevant credences. Yet it is a very important one. As it will emerge, our answer to it will have significant

implications for the structure of our entire theory.

In line with my approach so far, I shall rely on a representation theorem to understand axiological credences. In their heydays – culminating in Savage's *Foundations of Statistics* – the alleged ability of representation theorems to provide a notion of credences may have been the pride of decision theory. But recently, many philosophers have become very sceptical about these theorems. Many believe that representation theorems are neither necessary nor sufficient for understanding credences – that we can understand the notion of credences even without the help of these theorems, and that (even if we could not) we could not understand them with the help of these theorems either. In this chapter, I shall argue against both of these claims. I shall defend the alternative view that representation theorems provide the best account of credences, and thus that they can and should play an important role in grounding EVM. In a way I shall explain, this in turn will have important implications for the *normative* structure of EVM. So in sum, the primary question of this chapter is what credences are. But at the core of it is an argument defending the use of representation theorems for understanding and grounding EVM. And one of its most important upshots will be a conclusion about the precise ways in which EVM can be normatively significant.

The chapter will proceed as follows. In section 4.1, I shall present some alternative replies to the question of how to understand credences, and argue that they are unsatisfying. This will provide a first motivation for looking at representation theorems as a solution.

In section 4.2, I shall first give a brief general introduction into decision-theoretic explications of credences. I shall then present an explication on the

basis of a representation theorem from state-dependent utility theory, and outline more fully than I have so far done, how – given this explication – my theory of axiological uncertainty can be applied in real-life decision making.

In section 4.3, I discuss some objections to this explication. I shall ultimately argue that representation theorems provide our best account of credences, and outline the implications of this claim for the normative significance of EVM.

In section 4.4, I first briefly discuss how the problem of credences could be addressed under alternative assumptions about what axiologies are – viz., Comparativism and Substantial Absolutism. And I then briefly examine a conceptually simpler theory of axiological uncertainty that eschews the concept of probability altogether – viz., Weighted Value Maximisation.

4.1 The Problem

What does it mean to say that you have a particular credence distribution over axiologies; and why should we turn to representation theorems to understand this? A first, flat-footed response to the problem of understanding credences would be to take the notion of an axiological credence distribution as a primitive. In decision theory, some authors have recently advocated this primitivism about credences more generally.[1] If we accept it in the theory of axiological uncertainty, then as far as the notion of probability is concerned, we do not need to go beyond the Expected Value Theorem to axiomatise

[1]Cf. Eriksson and Hájek (2007); Paseau (ms, 25f.) even says: it 'is by now relatively familiar that degrees of belief are most plausibly construed as theoretical primitives'.

EVM.

However, I think that primitivism about credences is very unsatisfying, at least in the theory of axiological uncertainty. The case of credences is parallel to that of intra- and intertheoretic comparisons of value. As I argued in chapter 2, we need an account of these comparisons because EVM presupposes a quantitatively significant concept of value, and unless more is said, we do not seem to understand that concept. The same holds for axiological credences. EVM presupposes a *quantitative* concept of probabilities. So to understand what 'EVM' means, we must know what it means to have a particular, quantitatively specified credence distribution over axiologies – to have, say, a 0.3 credence in some theory T_i. And we need an account of what that means. Suppose you said 'my credence in this form of utilitarianism was once 0.3, but now it is 0.2'. Unless you give me *some* account of what you mean by that, I would not understand what you meant – only that you chose a slightly swaggering way of expressing that you thought utilitarianism fairly likely, and now give it less credit, which I would also think if your numbers were 0.4 and 0.3. So unless we give some account of credences, I take it, it is simply unclear what it means for you to have a 0.3 credence in T_i, as opposed, say, to a 0.35, or a 0.4 credence in that theory.

This problem becomes particularly evident if we realise how *rich* the notion of an axiological credence distribution is. It is *triply* quantitative: specifying quantitative subjective probabilities attached to intratheoretically cardinal axiologies among which cardinally significant comparisons hold. The last point is particularly important. At least under Absolutism, there are infinitely many distinct axiologies corresponding to each axiological ordering.

So it is not enough to know what it is for you to find a particular *ordering* very plausible – or even to assign a 0.3 probability to that ordering. We need to know what it means for you to have a particular credence in axiologies that *compare* in a particular, cardinally significant way. Unless we know *that*, we do not know what it means that one of your options has a higher expected value than another, and thus what EVM means. And it is not plausible that our intuitive understanding of such a rich notion goes very far.

Note also that this problem does not arise specifically because I have been considering *precise* credences, as opposed to imprecise or fuzzy ones that cannot be represented by single real numbers. Just as with precise credences, we do not have an intuitive understanding of what it means to have a credence of '*roughly* 0.3' in T_i, as opposed to 'roughly 0.4', or 'roughly 0.2' – let alone of what it means to have a particular credence interval ('between 0.2 and 0.4'). Again: if you said 'my credence in this version of utilitarianism was once *about* 0.3, but now it is rather somewhere *around* 0.2', I would not understand what you meant. The general problem does not hinge on the precision of the probabilities.

It might be that we can make important progress in some areas of epistemology or even decision theory while taking credences as primitives. But ultimately, at least in the theory of axiological uncertainty, that is unsatisfying. If we want to defend EVM, we should be able to explain what it says.

There would be a very simple explanation of what we mean by EVM. Consider the

Simple Explication: That your credence in T_i is p_i means that when you consider the set of theories $\mathcal{T}$, consider how confident you feel about each axiology upon reconsidering the evidence for and against it, and try to associate a nonnegative number to each of them, such that the numbers add up to 1 and reflect your confidence, then you associate with T_i the number p_i.

On this account, credences are given through mere introspection. I have often encountered this idea in conversation ('why are you making such a fuss about credences; surely we have a fairly decent grasp of what they are!'). And at least if we have a way of identifying the relevant axiologies,[2] this may well be a possible explication. If you say that your credence in T_i is 0.3, and that you understand this statement in terms of the Simple Explication, I know what you mean. You mean that you sat down with your list of axiologies, tried to distribute numbers in accordance with your feelings of confidence, and T_i ended up getting the number 0.3.

However, it seems implausible that our best theory of axiological uncertainty is EVM as understood in terms of the Simple Explication. Suppose that my list of axiologies features versions of total utilitarianism, and the views that (in addition to wellbeing) beauty, biodiversity, friendship, autonomy, virtue or equality have value; suppose it features axiologies with different conceptions of wellbeing; and suppose that it features multiple versions of these axiological orderings too. If I had to, I could consult my feelings of confidence and write down a respective set of numbers. And I would certainly have a grasp on *some* facts about how confident I am, such as that

[2]Cf. footnote 32 in this chapter (and the main text to this footnote) for this caveat.

I am more confident in total than in average utilitarianism. But very soon, this project of assigning numbers to axiologies will seem extremely arbitrary and groundless to me. In other words, the problem with this explication is that it does not seem to pick out something of ultimate normative importance. It would seem implausible and slightly reckless to claim that I should ground the most important decisions in my life solely on this intuitive list of numbers. If our theory of axiological uncertainty should guide us in our decision making, we should find a more robust and important understanding of credences.

A third response to our challenge would be that we should turn to scientists and psychologist to explain our notion of credences. The concept of 'credences' after all seems to be an empirical, psychological one. So one might think that only serious empirical work will give us an adequate and scientifically informed account of credences.[3]

However, I think that the most plausible way of integrating scientific or psychological research into our account of credences still essentially makes use of something like representation theorems. So let me now turn to representation theorems. I shall come back to this idea of a scientifically informed account of credences on page 198.

[3]Such a view is hinted at in Meacham and Weisberg (2011, 642; 661).

4.2 Subjective Expected Value

In this section, I shall offer what I take to be the most promising account of axiological credences. Section 4.2.1 provides a brief introduction into decision-theoretic explications of subjective probabilities, and explains the account I shall explore on an intuitive level. Section 4.2.2 outlines the formal theorem on which my explication is based. And section 4.2.3 outlines how – given this explication – my theory of axiological uncertainty can be applied in real-life decision making.

4.2.1 De Finetti, Ramsey and Savage, and the Structure of the Argument

The problem of explaining degrees of beliefs in *non*-normative propositions has a long history in decision theory, and various different accounts have been proposed. One of the first to address that problem was Bruno de Finetti (1980), who suggested what is now standardly called a *betting interpretation* of degrees of belief. De Finetti says:

> Let us suppose that an individual is obliged to evaluate the rate p at which he would be ready to exchange the possession of an arbitrary sum S (positive or negative) dependent on the occurrence of a given event E, for the possession of the sum pS; we will say by definition that this number p is the measure of the degree of probability attributed by the individual considered to the event E, or, more simply, that p is the probability of E (according to the individual considered [...]). (1980, 62)

According to this interpretation, that you have a degree of belief of 0.25 in the proposition that it will rain this afternoon means that you are indifferent

between the bet that gives you £1 if it rains, and the sure thing gain of £0.25. So de Finetti derives your credences from your preferences about monetary lotteries. In effect, he takes quantities of money to represent what you care about, and interprets you as maximising the expectation of that quantity.

This leads to an obvious problem – viz., that what you care about may *not* adequately be represented by money. De Finetti's account seems to misrepresent your beliefs if money has diminishing marginal value for you. Even if you have the same degree of belief in a coin's landing heads or tails, you might prefer a sure gain of £1 million to the bet that gives you £2 millions if the coin lands tails (and nothing otherwise), because you care more about the first than about any subsequent million.[4] For this reason, the account given by Frank Ramsey (1926) was arguably more promising. Ramsey introduced a notion of 'utility' that need not conform linearly to any other quantity. More generally, he derived *both* probabilities and utilities from preferences alone. He provided a simple representation theorem showing that if your preferences satisfy certain conditions you can be represented as maximising expected utility with respect to a particular utility function (unique up to positive affine transformation) and a (unique) probability function. According to Ramsey, this probability function can be interpreted as specifying your degrees of belief.

Ramsey's account was in turn further refined by Leondard Savage (1954), who provided a more general and conceptually more sophisticated representation theorem. And ever since, decision theorists have worked on refining

[4]Cf. e.g. Eriksson and Hájek (2007) and Hájek (2012) for other famous objections against De Finetti's interpretation.

Savage's theorem still further – for example, in allowing for state-dependent preferences and utilities.

The main tenet of all these decision-theoretic explications is, very roughly, that to have a particular credence in a proposition is to give that proposition a particular weight in your preferences under uncertainty. Representation theorems show that if your preferences satisfy certain conditions, you can be *represented as* maximising expected utility with respect to particular utility and probability functions; and according to the proponents of these theorems, you can then not only be represented *as* having these utilities and probabilities, but you actually have them: the probability function specifies your actual credences. The proposals have become more sophisticated in the range of decisions or preferences they consider, and in the precise role they require your belief to play in your decision making. But this rough core is common to them all.

As various people have emphasised,[5] proponents of representation theorems thus need an account of why your being *representable as* having certain utilities and probabilities should imply that you actually have them. There are numerous different ways in which one might try to bridge this gap. For example, one might say that it is an empirical truth of some sort that the relevant probability function would specify our real credences if we satisfied the axioms; or one might say that we take it to specify our credences as a matter of brute stipulation.[6] I cannot discuss all of these interpretations. So as in chapter 2, I shall simply invoke the argument from representation

[5]Cf. e.g. Zynda (2000), Hájek (2008), Meacham and Weisberg (2011).

[6]Cf. Eriksson and Hájek (2007), Meacham and Weisberg (2011) or Buchak (2013, ch.1) for overviews.

theorems that I find most promising. On this interpretation, again, we use representation theorems as an *explication* for credences. That is, the claim is that the probabilities and utilities from representation theorems capture what is useful about the folk notions of belief and value, and drop what is problematic or undesirable about them. So that shall be my main claim in what follows.

However, if we are to explain your *axiological* credences, we cannot simply focus on your preferences. At least, I cannot see a convincing way of doing that. The reason is that you may care about all sorts of things besides moral value – your own self-interest, the welfare of your nearest and dearest, or whatever – and you may or may not care very much about moral value. So for example, if you give total utilitarianism only very little weight in your ordinary *preferences*, that need not mean that you find it implausible *as* an axiology. It may simply mean that you do not care very much about axiological value in general, even though *as* an axiology, you find total utilitarianism plausible.

So in order to define axiological credences, we somehow need to separate your axiological and non-axiological concerns. One way to do so is to focus not on which options you find preferable to which, but on which options you find *u-better* than wich. In other words, instead of explicating credences in terms of preferences, we can explicate them in terms of *u-value judgments*. Very roughly, we might say that your credence p_i in an axiology T_i is the weight you give that axiology in your u-value judgments. There might be other ways of separating axiological and non-axiological concerns, and I shall briefly explore some alternatives in section 4.3.3. But I take this to be the

most promising account, and so it is this idea that I shall now develop at some length. I shall refer to this explication as the *judgmement-based explication* of axiological credences, and to a form of EVM understood in terms of it as *judgment-based EVM*. So what I shall look for is a set of conditions $\mathcal{C}$ such that, if your u-value judgments satisfy conditions $\mathcal{C}$, you can be represented as making u-value judgments in accordance with EVM. And I shall treat it as a matter of explication that you then *are* satisfying EVM.

Note that this conditional is not a normative claim. Its consequent says that you do satisfy EVM, not that you *should* do so. To establish that conclusion, we need an extra normative premise, saying that your u-value judgments *should* satisfy the relevant conditions $\mathcal{C}$. So the structure of the overall argument for judgment-based EVM is as follows:

(A) Your u-value judgments should satisfy conditions $\mathcal{C}$;

(B) as a matter of logical equivalence, your u-value judgments satisfy conditions $\mathcal{C}$ if and only if you are following EVM; therefore

(C) you should follow EVM; EVM is true.

I shall discuss the overall form and significance of such an argument in section 4.3.2. First, let me now provide a set of conditions $\mathcal{C}$ that can fill this scheme.

4.2.2 The Subjectivist Expected Value Theorem

I shall again take a theorem from state-dependent utility theory, extend it slightly and apply it to our context. As I mentioned on page 167, I shall

intend this theorem as an account of credences under Non-Substantial Absolutism. I shall come back to the question of how to understand credences under alternative assumptions about axiologies in section 4.4.1.

To explicate probabilities, we need a theory that is based on a slightly different formal framework than I have so far been using, such that the main binary relation does not range on options in $\mathcal{Q}$. So let me introduce a new set, $\mathcal{K}$, defined as $\mathcal{K} = \{\mathfrak{a} : Z \to \mathbb{R}_+ \mid \sum_{x \in X} \mathfrak{a}(i,x) = 1 \quad \forall i \in I\}$.[7] I shall use lower case Fraktur-letters, $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$... to refer to members of $\mathcal{K}$, and shall for simplicity again call them 'options'. Options in $\mathcal{K}$ differ from options in $\mathcal{Q}$:[8] in an option in $\mathcal{K}$, the numbers assigned to outcomes sum to 1 *within* each axiology, not across axiologies. So an option in $\mathcal{K}$ does not specify an underlying probability distribution over axiologies. It *does*, however, specify an underlying probability distribution over *outcomes* (relative to each axiology). In this sense, in considering options in $\mathcal{K}$, I am still taking *non-normative* probabilities as given primitives. In principle, this is problematic – for exactly the same reasons as it was problematic to take axiological credences as a primitive. However, my main concern is with *axiological* uncertainty. That problem is complicated enough. So to simplify my treatment of it, I shall just assume non-normative probabilities as given primitives throughout this thesis.

We can again define $p\mathfrak{a} + (1-p)\mathfrak{b} \in \mathcal{K}$ as the option that leads to $\mathfrak{a}$ with probability p, and to $\mathfrak{b}$ with probability $(1-p)$, hence $(p\mathfrak{a} + (1-p)\mathfrak{b})(i,x) = p\mathfrak{a}(i,x) + (1-p)\mathfrak{b}(i,x)$ for all (i,x) in Z. And again, even though options

[7] As a reminder, Z was defined as $Z = \{(i,x) \mid i \in I, x \in X\}$, I being the index set of axiologies, and X the set of outcomes.

[8] $\mathcal{Q}$ was defined as $\mathcal{Q} = \{\boldsymbol{a} : Z \to \mathbb{R}_+ \mid \sum_{(i,x) \in Z} \boldsymbol{a}(i,x) = 1\}$.

like $\mathfrak{a}$, $\mathfrak{b}$, and $\mathfrak{c}$ (in $\mathcal{K}$) are formally distinct from options like a, b and c (in $\mathcal{O}$[9]), and I said that axiologies order the latter, I shall sometimes say that $\mathfrak{a}$ *is at least as good as* $\mathfrak{b}$ according to an axiology T_i. By this I shall mean, intuitively, that the prospect represented by $\mathfrak{a}$, given T_i, is at least as good according to T_i as the prospect represented by $\mathfrak{b}$, given T_i. To define this concept formally, define for each axiology T_i a function $K_i : \mathcal{K} \to \mathcal{O}$; $\mathfrak{a} \mapsto K_i(\mathfrak{a})$, such that

$$K_i(\mathfrak{a})(x) = \mathfrak{a}(i, x). \tag{4.1}$$

The mapping K_i thus formally turns an option $\mathfrak{a}$ into the prospect that $\mathfrak{a}$ represents, given T_i. So for some $\mathfrak{a}$ and $\mathfrak{b}$, I shall say that $\mathfrak{a}$ is at least as good as $\mathfrak{b}$ according to T_i if $K_i(\mathfrak{a}) \succeq_i K_i(\mathfrak{b})$ – and similarly for 'better' and 'equally good', and (when I am making a descriptive claim) for the fact that you *judge* $\mathfrak{a}$ u-better than $\mathfrak{b}$ according to some theory.

It depends on your axiological credence distribution whether an option $\mathfrak{a}$ is u-better than an option $\mathfrak{b}$. If $\mathfrak{a}$ is better than $\mathfrak{b}$ according to some theory T_i, and you are certain that T_i is true, $\mathfrak{a}$ will be u-better *for you, now.* If instead you are certain that some theory T_j is true, and according to this theory $\mathfrak{b}$ is better than $\mathfrak{a}$, then $\mathfrak{b}$ will be u-better for you, now. So there is no uniquely correct u-value relation on $\mathcal{K}$. Different people can make different u-value judgments about $\mathcal{K}$, and these can all be correct *as* u-value judmgents. There are only correct u-value relations *relative* to a particular agent at a particular time – or more specifically, relative to her credence distribution.

[9] $\mathcal{O}$ was defined as $\mathcal{O} = \{a : X \to \mathbb{R}_+ | \sum_{x \in X} a(x) = 1\}$.

So if I were precise, I would index u-value relations on $\mathcal{K}$ to agents and times. To avoid inconvenient notation, I shall not do that. I shall simply denote the u-value relation on $\mathcal{K}$ by '$\dot{\succeq}_U$' (or '$\dot{\succ}_U$' and '$\dot{\sim}_U$') – using a little dot to distinguish it from the relation $\succeq_U$ on $\mathcal{Q}$. It is important to bear in mind that this relation is always relative to a particular agent at a particular time. For that reason I shall sometimes speak of 'your $\dot{\succeq}_U$'. When I am making a descriptive claim, $\dot{\succeq}_U$ stands for what your (well-considered)[10] u-value judgments actually *are*. When I am making a normative claim, $\dot{\succeq}_U$ stands for what your u-value judgments should be. And that normative claim about your judgments is equivalent to a claim about value: that you *should* judge $\mathfrak{a} \dot{\succ}_U \mathfrak{b}$ simply means that $\mathfrak{a}$ *is* u-better than $\mathfrak{b}$, relative to your credence distribution. I shall use 'p_i' to represent your credence in T_i.

For any binary relation $\dot{\succeq}$ on $\mathcal{K}$, we can define the von Neumann-Morgenstern axioms and the term '*vNM-conformable*' – just as I defined them for relations on $\mathcal{Q}$ and $\mathcal{O}$:

Transitivity$_{\mathcal{K}}$: if $\mathfrak{a} \dot{\succeq} \mathfrak{b}$ and $\mathfrak{b} \dot{\succeq} \mathfrak{c}$, then $\mathfrak{a} \dot{\succeq} \mathfrak{c}$;

Completeness$_{\mathcal{K}}$: for any $\mathfrak{a}$ and $\mathfrak{b} \in \mathcal{K}$, $\mathfrak{a} \dot{\succeq} \mathfrak{b}$ or $\mathfrak{b} \dot{\succeq} \mathfrak{a}$;

Independence$_{\mathcal{K}}$: if $\mathfrak{a} \dot{\succ} \mathfrak{b}$ and $p \in]0,1[$ then $p\mathfrak{a} + (1-p)\mathfrak{c} \dot{\succ} p\mathfrak{b} + (1-p)\mathfrak{c}$ for any $\mathfrak{c} \in \mathcal{K}$;

Continuity$_{\mathcal{K}}$: if $\mathfrak{a} \dot{\succ} \mathfrak{b}$ and $\mathfrak{b} \dot{\succ} \mathfrak{c}$ then there exist p and $q \in]0,1[$, s.t. $p\mathfrak{a} + (1-p)\mathfrak{c} \dot{\succ} \mathfrak{b}$ and $\mathfrak{b} \dot{\succ} q\mathfrak{a} + (1-q)\mathfrak{c}$.

[10] The notion of 'well-considered' here is intended to rule out obvious misrepresentations of your credences – say, as when you consistently miscalculate the implications of some axiology, but accidentally nonetheless make vNM-conformable judgments.

However, we need more than these conditions to provide a judgment-based explication of credences. That your u-value judgments are vNM-conformable does not imply that there is a *unique* probability distribution over axiologies, with respect to which you are maximising expected value. Suppose that in making vNM-conformable u-value judgments you give considerable weight to the total utilitarian ordering. At least under Absolutism, this can be explained in two ways: it might be that you have a relatively high credence in total utilitarianism; or it might be that you have credence in a version with a very inflated value-function – that is, a version on which choosing the best option *matters* very much. So in order to provide a unique separation of your probability and your value-functions, we have to add some further conditions.

There is a debate in state-dependent utility theory about how best to achieve this.[11] I shall not enter that formal debate. Instead, I shall again simply use a framework and theorem that serve my purposes. And as in chapter 2, it is a theorem due to Edi Karni and David Schmeidler (1980).[12] The basic strategy behind this theorem (as put in terms of u-value judgments) is to consider not only your u-value judgments about $\mathcal{K}$, but *also* your judgments about options in which the probability distribution over axiologies is given – that is, judgments of the form 'If the probability distribution over theories was P, I would judge...', or judgments about $\mathcal{Q}$. Assuming that your judgments about $\mathcal{K}$ and $\mathcal{Q}$ are both vNM-conformable and consistent – that is, that they are induced by the same utilities and the difference between them

[11] Cf. e.g. the discussion in Karni and Mongin (2000).

[12] The theorem and proof (with a slight obvious typing error in part '(ii)') is reproduced in Karni (1985, 17); cf. also Karni et al. (1983) and Karni and Mongin (2000) for discussion.

is explained fully by the different underlying probability distributions – we can then derive from your judgments relevantly unique utility and probability functions. This is not surprising. By considering your conditional judgments ('If the probability distribution over theories was P, I would judge...'), we can first detect your values: the function representing these judgments is a pure reflection of values, since the probabilities are already given in the options. And knowing your values, we can then detect your probability distribution by considering your ordinary judgments ('Actually, I judge...').

However, this means that in our explication of axiological credences, we have to presuppose *some* understanding of axiological probabilities as primitive. We cannot eschew primitivism altogether. But I do not think that this renders our account inadequate, or circular. What I am explicating is a *subjective* notion of probabilities. I have not said anything explicitly about the notion of probability that $\mathcal{Q}$ presupposes. We could understand that notion objectively. And there seem to be contexts in which we could understand an objective notion of axiological probabilities. For example, suppose God has determined the true axiology at the beginning of days; since he had as yet no criterion to make a good choice, he did so by way of a perfect randomising device (based, e.g., on a random subatomic phenomenon whose unpredictability is due to quantum mechanics). Though highly unnatural, this story does not seem to be conceptually incoherent. If it is not, and if we intuitively understand the objective probabilities involved in certain randomising devices, then there is a way in which we could understand the options in $\mathcal{Q}$, even without understanding axiological *credences*. Judgments like 'If the probability distribution over theories was P, I would judge...'

could be understood as meaning: 'Supposing God had set his randomising device in such a way that the probability distribution was P, I would judge...'. This adds another complexity to my framework. But I think it is only that: an extra complexity. That my explication presupposes $\mathcal{Q}$ does not make it circular, or incoherent, or overly primitivist.

Actually, we shall restrict our attention to those options in $\mathcal{Q}$, on which all axiologies have a positive probability. That is, we shall consider $\mathcal{Q}^+ \subset \mathcal{Q}$, with $\mathcal{Q}^+ = \{\boldsymbol{a} : Z \to \mathbb{R}_+ \mid \sum_{(i,x)\in Z} \boldsymbol{a}(i,x) = 1 \text{ and } \sum_{x\in X} \boldsymbol{a}(i,x) > 0 \ \forall i \in I\}$. For some i in I, and binary relations $\dot{\succeq}$ and $\succeq$ on $\mathcal{K}$ and $\mathcal{Q}$, say that $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$ *agree outside* i if for all j in $I, j \neq i$, and all x in X, $\mathfrak{a}(j,x) = \mathfrak{b}(j,x)$; and similarly for $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^+$. Say that i is *obviously null* if: (i) for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$ that agree outside i, $\mathfrak{a} \dot{\sim} \mathfrak{b}$, and (ii) there exist $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^+$ that agree outside i such that $\boldsymbol{a} \succ \boldsymbol{b}$. Say that i *is obviously non-null* if there are $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$ that agree outside i such that $\mathfrak{a} \dot{\succ} \mathfrak{b}$. Finally, define a function $L : \mathcal{Q}^+ \to \mathcal{K}$; $\boldsymbol{a} \mapsto L(\boldsymbol{a})$, such that for each i in I,

$$L(\boldsymbol{a})(i,x) = \boldsymbol{a}(i,x) / \sum_{y\in X} \boldsymbol{a}(i,y). \tag{4.2}$$

L scrapes out the probabilities from $\boldsymbol{a}$, an option in $\mathcal{Q}^+$, thus turning it into an option in $\mathcal{K}$. With this in mind, we can define the

> **Consistency Axiom**: For all $i \in I$ and all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^+$ that agree outside i: if $L(\boldsymbol{a}) \dot{\succ} L(\boldsymbol{b})$, then $\boldsymbol{a} \succ \boldsymbol{b}$; and if i is obviously non-null and $\boldsymbol{a} \succ \boldsymbol{b}$, then $L(\boldsymbol{a}) \dot{\succ} L(\boldsymbol{b})$.

Finally, say that i is *non-uniform under* $\succeq$ if for some $\boldsymbol{a}$ and $\boldsymbol{b}$ that agree

outside i, $\boldsymbol{a} \succ \boldsymbol{b}$. And say that $\dot{\succeq}$ is *non-uniform* if there are $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$ such that $\mathfrak{a} \dot{\succ} \mathfrak{b}$. Given that, Karni and Schmeidler (1980, 9) state

> **Karni and Schmeidler's Theorem 2**: Suppose that a reflexive binary relation $\succeq$ on $\mathcal{Q}$ is vNM-conformable, that a reflexive binary relation $\dot{\succeq}$ on $\mathcal{K}$ is vNM-conformable and non-uniform, and that they jointly satisfy the Consistency Axiom. Then (i) there exists a state-dependent utility function $u : Z \to \mathbb{R}$ and a probability distribution P over I such that, for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$,
>
> $$\mathfrak{a} \dot{\succeq} \mathfrak{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} P(i)u(i,x)\mathfrak{a}(i,x) \geq \sum_{(i,x)\in Z} P(i)u(i,x)\mathfrak{b}(i,x), \tag{4.3}$$
>
> and for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$,
>
> $$\boldsymbol{a} \succeq \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} u(i,x)\boldsymbol{a}(i,x) \geq \sum_{(i,x)\in Z} u(i,x)\boldsymbol{b}(i,x). \tag{4.4}$$
>
> (ii) If v is another function for which (i) is true, then v is a positive affine transformation of u. (iii) If i is obviously null, $P(i) = 0$, if i is obviously non-null, $P(i) > 0$, and if each i is non-uniform under $\succeq$, then there is no other probability distribution $Q \neq P$ for which (i) is true.

To turn this into a result about u-value, I have to express more formally the assumption behind the judgment-based explication. On page 177, I expressed it very roughly as saying that your credence in an axiology is the weight you give that axiology in your u-value judgments. Let me state this more precisely.

I shall assume that besides judgments about the u-value relation, you also make judgments about which options are better than which, according to your axiologies. That is, I assume that for each theory T_i you also

make judgments $\succeq_i$ about $\mathcal{O}$, that all these $\succeq_i$ are vNM-conformable and non-uniform, and that your u-value judgments $\succeq_U$ on $\mathcal{Q}$ satisfy the Pareto Condition (from page 59) with respect to these judgments. Furthermore, I shall again assume that if a utility function u represents one of your $\succeq_i$ ordinally, it also represents it cardinally – where this means that the cardinal intratheoretic comparisons between certain outcomes, according to your theory T_i, are the same as the ratios among the utility differences between these outcomes. Now let $u(\cdot, \cdot)$ be a theory-dependent utility function from $I \times X$ to $\mathbb{R}$, and P a probability distribution over theories. I shall say that u *represents each of your axiologies cardinally* if for each of your axiologies T_i, the utility function $u(i, \cdot)$ represents that axiology cardinally. And I shall say that u *jointly represents your axiologies cardinally* if the crosscutting cardinal intertheoretic comparisons, according to your axiologies, are the same as the respective intertheoretic utility difference ratios according to u. Furthermore, I shall say that the pair (u, P) *represents your u-value judgments about* $\mathcal{K}$ *ordinally*, if your $\dot{\succeq}_U$ is such that (4.3) holds for u and P, for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$. Similarly, I shall say that u *represents your u-value judgments about* $\mathcal{Q}$ *ordinally*, if your $\succeq_U$ is such that (4.4) holds for u, for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$. Finally, suppose that you have the credence distribution P over theories that are jointly represented cardinally by u. I shall then say that the pair (u, P) *represents your axiological beliefs cardinally.* My assumption is that, if there is a pair (u, P), with P being unique and u being unique up to positive affine transformation, such that u represents your u-value judgments about $\mathcal{Q}$ ordinally, the pair (u, P) represents your u-value judgments about $\mathcal{K}$ ordinally, and u represents each of your axiologies cardinally, then (u, P) represents

your axiological beliefs cardinally. Hence if that is so, we can assume, say, that $P(T_i)$ represents your credence p_i in the theory that is represented by the function $G_i(\cdot) = u(i, \cdot)$. More briefly, I shall express this assumption, or explication, by saying that your credences acquire their cardinal significance in the context of your weighing axiologies under uncertainty.

Since I am now only interested in your u-value relation about $\mathcal{K}$, let me simplify Karni and Schmeidler's theorem slightly in applying it to our context. Given our explications, and the assumption that each of your $\succeq_i$ is vNM-conformable and non-uniform, the following theorem holds:

> **Subjectivist Expected Value Theorem**: If your $\dot{\succeq}_U$ and $\succeq_U$ are vNM-conformable and jointly satisfy the Consistency Axiom, if your $\dot{\succeq}_U$ is non-uniform and your $\succeq_U$ satisfies the Pareto Condition with respect to your $\succeq_i$, then for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$,
>
> $$\mathfrak{a} \,\dot{\succeq}_U\, \mathfrak{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \mathfrak{a}(i,x) p_i G_i(x) \geq \sum_{(i,x)\in Z} \mathfrak{b}(i,x) p_i G_i(x). \tag{4.5}$$

Again, I understand this theorem as a non-normative claim. It says that if your u-value judgments satisfy the relevant conditions, you are as a matter of fact following (4.5) – and that is, EVM.

Clearly, the converse is also true: if there exist a probability distribution P, and functions G_i that represent your axiologies cardinally, and your judgments satisfy (4.5) (and its equivalent regarding $\mathcal{Q}$) with respect to them, then your judgments will satisfy the relevant conditions. So we can turn this into an argument for the normative truth of (4.5) by the argument-scheme set out on page 178. Denote the conditions of the Subjectivist Expected Value

Theorem by '$\mathcal{C}_S$'. And let me say that you satisfy $\mathcal{C}_S$, or that your judgments are $\mathcal{C}_S$-*conformable* if your u-value judgments satisfy these conditions. If all your axiologies are vNM-conformable and non-uniform, we might then give the following argument for EVM: your u-value judgments should satisfy $\mathcal{C}_S$; if and only if they do, you are following EVM; so you should follow EVM. This is the main argument for a form of EVM based on the Subjectivist Expected Value Theorem. In the next chapter, I shall weaken the conditions $\mathcal{C}_S$ by allowing for incompleteness in your u-value judgments. But apart from that, this is the main argument for the theory of axiological uncertainty defended in this thesis; it – or its equivalent in the next chapter – is the main argument of the thesis.

Most of the conditions of the Subjectivist Expected Value Theorem are familiar from the Expected Value Theorem. So my discussion of these conditions in section 2.3 clarifies how plausible they are as normative constraints on your u-value judgments. It is worth mentioning, however, that in the context of the relation $\dot{\succeq}_U$ on $\mathcal{K}$, there is an additional problem with Completeness. Obviously, the problem of intertheoretic comparisons carries over to the present context: your $\dot{\succeq}_U$ need not necessarily satisfy Completeness, because some of your axiologies may be less than fully comparable. But perhaps there is another reason for why your judgments need not be complete. To assume that your $\dot{\succeq}_U$ should satisfy Completeness is to assume that your credences ought to be fully precise – i.e., that they ought to be representable by a single real number. If your credences can justifiably be 'fuzzy' – less than fully precise – then your $\dot{\succeq}_U$ arguably need not be complete; and this is so even if all your axiologies are complete and fully comparable.

There are indeed strong arguments to the effect that your non-normative credences need not be precise,[13] and these plausibly carry over to the axiological case. But the Subjectivist Expected Value Theorem rules out any fuzzy credences. Actually, to simplify my results, and since the problem of fuzzy credences is not specific to axiological uncertainty, I shall be ruling out fuzzy credences throughout this entire dissertation. So this is another important restriction of this thesis. It is plausible that your axiological credences can justifiably be fuzzy, but I shall ignore this.

4.2.3 Applying Judgment-Based EVM in Practice

Before I discuss objections to the argument I have just outlined, it might be worth to take stock and provide a brief summary of how, roughly, the theory of axiological uncertainty based on the Subjectivist Expected Value Theorem and the judgment-based explication of credences can be applied in real life cases. Note that, in principle, the theory itself is silent on this. As I mentioned on page 18, I am advocating judgment-based EVM as a *criterion of u-betterness*, not as a decision procedure, or practical method that we consciously have to apply to make choices. And in principle, it might be that we best satisfy this criterion of u-betterness by always doing what some good friend tells us, or by avoiding any though of u-value altogether. But it will nonetheless be helpful to outline a possible way of applying judgment-based EVM – the way that comes closest to taking the theory actually as a decision procedure, and that I find most promising.

[13] Cf. e.g. Joyce (2005; 2010); cf. e.g. Elga (2010) and White (2010) for arguments that subjective probabilities should be sharp.

So as an example, suppose you consider becoming a vegan. You believe that becoming a vegan increases animal welfare, but costs you resources (time and money) that you are now using to produce half as much human welfare. Suppose you are uncertain about the value of animal welfare, and want to know how your axiological uncertainty affects the u-value of your becoming a vegan. For simplicity, I shall focus on just two orderings: a version of standard total utilitarianism, and a version of total utilitarianism on which animal welfare has *some* weight, but less than human welfare. I shall call the latter species-weighted utilitarianism. I shall refer to your favourite version of standard total utilitarianism as T_{SU}, and to your favourite version of species-weighted utilitarianism as T_{WU}.

So how do you determine how your uncertainty between T_{SU} and T_{WU} affects the u-value of your becoming a vegan? First of all, you have to decide what weight T_{WU} assigns to animal welfare; that is, you have to determine the precise *ordering* that T_{WU} implies. Suppose you do that, and believe that on the most plausible form of species-weighted utilitarianism, human welfare counts *twice* as much as animal welfare in determining value. (Suppose also that we understand what that means.)

So now you have to decide how these two theories compare interthereticically, and how much credence you assign to them. To decide on intertheoretic comparisons, you have to consider your u-value judgments about options in $\mathcal{Q}$ – i.e., options with given objective probability distributions over theories. This will be easiest with unnatural toy examples like the following:

$\boldsymbol{a}_x$		$\boldsymbol{b}_x$	
T_{SU}	T_{WU}	T_{SU}	T_{WU}
$p_{SU} = x$	$p_{WU} = (1 - x)$	$p_{SU} = x$	$p_{WU} = (1 - x)$
Red killed	status quo	status quo	Red killed

Table 4.1

If you choose $\boldsymbol{a}_x$ then Red will be killed if T_{SU} is true, and nothing happens if T_{WU} is correct. If you choose $\boldsymbol{b}_x$ then Red will be killed if T_{WU} is true, and nothing happens if T_{SU} is correct. Now, if you judged that $\boldsymbol{a}_{2/3} \sim_U \boldsymbol{b}_{2/3}$, say, you would think that killing Red was twice as bad on T_{WU} as on T_{SU}. But suppose you believe that, most plausibly, the value of human welfare is the same on both theories; that is, most plausibly, $\boldsymbol{a}_x$ and $\boldsymbol{b}_x$ would be equally u-good if both T_{SU} and T_{WU} have equal probability – i.e. $\boldsymbol{a}_{1/2} \sim_U \boldsymbol{b}_{1/2}$.

So next, you need to determine what credence you have in those theories. To that end, you have to consider your u-value judgments about options in $\mathcal{K}$. This might again be easiest with unnatural toy examples like the following:

$\mathfrak{a}_n$		$\mathfrak{b}_m$	
T_{SU}	T_{WU}	T_{SU}	T_{WU}
n people killed	status quo	status quo	m people killed

Table 4.2

If you choose $\mathfrak{a}_n$ then n people will be killed if T_{SU} is true, and nothing happens if T_{WU} is correct. If you choose $\mathfrak{b}_m$ then m people will be killed if T_{WU} is true, and nothing happens if T_{SU} is correct. Let us suppose that all these people would otherwise have the same level of wellbeing. If you believe that $\mathfrak{a}_1 \dot{\sim}_U \mathfrak{b}_1$ you find it equally u-good (or u-bad) to risk the death

of a person on T_{SU} as on T_{WU}. So according to the explication I've given, you assign both T_{SU} and T_{WU} the same probability; that will be probability 0.5 if you do not have credence in any other theory, and less than 0.5 if you do. If you believe that $\mathfrak{a}_2 \dot{\sim}_U \mathfrak{b}_1$ you find it equally u-good to risk the death two people on T_{SU} as the death of one person on T_{WU}. So you assign T_{SU} a probability that is half as large as the probability you assign to T_{WU}. Suppose that is what you do: you find species-weighted utilitarianism twice as likely as standard utilitarianism.

So your judgments about these fairly simple toy examples imply precisely how your favourite versions of standard and species-weighted utilitarianism compare, and how much credence you have in them. As a parenthesis, let me note that the ability to construct unnatural but simple examples like $\boldsymbol{a}_x$, $\boldsymbol{b}_x$, $\mathfrak{a}_n$ and $\mathfrak{b}_m$ seems actually very helpful in thinking about these questions. So as I indicated briefly on page 47, I think that the slightly unnatural aspects of my framework may in fact be an advantage.

In any case, you can now use those probabilities and values to determine how your uncertainty between T_{SU} and T_{WU} affects the u-value of your becoming a vegan. We were assuming that, according to your non-normative beliefs, becoming a vegan increases animal welfare (by some amount $w > 0$, say), but costs you resources that you are now using to produce half as much human welfare ($w/2$). That is:

Becoming a vegan		*Not becoming a vegan*	
T_{SU}	T_{WU}	T_{SU}	T_{WU}
animal welfare $+w$, human welfare $-w/2$	animal welfare $+w$, human welfare $-w/2$	status quo	status quo

Table 4.3

In this case, and as far as these two theories are concerned, it is u-better for you to become a vegan. To see this, we can assume (without loss of generality) that the value of the status quo is 0 on both theories, that the disvalue of decreasing human welfare by $w/2$ is $-w/2$, and that these are the only theories you have credence in. The expected value of the status quo is then zero, while the expected value of becoming a vegan is

$$\tfrac{1}{3}(w - \tfrac{1}{2}w) + \tfrac{2}{3}(\tfrac{1}{2}w - \tfrac{1}{2}w) = \tfrac{1}{6}w > 0. \tag{4.6}$$

So relative to your credences and the comparisons you are making, and as far as these two theories are concerned, it is u-better for you to become a vegan.

This is at least the beginning of a fully worked out answer to the question whether you should become a vegan; and the beginning of the full story of how our theory can help you – in Savage's terms – to 'police [your] own decisions for consistency and, when possible, to make complicated decisions depend on simpler ones' (1954, 20). Of course, you might have credence in other versions of these theories which compare slightly differently; or you might have credence in other forms of species-weighted utilitarianism which imply a slightly different ordering; or you might have still other views that

affect this decision. On the whole, your calculation of the expected value of your options will be more accurate the more of these alternatives you take into account. Moreover, it might be that you have an intuition concerning the u-value of becoming a vegan. If your intuition is that it is u-better for you *not* to become a vegan, then your u-value judgments cannot all be correct. So you need to revise some of these judgments – plausibly, until you attain some kind of reflective equilibrium in which your u-value judgments satisfy all our constraints. What is important for now is that the procedure I specified in this section is no more – but no less – than the beginning of how to apply the theory of axiological uncertainty I am outlining in this thesis to real life questions.

4.3 Objections and Implications

It is time to address some objections. My argument in section 4.2.2 again assumed that representation theorems can serve the two foundational purposes that I outlined in section 2.1: that of clarifying our quantitative concepts, and that of justifying EVM. But as I mentioned on page 168, many philosophers have become very sceptical of the significance of representation theorems, particularly with respect to the notion of credences. For example, Christopher Meacham and Jonathan Weisberg claim that 'representation theorems cannot serve either of these foundational purposes', and that, 'we should [...] lay the foundations of decision theory on firmer ground' than that provided by representation theorems (2011, 641). So in section 4.3.1, I shall discuss

objections to the effect that my technical notion of axiological credences differs too much from our intuitive concept. In section 4.3.2, I shall discuss objections against the normative relevance of judgment-based EVM. And in section 4.3.3, I shall again examine whether there are better alternative explications of axiological credences.

In most cases, I shall respond to objections that have been levelled against preference-based representation theorems in decision theory, and that apply *mutatis mutandis* to my judgment-based theorem. However, in my responses, I shall only be concerned with the role of representation theorems in the theory of axiological uncertainty. I shall ultimately claim that the judgment-based explication is the best explication of axiological credences. But I shall *not* claim that preference-based representation theorems can serve these same foundational purposes in decision theory or epistemology more generally. That is another question.[14] Moreover, I shall again not claim that something like the judgment-based explication is the best explication under all accounts of axiologies. I will again simply presuppose Non-Substantial Absolutism.

4.3.1 Relationship to the Intuitive Concept

A technically explicated concept may in some respects diverge from the original, intuitive one: it should be more precise, but it may arguably also shift the original meaning slightly. However, if the technical concept is to be an

[14]In particular, I shall ignore the debate about probabilism, the view that our credences should satisfy the axioms of probability theory. According to the judgment-based explication and my main argument, probabilism is true. But I do not understand it as an *argument* for probabilism and against rivals to probabilism; instead, I understand it as an argument for how your credences and values should interact to determine the u-value relation.

explication – rather than a stipulative definition of a completely new term – it should not deviate too much from the explicandum. In decision theory, explications along the lines of de Finetti, Ramsey and Savage have come under criticism in this respect, and some of these worries carry over to our context.

A first worry is this. According to our intuitive notion of confidence, most of us have varying degrees of belief in different axiologies: many of us find some axiologies plausible, and others very implausible. Yet we generally do not satisfy the conditions $\mathcal{C}_S$; so we can rarely be ascribed degrees of belief as defined by our explication. Hence in this sense, our technical notion diverges radically from its explicandum. Among others, Meacham and Weisberg (2011) raised this objection against preference-based definitions of credences in decision theory.[15] Referring to the technical, explicated notions of utilities and degrees and belief as '*utilities**' and '*degrees of belief**', they say: 'If people don't have degrees of belief* and utilities*, these terms will not apply to most of the same cases as the original concepts' (2011, 653). And they argue that these technical terms therefore 'cannot play a useful role in descriptive accounts of our mental states, predictive accounts of our behaviour', nor even in 'prescriptive accounts of what our behaviour ought to be' (2011, 653):

> To make degrees of belief* and utilities* relevant to epistemology and normative decision theory, these states must be linked to the states that are the topic of our normative theorizing in these domains. And since agents like us generally don't have degrees of belief* and utilities*, it's hard to see how they're relevant. (2011, 655)

The same worry applies *mutatis mutandis* in our context.

[15] Eriksson and Hájek (2007, 200f.; 203f.) and Zynda (2000, 62) make the same point.

However, this worry is misguided. It may well be true that my technical concept cannot play a useful role in descriptive or predictive accounts of our mental states or behaviour. But I am not giving such an account. I am giving a *normative* account about what your u-value judgments should be. And it is not clear why our generally not having credences (in the technical sense) should be relevant for that. 'The states that are the topic of our normative theorizing' are credences, so my explication should be similar to our intuitive understanding of credences, for sure. But I am not interested in what your credences are, or how we could describe them quantitatively, if you do not satisfy the axioms. All that my argument requires is that you should satisfy the axioms, and that *if* you do, your credences are such that you satisfy EVM. For this argument, it is enough if the judgment-based explication comes near to our intuitive concept – to the states that are the topic of our normative theorizing – *when you satisfy the axioms*. And Meacham and Weisberg's argument does absolutely nothing to challenge that. Against *normative* interpretations of EVM, their argument is simply a non-starter.[16]

So if the judgment-based explication is to be challenged, it must be ques-

[16]Strangely, Meacham and Weisberg seem to recognise this. They say: '[a normative interpretation of EVM] is not concerned with what your degrees of belief and utilities are when [the axioms are] not satisfied, since it is concerned only with the case where you do what you ought to do, i.e. where you satisfy [the axioms]. So [a normative interpretation of EVM] only needs it to be the case that the [...] representation is the correct one when it exists' (2011, 655). However, as far as I see, all the arguments that they level against a normative explication-based interpretation of EVM depend on this fact that we usually do not have degrees of belief* and utilities* – as is blatantly explicit in the last quote above. Perhaps they believe that some of the objections they raise against other interpretations also affect a normative explication-based interpretation of EVM. Or perhaps they see their claim that 'adopting degrees of belief* and utilities* trivializes normative decision theory' (2011, 653) as independent from their claim that degrees of belief* and utilities* are too unconnected to our intuitive concepts. I shall come back to their trivialising-objection on page 208, and shall show that these two claims are *not* independent.

tioned whether my explication is adequate when your u-value judgments are $\mathcal{C}_S$-conformable. But this may indeed be questioned. One relevant worry is that our ordinary concept of degree of belief is much *richer* than the notion I have introduced. Among others, David Christensen (2001) has stressed this point with regards to decision theory.[17] He argues that 'the preference-based definition leaves out important parts of our pretheoretic notion' of degrees of belief (2001, 361). For one thing, 'a person's beliefs [...] affect the way she behaves in countless ways that have nothing directly to do with the decision theorist's paradigm of cost-benefit calculation' (2001, 361); for another thing, degrees of belief not only help to explain behaviour, but also 'other psychological states and processes' (2001, 361). For example, your self-deprecating beliefs may explain why you are performing poorly in a competition, or why you are being sad or afraid, or why you release stress hormones or are physically unhealthy. Beliefs are involved in a plethora of explanatory connections, even when our preferences satisfy the axioms. This being so, Christensen points out, 'the move of settling on just one of these connections – even an important one – as definitional comes to look highly suspicious' (2001, 362). And this worry carries over to our context. Your axiological beliefs arguably play a much richer role than just determining your u-value judgments: they may explain your behaviour, your immediate reactions or attitudes, your emotions, and so on.

This is a more pertinent worry. But again I think it can be answered.

[17]The same point is endorsed by Eriksson and Hájek (2007, 208). Meacham and Weisberg (2011, 646) also highlight the rich explanatory connections of beliefs (though not as an objection against what they call the 'Explicative View', and what is basically the view I suggested); cf. also Hájek (2008, 803ff.).

Christensen focuses on Patrick Maher's (1993) understanding of probabilities and utilities, according to which

> an attribution of probabilities and utilities is correct just in case it is part of an overall interpretation of the person's preferences that makes sufficiently good sense of them and better sense than any competing interpretation does. (1993, 9)

Accordingly, Christensen says:

> a given interpretation of an agent's degrees of belief might maximize expected-utility fit with the agent's preferences while a different interpretation might fit much better with other psychological-explanatory principles. In such cases of conflict, where no interpretation makes all the connections come out ideally, there is no guarantee that the best interpretation will be the one on which the agent's preferences accord perfectly with maximizing [expected utility]. (2001, 362)

However, Maher's understanding of representation theorems – or what Christensen interprets as Maher's understanding – differs from mine. If we assume that there always *is* a best overall interpretation of an agent, it may indeed presuppose an implausible cosmic accident to assume that it always coincides with the judgment-based explication. But I have *not* offered that explication as the unequivocally best overall interpretation of an agent. At least if 'best' means 'descriptively' or 'empirically best' (as opposed to 'best for our purposes'), I think it is dubious to assume that there generally *is* an overall best interpretation of an agent.

Suppose that according to the judgment-based explication, I have much more credence in standard than in anthropocentric utilitarianism. But suppose that I in various ways react more strongly to the suffering of humans

than to the suffering of animals, and so my behaviour, my emotional reactions, my implicit attitudes and hormone levels all suggest that I have more credence in anthropocentric than in standard utilitarianism. In which theory do I then have a higher degree of belief? Do my reactions and behaviour show that I do not *really* have a much higher credence in standard utilitarianism and that I actually have anthropocentric credences; or do I act and react *against* my true credences, due to akrasia or biases or whatever? Very plausibly, our intuitive concept of credence is not precise enough to generally imply anything definitive in such cases. In other words, it seems dubious to assume that there generally is a 'best interpretation'.[18]

That is why I introduced the judgment-based definition as an *explication* of our intuitive concept, rather than an analysis. It spells out what it might mean (at least under certain conditions), that your degree of belief in an axiology is the weight you give that axiology under uncertainty. This is not the only role that axiological degrees of belief play. But it is undeniably a major one. And if you satisfy the conditions $\mathcal{C}_S$ and thus give each axiology a constant weight under uncertainty, I think it is not outright wrong or completely misleading to call the relevant weights your 'credences'.

Yet even if that is conceded, one might perhaps wonder why – of all possible connections – we should focus precisely and exclusively on u-value judgments to explicate credences. But I think there are several reasons to do so. First of all, it is important to bear in mind what role these credences play

[18]There is a literature on whether, if one professes to believe that p but acts contrary to that professed belief, one truly believes that p (cf. e.g. Schwitzgebel (2010) for an overview, and a position similar to the one expressed above); I am not aware of a discussion of such cases with regards to a *graded* notion of belief.

in judgment-based EVM. What I am ultimately claiming is that you should satisfy EVM with respect to them. So we need a notion of 'credences' that picks out something that should be relevant in your decision making. But many aspects that Christensen emphasises do not seem to be *normatively* very relevant. For example, a growing literature in psychology shows that people have implicit attitudes and biases that contradict what they overtly claim to believe and what they (presumably) take to reflect their evidence – often even sexist or racist attitudes.[19] Many of us will not have the moral emotions that we take to be fitting.[20] And presumably, many of us fail to *act* in accordance with what we take to be morally appropriate.[21] This is not surprising, given that explicit beliefs can change very quickly upon receiving new evidence, whereas our behaviour, emotions and implicit attitudes are much more resistant to changes. But even if in some encompassing sense, most people therefore do have slightly sexist or racist beliefs, or often do not *really* have the 'beliefs' implicit in their u-value judgments, that does not seem to be very relevant for the theory of axiological uncertainty. Surely, you should not *perforce* satisfy EVM with respect to your most deep and subconscious attitudes and biases, simply because these may in a broad sense reflect your true credences. You should satisfy EVM with respect to the 'credences' that you take to be epistemically appropriate. And it seems that in focusing on your *judgments*, the judgment-based explication does indeed

[19]Cf. e.g. Greenwald and Banaji (1995), Gaertner and McLaughlin (1983) and Dovidio and Gaertner (2000) for relatively early works; Strohminger et al. (2014) for a recent methodological survey about implicit *moral* attitudes.

[20]Cf. e.g. Greene et al. (2001, 2107) on 'participants who judge in spite of their emotions'.

[21]Cf. e.g. the findings of Schwitzgebel and Rust (2014) and Schwitzgebel (2014), suggesting that more stringent moral views do not imply more stringent moral behaviour.

pick out such credences.

Furthermore, as I shall argue in the next section, the main reason why we should be interested in whether or not the weights in our representation theorems can be called 'credences' is that we want to establish a connection between the theory of axiological uncertainty on the one hand, and axiological epistemology on the other. We want these weights to reflect the entities that are relevant to epistemology. But at least *prima facie*, it does not seem that you are making an *epistemological* mistake if your u-value judgments reflect reasonable credences, but you do not have the emotions, attitudes, hormone levels or the behaviour that accord with these judgments. Plausibly, you are then weak-willed, or emotionally biased, or whatever. So it does not seem that if we want to make a connection to epistemology, we necessarily have to invoke a richer account of credences.

Finally, it is not clear whether it is even possible to provide a much richer and scientifically more informed explication of credences, while still guaranteeing their *quantitative* significance. One might perhaps attempt to find some quantitative empirical state that allows for a pertinent explication of credences – numbers of neuron-firings, or numbers of hormones, or whatever. But – for reasons similar to those I gave to the parallel idea concerning cardinal strengths of attitudes in section 3.4.2 – I very much doubt that this is possible. So the most promising way to do get an empirically richer explication, I take it, might be via representation theorems that feature conditions on other aspects of beliefs.[22] For example, we could introduce conditions

[22]The following proposal is the theorem-based scientifically informed account that I alluded to on page 173.

of the sort: 'if you have a favouring attitude for $\mathfrak{a}$ over $\mathfrak{b}$, then for any $\mathfrak{c}$ in $\mathcal{K}$ and any $p \in]0,1[$, you have a favouring attitude for $p\mathfrak{a} + (1-p)\mathfrak{c}$ over $p\mathfrak{b} + (1-p)\mathfrak{c}$'. These conditions may then imply an attitude-relative-'credence'-function. Perhaps we could do the same for your behaviour, your hormone level, and so on. And we could then take some weighted average of these various relativized 'credence'-functions to get your overall credence-function. Or we could assume bridging-principles of the form 'if you judge that $\mathfrak{a}$ is u-better than $\mathfrak{b}$, you have a favouring attitude for $\mathfrak{a}$ over $\mathfrak{b}$', guaranteeing that all these credence functions are the same. No one has done this, and it would certainly involve serious difficulties.[23] But even if that were possible, the complexity of such an account of credences would be a major drawback. I cannot see why, on the whole, anything should be gained by it.

In sum, Christensen's argument does neither show that the judgment-based explication – understood as an *explication* – is inadequate, nor does it point to a more promising account of credences. On the contrary, for our purposes, it would be unfortunate to adopt a much richer notion: it would grant weight to aspects that should not be normatively relevant, aspects that are not necessarily the business of epistemology, and it would clearly render our theory much more complex.

However, there is a third, related worry. One might worry not only that there are *additional* connections between our credences and our mental or physical states, but that sometimes the presumed connections with our u-value judgments do not even exist. As an analogue in decision theory, con-

[23]For example, as I pointed out on page 139, different people seem to react differently to the belief that one option is u-better than another; no such reaction might be a necessary condition for that belief, and the choice of any one might be somewhat arbitrary.

sider an objection by Lina Eriksson and Alan Hájek (2007):

> At the core of [decision-theoretic explications] is the idea that credences should somehow be defined or understood in terms of preferences. But credences and preferences are certainly separable in thought, and sometimes in practice. Imagine a Zen Buddhist monk who has credences but no preferences [i.e., 'is indifferent among all things']. [...] If the monk is conceptually possible, then any account that conceptually ties credences to preferences is refuted. [...] Or consider a chronic apathetic who has lost all his desires, but who has kept all his credences. To be sure, these characters are not recognizably like us, although some of us may approximate them over certain domains, and to the extent that we do, bets and preferences more generally ill-reflect our true credences. (2007, 194)

Others have provided similar arguments.[24] And a similar point seems to apply in our context: axiological credences and u-value judgments are 'separable in thought', it seems, since we can imagine someone who has various axiological beliefs but for some reason does not make any u-value judgments.

But what precisely is the argument here? It is worth saying that, *pace* Eriksson and Hájek, mere 'approximation' of monk-like apathy does not seem to distort preference-based explications. As far as I see, it will simply result in an attenuated utility function, and this seems precisely appropriate. Morevoer, it is important to note that our theorem will simply not *apply* to perfect monk-like agents. Karni and Schmeidler's Theorem 2 explicitly assumes that you *have* strict preferences; and the Subjectivist Expected Value Theorem presupposes that you *do* make (non-uniform) u-value judgments. So in this straightforward sense at least – and quite like Meacham and Weisberg's worry – the case of the monk cannot show that the explication is

[24]Cf. Christensen (2001, 363).

flawed *when* it applies.[25]

However, Eriksson and Hájek do seem to claim this, and they seem to raise a deeper worry. As applied to our context, their argument seems to be that the mere possibility for axiological credences and u-value judgments to come apart shows that any *definition* of the former in terms of the latter must be flawed. If they *can* come apart, the thought seems to be, then any connection between them will at best be a contingent matter and cannot be a matter of definition: 'if the monk is conceptually possible, then any account that *conceptually ties* credences to preferences is refuted'.

However, that axiological credences are possible even without u-value judgments does not show that there cannot be a conceptual connection between credences and u-value judgments in cases where the latter *are* present. That is simply a *non-sequitur*. A condition may be sufficient but not necessary for the application of a concept, and that might be a conceptual truth. For example, it may be a conceptual truth that *if* you have the ability repeatedly to perform the first prelude from the *Well-Tempered Clavier*, then you know how to perform it – even if you can also know how to perform it without having that ability (e.g., when your arms are broken). Our explication only says that *if* your u-value judgments are $\mathcal{C}_S$-conformable, you have the credences that make you satisfy EVM. For all that Eriksson and Hájek

[25]True, there is arguably a difference between the monk and someone with, say, intransitive preferences: the latter is (presumably) making some sort of mistake, whereas the monk need not make any mistake. So the monk shows that Karni and Schmeidler's Theorem 2 cannot ground a fully general normative theory. But I have already admitted that many constraints in the Subjectivist Expected Value Theorem are too strong as normative constraints, and are only plausible given certain restrictions or simplifications (e.g., the assumption that all axiologies are vNM-conformable, or fully comparable, or non-uniform).

say, that may well be a conceptual truth. Their worry may illustrate that we do have alternative criteria for ascribing beliefs, other than preferences or u-value judgments. But that was Christensen's point, and I have argued that it does not render our explication false or inadequate. And Eriksson and Hájek's worry does not show that preferences or u-value judgments are *no* conceptual criterion for ascribing beliefs.

It may be that Eriksson and Hájek have in mind yet another argument. Perhaps their thought is something like this: (i) we in any case need an explication of the perfect monk's credences; (ii) *that* explication cannot be preference based; (iii) whatever explication we give for the monk would also apply to all other agents; and therefore the preference based explication becomes redundant. However, for reasons I have outlined, it seems dubious to me that we *can* give an adequate explication of someone's axiological credences if he does not make any u-value judgments. And I also do not see why we should truly *need* that. It is true that our explication is restricted in application, and that it would be preferable to have a more general one. But unless these restrictions are too severe, and unless we have a better proposal, it seems unreasonable to dismiss our theory *simply because* it is less than fully general. So to turn the present Eriksson-Hájek-type worry into a full objection, one would have to argue more thoroughly for why we need an explication that does not presuppose u-value judgments. I shall turn to some such considerations in the next section. But as it stands, the present worry does again not 'refute' our explication.[26]

[26]There are other objections against preference-based explications in decision theory; Eriksson and Hájek (2007) offer an overview. Another worry that carries over to our context is that explicating credences in terms of preferences gets the order of explanation

4.3.2 Normative Relevance

Let me now address a different set of worries about the judgment-based explication. These worries concern what we may call the 'normative relevance' of the resulting theory.

A first such worry is that judgment-based EVM becomes *trivial*. A theory of axiological uncertainty should arguably be able to positively guide you in your decision making, and hence to constrain your judgments about what is u-best. But my argument seems to take your u-value judgments simply as *given* to *define* your credences and values. So it may seem that judgment-based EVM would always simply *guarantee* that your u-value judgments are correct, and thus cannot guide you in your decision making at all.[27]

However, this would be a misunderstanding. Judgment-based EVM *does* put firm constraints on you. These constraints are simply the *axioms*. In the version I have outlined, it is the conditions $\mathcal{C}_S$ that constrain you, and your u-value judgments cannot be true if they violate these conditions. By the same token, you may use these conditions to guide your decision making. Judgment-based EVM does not simply guarantee that your u-value

wrong (cf. Eriksson and Hájek (2007, 207f.)). I have replied to the equivalent objection concerning value on page 89; a similar reply could be given concerning credences. As far as I see, the remaining objections do not apply to my explication of axiological credences. In particular, a prominent objection against explicating degrees of belief in terms of preferences is that this explication is overly 'pragmatic', reducing a doxastic attitude (credence) to a conative one (preferences) (cf. Joyce (1999, 89ff.) and Eriksson and Hájek (2007, 194)). However, my explication reduces a doxastic attitude (axiological credences) to another doxastic attitude (u-value judgments). So this objection does not apply.

[27] Sepielli (2010, 169) raised this objection against preference-based explications in decision theory: '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.' This is 'highly counterintuitive' (2010, 168), he says.

judgments are correct.

What is true is that – as far as the correctness of your u-value judgments *as* u-value judgments is concerned – these axioms are all that judgment-based EVM implies. Normatively speaking, the version of judgment-based EVM I have outlined *reduces* to the conditions $\mathcal{C}_S$. Moreover, the conditions $\mathcal{C}_S$ only rule out particular *sets* of u-value judgments; they do not rule out any individual judgment considered by itself. Following Ramsey (1926, 41) and Savage (1954, 20), we might say they are only 'consistency constraints'. But I think that this would be somewhat misleading, since it is not straightforwardly *inconsistent* to violate our conditions $\mathcal{C}_S$. I shall thus call them *global constraints*. So normatively speaking – and as far as the correctness of your u-value judgments *as* u-value judgments is concerned – judgment-based EVM reduces to a set of global constraints. This does not mean that any set of judgments satisfying $\mathcal{C}_S$ is as good as any other *in all respects*. Presumably, some such judgments will reflect inadequate credences and thus be epistemically problematic. But if judgment-based EVM is true, then any set of u-value judgments satisfying $\mathcal{C}_S$ is correct *as* u-value judgments, *relative* to some set of credences. When we are trying to form u-value judgments, relative to our credences, we ultimately only have our global constraints $\mathcal{C}_S$.

Even if that is not outright trivial, perhaps some people will still find it disappointing. Meacham and Weisberg apparently do. They say that expected utility maximization* – the view that one ought to maximise the expectation of utility* relative to one's degrees of belief* (in the technical, representation-theorem-based senses of these terms) – is 'prescriptively useless' (2011, 656):

> adopting degrees of belief* and utilities* trivializes normative decision theory. Normative decision theory applies only to agents who have degrees of belief and utilities. But agents who have degrees of belief* and utilities* are automatically [...] expected utility maximizers with respect to them. If we take the 'degrees of belief' and 'utilities' that appear in normative decision theory to be degrees of belief* and utilities*, it will be true by definition that all agents subject to the norms of decision theory satisfy them. (2011, 653)

This passage is slightly misleading. On one reading, it is again simply false that 'normative decision theory' becomes 'useless' and 'trivialised' by adopting these explications. But Meacham and Weisberg are careful to note that only agents who *have* degrees of belief* and utilities* – i.e., satisfy the axioms of decision theory – are expected utility* maximisers. So presumably, their main claim (which they also make repeatedly), must be that a reduction to global constraints somehow makes decision theory 'uninteresting' (2011, 642; 645; 655; 661), if not outright trivial.

But it is not clear *why* a reduction to global constraints should make decision theory uninteresting, and unfortunately, Meacham and Weisberg do not say anything about why it should.[28] It is worth noting that the shift from an understanding of EVM on which it implies local constraints to one on which

[28]In a footnote, they consider the view that 'we should understand normative decision theory as the injunction to *have* degrees of belief and utilities', i.e. to satisfy the relevant conditions. But all they say in response is that 'this proposal represents a substantive shift in the content of normative decision theory. We're no longer dealing with the same norms, and these replacements can't do the same work as the originals. For example, normative decision theory is supposed to say which of an agent's options she ought to take. But the injunction to have degrees of belief and utilities will be silent on this, since every option will maximize expected utility relative to some pair of probability and utility functions' (2011, 648, n.15). This is merely to say that if we believed that EVM could imply local constraints, judgment-based EVM constitutes a shift in our understanding. But it still does not show why that should make it uninteresting, or why we thus have to 'lay the foundations of decision theory on firmer ground' than that provided by representation theorems.

it implies global constraints is much less significant than it may at first seem. On the one hand, at least if we accept a reflective-equilibrium view of how we ought to form judgments in ethics, we should be familiar with the idea that we ultimately only have some sort of global constraints. For example, suppose we understand EVM in terms of the Simple Explication, on which credences are given through mere introspection and which thus allows EVM to imply local constraints. And suppose your introspective list of numbers implies a set of u-value judgments that you find extremely implausible (e.g., giving far too much weight to average utilitarianism). Presumably, we would then not say that you should stick slavishly to these judgments. Rather, we would say that you have to adjust your 'credences' until they come into a reflective equilibrium with a set of judgments that you find plausible. So we would in any case, ultimately, embed EVM in something like global constraints. It is only that in judgment-based EVM this reduction to global constraints is, as it were, internal to the theory. But it is not clear why that should be problematic.[29]

On the other hand, nothing in the judgment-based explication implies that you may not *also*, say, consult your feelings of confidence about axiologies when determining a set of u-value judgments. For example, suppose that when trying to form a $\mathcal{C}_S$-conformable set of u-value judgments, you start with an intuitive list of numbers – the probability distribution P – that you take to reflect your credences. Surely, judgment-based EVM does not pro-

[29]In fact, at least in the present case, I think it is preferable to have the global constraints built into the theory. Note that according EVM understood in terms of the Simple Explication, your first (extremely implausible seeming) u-value judgments were *true* as u-value judgments. That seems to show that, in itself, this theory is of little practical significance.

hibit you to satisfy EVM with respect to these numbers. On the contrary, at least given something like Lewis's (1980) Principal Principle – roughly, the principle that your subjective probabilities conditional on objective chances should equal the objective chances – judgment-based EVM actually requires you to (at least initially) satisfy EVM with respect to P. On pain of inconsistency, your $\dot{\succeq}_U$ must then be equivalent to your $\succeq_U$ about $\mathcal{Q}_P$; and if it is, then according to the judgment-based explication, you are satisfying EVM with respect to P. If these intuitive numbers P lead you to a set of $\mathcal{C}_S$-conformable u-value judgments that you find plausible, there is nothing wrong with that. Judgment-based EVM will agree that these are your credences, and that your judgments are correct. So the reduction to global constraints does not prohibit the use of local constraints in practice.

But perhaps there is a more specific worry about why judgment-based EVM is 'uninteresting'. One might worry that we *cannot* make any u-value judgments without prior theoretic guidance. Indeed, this – one might argue – is why we need a theory of axiological uncertainty in the first place: without such theoretical guidance, we do not *know* how to make u-value judgments. So it may seem that a theory of axiological uncertainty must be able to constrain or guide us without presupposing that we can make any independent u-value judgments – otherwise the theory presupposes that we know already precisely what it was supposed to tell us.[30] But I do not think this presupposition of u-value judgments is very problematic. It is simply not true that we have no intuitions about u-value. On the contrary, at least after a moment of reflection, we are perfectly able to make intuitive u-value judgments even

[30]This is argued in Hedden (forthcoming, 13).

without relying or even knowing any philosophical theory about it – quite as we are able to make intuitive value judgments without relying on any theory. I said in chapter 1 that we need a theory of u-value. But I did not mean to suggest that we would be at complete loss about the u-value relation without such a theory. We would not. Global constraints may well provide us with all the guidance that we need.

One might also worry that a reduction to global constraints somehow makes EVM too permissive – that it is simply *wrong* that any judgments that satisfy something like 'consistency' constraints will be correct. But recall that I am only claiming that such judgments are correct *as* u-value judgments. They might of course still reflect unreasonable credence distributions, and then they might involve some epistemic mistake. In fact, note that on *any* theory of axiological uncertainty relying on subjective probabilities, your u-best option depends on your credences. So any such theory will not imply a local u-value judgment unless something about your credence-distribution is known. And in that sense, any such theory will reduce to 'global constraints' between your credences and the u-value relation. The difference is merely that judgment-based EVM takes the further information – about what your credences are – from your u-value judgments, rather than from your brute introspection, say. But again, it is not clear why that should be problematic.

So let me address a final objection about judgment-based EVM. If the normative implications of our view of axiological uncertainty boil down to the conditions $\mathcal{C}_S$, one may wonder why it should even *matter* whether we can interpret your u-value judgments in this way or another. As I myself seem to have admitted, whether or not we can interpret $\mathcal{C}_S$-conformable u-

value judgments as satisfying EVM is purely a matter of description. So why should we even care about that interpretation, and thus the relevant representation theorem and the bulk of my argument in this chapter? It seems that judgment-based EVM makes the *formula* for EVM entirely *redundant*.

I have some sympathy with this worry. As far as the truth of your u-value judgments *as* u-value judgments is concerned, I think the *formula* for EVM does indeed not add an additional norm; in principle, we could reduce our theory to the conditions $\mathcal{C}_S$. But again, even if no $\mathcal{C}_S$-conformable set of judgments violates norms of the theory of axiological uncertainty, some of them arguably do violate epistemic norms. Representation theorems are relevant because they allow us – or at least, promise to allow us – to express this connection between the theory of axiological uncertainty on the one hand, and axiological epistemology on the other, in a very neat way.

Most importantly, without representation theorems, it is not perspicuous from the axioms alone what precise means we have for stating that some $\mathcal{C}_S$-conformable sets of judgments are inadequate in light of our evidence. For example, it may be true that in light of our evidence concerning speciesism, judging that it is equally u-good to keep animals in species-appropriate conditions or not, or u-better to benefit people significantly more than animals, would be epistemically unreasonable. But it is not perspicuous how we can capture such facts in a more principled and unified way. Representation theorems show that if your judgments are $\mathcal{C}_S$-conformable, then to each axiology you attach a constant weight. So representation theorems – and *only* they – show that the talk of giving 'weights' to axiologies really does make sense. They allow us to say not only that you should not make this or that

u-value judgment concerning animal welfare, but more simply and generally, that you should not give much *weight* to speciesism. Or again, they at least promise to do so. We do not currently have any general principles about which evidence justifies which distribution of weights. But in some cases, we at least know that given this or that objection to theory T_i, one should give more weight to theory T_j than to theory T_i. And, given representation theorems, we know that we can translate such principled facts into claims about u-value judgments.

In principle, it is a matter of words whether or not we call these weights your 'credences'. We could have a theory that says simply that you should give each theory some weight, and that these weights should reflect our evidence in a particular way. But if we can interpret them as credences, that is a very natural and congenial way to establish the connection between norms of the theory of axiological uncertainty on the one hand, and epistemic norms on the other. If nothing else, it provides some flesh to an otherwise abstract claim about 'weights', and thus becomes explanatorily more powerful. It does justice to the intuitive idea that some axiologies are more likely than others, and that these likelihoods should be relevant in determining the u-value relation. And it allows us to say that the normative relationship between our evidence and these weights is one between evidence and a doxastic state, which may provide more unity to our overall picture of what an epistemic norm is, or what norms there exist.

In conclusion, even if the formula of EVM as the result of representation theorems does not add any additional constraints on your u-value judgments *as* u-value judgments, this does not show that such theorems, or the truth of

the formula for EVM, are irrelevant. Judgment-based EVM is neither trivial, nor uninteresting, nor redundant.[31]

4.3.3 Alternative Explications of Credences

Even if my arguments in the previous sections were sound, perhaps all other things being equal, some people might still find it preferable to have a theory of u-value that implies local constraints and is not judgment-based. So to round off the case for judgment-based EVM, let me briefly investigate again whether there are alternative explications. More specifically, let me reconsider the idea of a *preference*-based explication of axiological credences.

I have suggested in section 4.2.1 that in order to explicate axiological credences, we need to separate your axiological and non-axiological concerns. To see this with a concrete proposal, consider the following explication, along the lines of de Finetti (1980):

> **First Preference Explication**: That your credence in T_i is p_i means that you are indifferent between £p_i and a bet that gives you £1 if T_i is true.

Let me ignore the general problems of betting interpretations (such as the diminishing marginal value of money), and focus on the fact that this explication is *preference*-based. The explication is flawed if you care about axiological value, because the value of £1 is theory-dependent. Consider a

[31] We may note an additional point against the redundancy of representation theorems. In practice, instead of working with the axioms alone, it will often be easier to derive one's (provisional) probabilities and values from certain simple cases, and then apply the formula for EVM to see what the axioms imply in other cases. This is what I did in section 4.2.3. In this respect, representation theorems are also helpful for practical purposes.

uniform axiology on which all outcomes are equally good, and some utilitarian theory. If the uniform theory is true, then it does not matter morally whether or not you get the prize, but if the utilitarian theory is true, you may well do some extra good with the money. So insofar as you care about moral value, you will prefer a £1-bet on utilitarianism to a £1-bet on the uniform theory, *no matter* what your credences are (as long as you have some nonzero credence in the former). And the same general problem arises among non-uniform axiologies. So insofar as you care about moral value, its theory-dependence will distort this First Preference Explication.

But instead of bracketing your non-axiological concerns and focusing on your u-value judgments, let us now try to bracket your concern for moral value and focus on your prudential concerns specifically. We might thus suggest something like the

> **Second Preference Explication**: That your credence in T_i is p_i means that you are indifferent between £p_i and a £1-bet on T_i, insofar as you are concerned only with the *prudential* value of these bets for you.

However, this account still suffers from the problem of theory-dependence: even the prudential value of £1 can depend on the true axiology. This will be the case, at least, for axiologies on which your wellbeing has value. All these theories come in different forms, depending on what theory of wellbeing they presuppose. They may either presuppose an objective list theory of wellbeing, or hedonism, or a desire-based theory – or perhaps some still other view. And the prudential value of £1 might differ on these theories, say, because you can easily buy some pleasure with that money but not true, valuable friendship.

To the extent that the prudential value of outcomes depends on the true axiology, the Second Preference Explication is distorted.

To remedy this problem, we might try to bracket this theory-dependence of value. We might suggest something along the lines of the

> **Third Preference Explication**: That your credence in T_i is p_i means that you are indifferent between £p_i and a £1-bet on T_i, insofar as you are concerned only with the prudential value of these bets for you, and assume that the prudential value of money is the same regardless of which axiology is true.

This proposal would in principle overcome the worry about theory-dependent value. What is slightly unfortunate is that it reintroduces preferences of the kind we encountered with Savage's theory in section 2.2.1. Savage's theory presupposed that you have preferences about an act that leads to a perfectly sunny hike even when the state of nature is 'rain'; it presupposes that you have preferences about impossible state-outcome combinations. The same is true of the present account. Prudential value *is* axiology-dependent, but – on the Third Preference Explication – you must assume that it is not; you need to have preferences about impossible theory-outcome combinations. And this is not just a technical assumption, but features as the very core of the proposal.

But there is another, and more fundamental problem with any of these alternative explications. The problem is how we individuate the relevant theory 'T_i' – how we determine *which* theory we are referring to when explicating your credence in theory 'T_i'. At least under Non-Substantial Absolutism, we

cannot simply ask how much you would bet, say, on 'utilitarianism', because there are infinitely many axiologies corresponding to each ordering. This means that we need a criterion for how the theories you have credence in *compare* to one another. Unless we have such a criterion, we do not have an explication of credences that allows us to determine the expected value of your options. But under Non-Substantial Absolutism, we can individuate an axiology only by how it enters the u-value relation. And consequently, even this preference-based explication would presuppose that we can make u-value judgments – at least, u-value judgments about $\mathcal{Q}$.[32] So I cannot see why it should be preferable to explicate credences via preferences among highly counterfactual bets, in line with the Third Preference Explication, rather than – much more straightforwardly – via u-value judgments.

Conclusion

Let me draw a conclusion. As I mentioned on page 195, I have been presupposing Non-Substantial Absolutism throughout this discussion. Under that assumption, for the reasons I gave in this last section, I conjecture that we cannot provide a preference-based explication of axiological credences that comes close to the simplicity of the judgment-based explication and would be preferable to it. For the reasons I outlined in previous sections, I am sceptical about taking axiological credences as primitives, obtaining them through pure introspection, or about providing an empirically much richer or scientifically more informed explication of credences. To that extent, I am more sceptical than the representation-theorem-sceptics may be

[32]Note that the same individuation-problem arises for the Simple Explication.

about alternatives to the judgment-based explication. I also argued that the judgment-based explication is indeed a possible and good explication: it is not somehow wrong or purely stipulative, and it does neither trivialise EVM, nor make it uninteresting or redundant. So I am less sceptical than the representation-theorem-sceptics may be about the judgment-based explication itself. Whether representation theorems can ground EVM is a large question, and I cannot claim to have proved their ultimate significance beyond doubt. But I hope I have done enough to show that at least in the theory of axiological uncertainty, we should be much less quick in dismissing the significance of representation theorems than some recent sceptics have been. At least under the assumption of Non-Substantial Absolutism, judgment-based EVM may be the best form of EVM that there is.

4.4 Further Explorations: Alternative Explications and Eschewing Credences

To end this chapter, let me again add some notes that go beyond the main theory I have been exploring. One major open question is that of how to explicate credences under alternative accounts of intertheoretic comparisons. I shall consider this question in section 4.4.1. In section 4.4.2, I shall outline an axiomatisation of a view that eschews the concept of credences altogether.

4.4.1 Credences under Alternative Accounts of Axiologies

How could we explicate axiological credences under accounts of intertheoretic comparisons other than Non-Substantial Absolutism? In an interesting sense, the theory of Non-Substantial Absolutism is the most general theory, as it can easily be extended to cover alternative accounts. More specifically, to explicate credences under alternative accounts of intertheoretic comparisons, we could simply extend the Subjectivist Expected Value Theorem in the ways in which I have extended the Expected Value Theorem in chapter 3. For example, on page 124 I introduced the

> **Best/Worst Comparison Principle**: For all T_i and T_j, $\boldsymbol{a}_{[T_i,a_i^w]} \sim_U \boldsymbol{a}_{[T_j,a_j^w]}$ and $\boldsymbol{a}_{[T_i,a_i^b]} \sim_U \boldsymbol{a}_{[T_j,a_j^b]}$.

This principle will make the Expected Value Theorem imply EVM with a given comparativist view about how axiologies compare. I stated other such principles, as well as principles that will make the Expected Value Theorem imply EVM under a strength-sensitive FA-account (viz., the Dyadic Attitude Principle and the Monadic Attitude Principle).

These principles could straightforwardly be introduced into the Subjectivist Expected Value Theorem, as conditions on your u-value judgments about $\mathcal{Q}$. They would ensure that the value-functions that you assign to your theories satisfy whatever constraints we think they should satisfy – e.g., the Best/Worst Normalisation principle, or the constraint that value-facts correspond to pertinent facts about attitudes. If we do add them to the Subjectivist Expected Value Theorem, that extended theorem will al-

low us to give a judgment-based explication of credences, and thus defend judgment-based EVM, even for these alternative accounts of intertheoretic comparisons.

However, under alternative accounts of intertheoretic comparisons, alternatives to the judgment-based explication may be somewhat more promising. Consider again the Third Preference Explication from page 217. The main problem with this explication was that in order to individuate T_i, we would again have to refer to the u-value relation and so it would not have any clear advantage over the judgment-based explication.

However, under alternative accounts of what axiologies are, this may not be so. For example, suppose Comparativism is true, and axiologies are merely orderings. In that case, we can individuate T_i simply by the ordering it implies. So we might give the Third Preference Explication or another, slightly more sophisticated preference-based explication that avoids its more familiar problems (like the diminishing marginal value of money). If we have a specific comparison principle, we might then be able to tell you which of your options is u-best without presupposing that you make any u-value judgments. Something similar might be true with respect to forms of Absolutism based on FA-accounts. If intertheoretic comparisons are grounded in facts about fitting attitudes, we might individuate a theory T_i as 'the theory that implies the ordering O_i and the set $\mathcal{A}_i$ of facts about attitudes'. We might then give something like the Third Preference Explication to explicate your credences, and thus be able to tell you which of your options is u-best without presupposing that you make any u-value judgments.

However, it is not at all clear to me whether that would be preferable. If we use something like the Third Preference Explication, we have to presuppose that you have preferences about counterfactual bets on axiological orderings, or that you have beliefs about fitting strengths of attitudes, as well as preferences concerning counterfactual bets involving them. I am not at all convinced that this would be preferable to a judgment-based explication. On the contrary, it seems more common for people to have beliefs about u-value relations rather than, say, about precisely which attitudes would be fitting. And if my arguments from section 4.3 were sound, the reduction of judgment-based EVM to global constraints is much less problematic than people have suggested. So I am inclined to think that even under alternative accounts of intertheoretic comparisons, judgment-based EVM might be the best form of EVM.

4.4.2 Weighted Value Maximisation

I have argued that, *ceteris paribus*, it is preferable to interpret the weights that axiologies have in determining the u-value relation as probabilities or credences. By the same token, I think it is preferable to distinguish between the probability we assign to an axiology, and the value-function of this axiology – although both ultimately appear simply as the *weight* that this axiology has in determining the u-value relation. If we *can* make this distinction, I take it, that makes our theory explanatorily more powerful. We can then give a more specific explanation of what these 'weights' are. As I said, that explanation does justice to the intuitively plausible idea that some

axiologies are more likely than others and that these likelihoods should be relevant in determining the u-value relation; and it also does justice to the idea – which I also take to be intuitively plausible – that different axiolgies assign different values to outcomes, and that these values too should be relevant in determining the u-value relation. I have also argued that we can indeed make that distinction, both on a technical and on a conceptual level.

However, I have not claimed that my arguments to that effect were absolutely conclusive. So now that we have the formal framework in place, it is worth mentioning briefly that the alternative view, which eschews the distinction between probabilities and values altogether, can be axiomatised straightforwardly in that framework. We can do that with a precise analogue of the Expected Value Theorem, concerning $\mathcal{K}$ instead of $\mathcal{Q}$. So let me introduce the

> **Pareto Condition for** $\mathcal{K}$: If for some $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$, $K_i(\mathfrak{a}) \sim_i K_i(\mathfrak{b})$ for all T_i, then $\mathfrak{a} \dot{\sim}_U \mathfrak{b}$; and if $K_i(\mathfrak{a}) \succeq_i K_i(\mathfrak{b})$ for all T_i, and $K_j(\mathfrak{a}) \succ_j K_j(\mathfrak{b})$ for some T_j, then $\mathfrak{a} \dot{\succ}_U \mathfrak{b}$.

For two functions $u : Z \to \mathbb{R}$ and $v : Z \to \mathbb{R}$, say that v is a *positive unit-comparable transformation of* u if there is an $s \in \mathbb{R}, s > 0$, and a function $t : I \to \mathbb{R}$, such that $v(i, x) = su(i, x)+t(i)$ for all i in I and x in X. Given the assumption that each of your $\succeq_i$ is vNM-conformable, the following theorem holds:

> **Subjectivist Weighted Value Theorem**: If your $\dot{\succeq}_U$ is vNM-conformable and satisfies the Pareto Condition for $\mathcal{K}$ with respect to your $\succeq_i$, there is a theory-dependent utility function $u : Z \to \mathbb{R}$ that represents each axiology

cardinally, and is unique up to positive unit-comparable transformation, such that for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$,

$$\mathfrak{a} \dot{\succeq}_U \mathfrak{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \mathfrak{a}(i,x)u(i,x) \geq \sum_{(i,x)\in Z} \mathfrak{b}(i,x)u(i,x). \tag{4.7}$$

This theorem is put in terms of a theory-dependent *utility* function. It does *not* say that we can treat the respective functions $u(i,\cdot)$ as products p_iG_i of your credence in T_i with T_i's value-function. So (4.7) does not express EVM. I thus say it expresses *Weighted Value Maximisation*. What the Subjectivist Weighted Value Theorem says is that, if your u-value judgments satisfy the von Neumann-Morgenstern axioms and the Pareto Condition for $\mathcal{K}$, then to each axiology you are attaching a constant weight, in the form of a utility function.

Again, this theorem is a non-normative claim. It says that if your u-value judgments satisfy the relevant conditions, then you are, as a matter of fact, following (4.7). We can turn this into an argument for the normative truth of (4.7) by claiming that your u-value judgments ought to satisfy these conditions. However, that your $\dot{\succeq}_U$ *should* satisfy the Pareto Condition for $\mathcal{K}$ is only plausible if you do not rule out the truth of any axiology under consideration completely – or intuitively, if you assign 'nonzero credence' to all axiologies. If there is some T_j whose truth you rule out completely, then the second clause of the condition is implausible with regards to that theory. So to ground a normative argument on the Subjectivist Weighted Value Theorem, we would have to assume that the set of theories under consideration is such that you do not completely rule out the truth of any of

them. The obvious way to do so would be to restrict the theories to those for which your $\dot{\succeq}_U$ does, as a matter of fact, satisfy the Pareto Condition for $\mathcal{K}$.

This is a genuine theory of axiological uncertainty. To my knowledge, no one has defended that theory so far. Philosophers writing on axiological or normative uncertainty have always focused on *expected* value maximisation. For the reasons I mentioned, I think that if we *can* establish it, EVM is preferable to Weighted Value Maximisation. But once we acknowledge that the question of what credences are is not entirely unproblematic, Weighted Value Maximisation becomes a relevant alternative.

Conclusion

In this chapter, I examined the question of how to understand axiological credences. I argued that we should not take such credences as undefined primitives, suggested that we should not accept an explication based on pure introspection, and expressed doubts about empirically very rich explications or explications on the basis of preferences. As an alternative, and based on a representation theorem by Karni and Schmeidler, I then suggested the judgment-based explication. According to this explication, your credence in an axiology is the weight you give this axiology in your u-value judgments.

I argued that this explication is not somehow wrong or purely stipulative, and that it does neither trivialise EVM, nor make it uninteresting or redundant. If I am right, judgment-based EVM may be the best form of EVM that

there is – at least, but perhaps not only, on Non-Substantial Absolutism.

Chapter 5

The Problem of Incompleteness

Introduction

I have argued in chapter 2 that – barring the problem of intertehoretic comparisons and taking axiological probabilities as primitives – representation theorems from decision theory can help us formulate and defend EVM. In chapter 3, I have argued that intertheoretic comparisons are generally possible but that, at least given Absolutism, some plausible axiologies are not fully comparable and the u-value relation is not complete if it ranges over them. In the last chapter, I have argued that we cannot take axiological credences as a primitive and provided an explication for them. The present chapter makes some steps towards bringing together the upshots from the last three chapters. That is, it extends the Expected Value Theorem by dropping the completeness assumption and the presupposition of axiological probabilities as given primitives.

Unfortunately, although there are numerous results in expected utility

theory without the completeness assumption, and numerous results in state-dependent expected utility theory, there is – as far as I see – no pertinent existing result that combines these two fields: no representation theorem implying a unique separation of subjective probabilities and utilities for incomplete, state-dependent preferences.[1] So in the present chapter, I shall need some space to develop a result that at least goes in this direction. This is what I shall do in section 5.1. The emerging theorem will be the most encompassing result of this thesis.

Having dropped the Completeness Condition, we will then be in a better position to assess the prospects of understanding EVM as a general theory of *moral* uncertainty – rather than simply a theory of axiological uncertainty. So in section 5.2, I shall explore whether we can extend the framework of this thesis to cover moral uncertainty generally. I will conclude that the prospects of this extension look rather bleak.

Section 5.3 will then draw a conclusion. It will briefly summarise the main positive and negative upshots of this thesis, and indicate some possible paths for further research.

5.1 Axiomatising Incomplete U-Value Relations

In this section, I will first introduce a representation theorem due to Robert Nau (2006). I shall do that in section 5.1.1. Nau axiomatised incomplete

[1] I thank Edi Karni for confirming this to me (as of fall 2013).

preferences by use of state-dependent utility functions. However, Nau's theorem does not imply a unique separation of utilities and probabilities, and since the overall argument I intend to give depends on such a separation, I have to supplement that result with additional assumptions. So in section 5.1.2, I shall again simply assume probabilities as primitives, and turn Nau's result into an axiomatisation of an incomplete u-value relation with given axiological probabilities. Building on that, in section 5.1.3, I introduce an additional axiom and state a representation theorem that implies unique subjective probabilities and a relevantly unique set of state-dependent utility functions.

5.1.1 Nau's Theorem

For simplicity (and as I have done with Karni and Schmeidler's Theorem 2), I shall simply state Nau's result with respect to the sets of options I have already introduced, rather than defining new options that explicitly refer to 'states' as opposed to axiologies. His result again concerns prospects that do not presuppose an implicit probability distribution over states; it concerns the set $\mathcal{K}$. Let $\mathfrak{a}_x$ denote the constant act that yields consequence x with probability 1 for every i – that is, $\mathfrak{a}_x(i, y) = 1$ if $y = x$, and $\mathfrak{a}_x(i, y) = 0$ if $y \neq x$, for all i. We shall assume that X contains a worst and a best outcome, labeled $\overline{x}$ and $\underline{x}$ respectively; and we shall assume slightly different independence- and continuity constraints from those we have been using so far. That is, our axioms for a reflexive binary relation $\dot{\succeq}$ on $\mathcal{K}$ are defined as follows:

Transitivity$_{\mathcal{K}}$: If $\mathfrak{a} \dot{\succeq} \mathfrak{b}$ and $\mathfrak{b} \dot{\succeq} \mathfrak{c}$, then $\mathfrak{a} \dot{\succeq} \mathfrak{c}$;

Mixture-Independence$_{\mathcal{K}}$: $\mathfrak{a} \dot{\succeq} \mathfrak{b}$ if and only if $p\mathfrak{a} + (1-p)\mathfrak{c} \dot{\succeq} p\mathfrak{b} + (1-p)\mathfrak{c}$, for all $p \in]0,1[$ and all $\mathfrak{c}$ in $\mathcal{K}$;

Sequence-Continuity$_{\mathcal{K}}$: if $\{\mathfrak{a}_n\}$ and $\{\mathfrak{b}_n\}$ are convergent sequences such that $\mathfrak{a}_n \dot{\succeq} \mathfrak{b}_n$ for all n, then $\lim(\mathfrak{a}_n) \dot{\succeq} \lim(\mathfrak{b}_n)$;

Existence of Best and Worst$_{\mathcal{K}}$: For all $x \in X \setminus \{\overline{x}, \underline{x}\}$, $\mathfrak{a}_{\overline{x}} \dot{\succeq} \mathfrak{a}_x$, and $\mathfrak{a}_x \dot{\succeq} \mathfrak{a}_{\underline{x}}$;

Non-Triviality$_{\mathcal{K}}$: $\mathfrak{a}_{\overline{x}} \dot{\succ} \mathfrak{a}_{\underline{x}}$ (i.e., not $\mathfrak{a}_{\underline{x}} \dot{\succeq} \mathfrak{a}_{\overline{x}}$).

For simplicity, I shall use the term '*N-conformable*' for all reflexive binary relations on $\mathcal{K}$ that satisfy these axioms (for some $\overline{x}$ and $\underline{x}$ in X). Note that Existence of Best and Worst implies that $\dot{\succeq}$ cannot be *totally* incomplete. But apart from this assumption (and from what follows from it together with the other axioms), these conditions do not assume that $\dot{\succeq}$ is complete. I shall briefly comment on these axioms when we apply them to the u-value relation. But let me first state Nau's theorem.

Instead of defining a uniqueness criterion for our utility functions – which is somewhat more complicated if we are dealing with incomplete relations[2] – we shall simply introduce a *normalisation* for these functions, and then focus on normalised functions only. So define a normalized set of state-dependent

[2]Cf. Nau (2006) for a fully spelt out uniqueness criterion.

utility functions

$$W^* = \{w : Z \to \mathbb{R} \mid w(i,\underline{x}) = 0 \ \ \forall i \in I;\ 0 \leq \sum_{i \in I} w(i,x) \leq 1$$
$$\forall\, x \in X \setminus \{\overline{x}, \underline{x}\}; \sum_{i \in I} w(i,\overline{x}) = 1\}. \tag{5.1}$$

Call a collection of preferences $\{\mathfrak{a}_n \succsim \mathfrak{b}_n\}$ a *basis* for $\succsim$ under an axiom system if every preference $\mathfrak{a} \succsim \mathfrak{b}$ can be deduced from $\{\mathfrak{a}_n \succsim \mathfrak{b}_n\}$ by application of these axioms. Finally, for simplicity, for some state-dependent utility function $w : Z \to \mathbb{R}$, define

$$\mathbf{U}_w(\mathfrak{a}) = \sum_{(i,x) \in Z} \mathfrak{a}(i,x) w(i,x). \tag{5.2}$$

Given these definitions, Nau (2006, Theorem 2) proves

> **Nau's Theorem**: If a reflexive binary relation $\succsim$ on $\mathcal{K}$ is N-conformable, there exists a nonempty closed convex set $W \subset W^*$ of state-dependent utility functions, such that for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$,
>
> $$\mathfrak{a} \succsim \mathfrak{b} \quad \text{iff} \quad \sum_{(i,x) \in Z} \mathfrak{a}(i,x) w(i,x) \geq \sum_{(i,x) \in Z} \mathfrak{b}(i,x) w(i,x) \quad \forall w \in W. \tag{5.3}$$
>
> In particular, if $\{\mathfrak{a}_n \succsim \mathfrak{b}_n\}$ is a basis for $\succsim$ under these axioms, then W is the set of $w \in W^*$ satisfying $\{\mathbf{U}_w(\mathfrak{a}_n) \geq \mathbf{U}_w(\mathfrak{b}_n)\}$.

(5.3) is a representation in terms of a *set* of utility functions. This straightforwardly allows for incompleteness in $\succsim$: neither $\mathfrak{a} \succsim \mathfrak{b}$ nor $\mathfrak{b} \succsim \mathfrak{a}$ are true if there are u and v in W with $\mathbf{U}_u(\mathfrak{a}) > \mathbf{U}_u(\mathfrak{b})$ and $\mathbf{U}_v(\mathfrak{b}) > \mathbf{U}_v(\mathfrak{a})$.

However, note that in this result, the function w is not separated into a probability and a utility function. So the representation in (5.3) is not an *expected utility* representation; additional assumptions are required to assure that (5.3) can take the form of an expected utility representation with relevantly unique probabilities and utilities. Nau (2006) proceeds to introducing an assumption that guarantees the state-*in*dependence of preferences. Such an assumption is unfortunate for our purposes. So let me proceed differently.

5.1.2 The Expected Value Theorem for Incompleteness

To apply Nau's result to our context, I shall first turn it into a theorem that assumes probabilities as given primitives, as the Expected Value Theorem did. To that end, let me again consider the set $\mathcal{Q}$, and suppose $\succeq$ is a reflexive binary relation on $\mathcal{Q}$. We shall now assume that Z contains a best and a worst state-outcome pair, labeled $(\overline{i,x})$ and $(\underline{i,x})$ respectively. Given this assumption, we can define a normalised set of utility functions

$$U^* = \{u : Z \to \mathbb{R} \mid u(\underline{i,x}) = 0;\ 0 \leq u(i,x) \leq 1\ \ \forall\ (i,x) \in Z \setminus \{(\overline{i,x}), (\underline{i,x})\};\ u(\overline{i,x}) = 1\}. \tag{5.4}$$

Similarly, for some state-dependent utility function $u : Z \to \mathbb{R}$, define

$$\mathbf{U}_u(\boldsymbol{a}) = \sum_{(i,x)\in Z} \boldsymbol{a}(i,x)w(i,x). \tag{5.5}$$

With the option $\boldsymbol{a}_{(i,x)}$ as defined on page 124,[3] we shall define the relevant axioms governing $\succeq$ as follows:

Transitivity$_{\mathcal{Q}}$: if $\boldsymbol{a} \succeq \boldsymbol{b}$ and $\boldsymbol{b} \succeq \boldsymbol{c}$, then $\boldsymbol{a} \succeq \boldsymbol{c}$;

Mixture-Independence$_{\mathcal{Q}}$: $\boldsymbol{a} \succeq \boldsymbol{b}$ if and only if $p\boldsymbol{a} + (1-p)\boldsymbol{c} \succeq p\boldsymbol{b} + (1-p)\boldsymbol{c}$, for all $p \in]0,1[$ and all $\boldsymbol{c}$ in $\mathcal{Q}$;

Sequence-Continuity$_{\mathcal{Q}}$: if $\{\boldsymbol{a}_n\}$ and $\{\boldsymbol{b}_n\}$ are convergent sequences such that $\boldsymbol{a}_n \succeq \boldsymbol{b}_n$ for all n, then $\lim(\boldsymbol{a}_n) \succeq \lim(\boldsymbol{b}_n)$;

Existence of Best and Worst$_{\mathcal{Q}}$: For all $(i,x) \in Z \setminus \{(\overline{i,x}), (\underline{i,x})\}$, $\boldsymbol{a}_{(\overline{i,x})} \succeq \boldsymbol{a}_{(i,x)}$, and $\boldsymbol{a}_{(i,x)} \succeq \boldsymbol{a}_{(\underline{i,x})}$;

Non-Triviality$_{\mathcal{Q}}$: $\boldsymbol{a}_{(\overline{i,x})} \succ \boldsymbol{a}_{(\underline{i,x})}$ (i.e., not $\boldsymbol{a}_{(\underline{i,x})} \succeq \boldsymbol{a}_{(\overline{i,x})}$).

For simplicity, I shall again use the term '*N-conformable*' for all reflexive binary relations on $\mathcal{Q}$ that satisfy these axioms (for some $(\overline{i,x})$ and $(\underline{i,x})$ in Z). Nau's Theorem then implies

Theorem 5.1: If a reflexive binary relation $\succeq$ on $\mathcal{Q}$ is N-conformable, there exists a nonempty closed convex set $U \subset U^*$ of state-dependent-utility functions, such that for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$,

$$\boldsymbol{a} \succeq \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \boldsymbol{a}(i,x)u(i,x) \geq \sum_{(i,x)\in Z} \boldsymbol{b}(i,x)u(i,x) \quad \forall u \in U. \tag{5.6}$$

In particular, if $\{\boldsymbol{a}_n \succeq \boldsymbol{b}_n\}$ is a basis for $\succeq$ under these axioms, then U is the set of $u \in U^*$ satisfying $\{\mathbf{U}_u(\boldsymbol{a}_n) \geq \mathbf{U}_u(\boldsymbol{b}_n)\}$.

This is a fairly immediate implication of Nau's Theorem. The proof is given in appendix 6.3.

[3] $\boldsymbol{a}_{(i,x)}(i,x) = 1$.

To apply this theorem to the context of axiological uncertainty, I shall now assume that all axiologies under consideration satisfy the following conditions:

Transitivity$_{\mathcal{O}}$: if $a \succeq b$ and $b \succeq c$, then $a \succeq c$;

Mixture-Independence$_{\mathcal{O}}$: $a \succeq b$ if and only if $pa+(1-p)c \succeq pb+(1-p)c$, for all $p \in]0,1[$ and all c in $\mathcal{O}$;

Sequence-Continuity$_{\mathcal{O}}$: if $\{a_n\}$ and $\{b_n\}$ are convergent sequences such that $a_n \succeq b_n$ for all n, then $\lim(a_n) \succeq \lim(b_n)$.

If a binary relation on $\mathcal{O}$ satisfies these conditions, I shall say it is *N^*-conformable*.

If we want to represent an axiology on which outcomes are less than fully comparable, we cannot interpret it as implying determinate value difference ratios of the form 'the value difference between x and y is n times as great as the value difference between z and t'. Similarly, if we want to represent two axiologies that are not fully comparable, we cannot interpret them as implying determinate intertheoretic value difference ratios of the form 'the value difference between x and y, according to T_i, is n times as great as the value difference between z and t, according to T_j'. Instead, I take it, as far as intratheoretic comparisons are concerned we are interested in judgments like

(A) according to T_i, the value difference between x and y is at least (or at most) n times as great as the value difference between z and t.

I shall call such judgments *rough cardinal intratheoretic comparisons (of*

value). As far as intertheoretic comparisons are concerned, we are interested in judgments like

(B) the value difference between x and y, according to T_i, is at least (or at most) n times as great as the value difference between z and t, according to T_j.

Again, our theorem guarantees that we can explicate level-comparisons too. That is, we can actually explicate more general statements like

(C) the difference between the value of x, according to T_i, and the value of y, according to T_j, is at least (or at most) n times as great as the difference between the value of z, according to T_h, and the value of t, according to T_k.

I shall call such statements *rough crosscutting cardinal intertheoretic comparisons (of value).* To state my explications, let U be a nonempty closed convex set of utility functions $u : X \to \mathbb{R}$. Suppose that for some axiology T_i, and for all a and b in $\mathcal{O}$,

$$a \succeq_i b \quad \text{iff} \quad \sum_{x \in X} a(x)u(x) \geq \sum_{x \in X} b(x)u(x) \quad \forall u \in U. \tag{5.7}$$

I shall then say that U *represents* T_i *ordinally.* And I shall use equivalent, self-explanatory definitions for the claims that some nonempty closed convex set U of theory-dependent utility functions represents the u-value relation $\succeq_U$ ordinally, or (in the next section) that some pair (U, P) represents your u-value relation $\dot{\succeq}_U$ ordinally. Now suppose that for some nonempty closed convex set U of utility functions, and some axiology T_i, the rough cardinal

intratheoretic comparisons between certain outcomes are true according to T_i if and only if they are true for all functions in U. I shall then say that U *represents* T_i *cardinally*. According to my explication, if a nonempty closed convex set $U \subset U^*$ of utility functions represents an axiology ordinally, then it also represents it cardinally. Similarly, suppose that some nonempty closed convex set U of theory-dependent utility functions is such that, for each axiology T_i, the set $U_i = \{u(i, \cdot) \mid u \in U\}$ represents that axiology cardinally. I shall then say that U *represents each axiology cardinally*. Suppose that for some nonempty closed convex set U of theory-dependent utility functions, the rough crosscutting cardinal intertheoretic comparisons are true, according to our theories, if and only if they are true for all functions in U. I shall then say that U *jointly represents all axiologies cardinally*. According to my explication, if a nonempty closed convex set $U \subset U^*$ of theory-dependent utility functions represents the u-value relation ordinally, and represents each axiology cardinally, then it jointly represents all axiologies cardinally. If that is so, we can assume, say, that our theories are represented by the set of value-functions $\mathcal{G} = \{G : Z \to \mathbb{R} \mid G = u, u \in U\}$ of theory-dependent utility functions, and accordingly, that each theory T_i is represented by the set of value-functions $\mathcal{G}_i = \{G_i : X \to \mathbb{R} \mid G_i = G(i, \cdot) = u(i, \cdot), u \in U\}$.

To apply these explications, let me add a third clause to the Pareto condition I introduced in chapter 2, and state the

Strong Pareto Condition: For any probability distribution P on I, and for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^P$, if $H_i(\boldsymbol{a}) \sim_i H_i(\boldsymbol{b})$ for all T_i with $P(i) > 0$, then $\boldsymbol{a} \sim_U \boldsymbol{b}$; if $H_i(\boldsymbol{a}) \succeq_i H_i(\boldsymbol{b})$ for all T_i with $P(i) > 0$ and $H_j(\boldsymbol{a}) \succ_j H_j(\boldsymbol{b})$ for

some T_j with $P(j) > 0$, then $\boldsymbol{a} \succ_U \boldsymbol{b}$; and if for some T_j with $P(j) > 0$, $H_i(\boldsymbol{a}) \sim_i H_i(\boldsymbol{b})$ for all $T_i \neq T_j$ with $P(i) > 0$, then $\boldsymbol{a} \succeq_U \boldsymbol{b}$ only if $H_j(\boldsymbol{a}) \succeq_j H_j(\boldsymbol{b})$.

The third clause of this condition guarantees that the value-functions in our representation will not represent *sharpenings* of our axiologies.

Given our explications, and the assumption that all axiologies under consideration are N*-conformable, the following theorem holds:

Expected Value Theorem for Incompleteness: If the u-value relation $\succeq_U$ is N-conformable and satisfies the Strong Pareto Condition, then for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$,

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \boldsymbol{a}(i,x)G_i(x) \geq \sum_{(i,x)\in Z} \boldsymbol{b}(i,x)G_i(x) \quad \forall G \in \mathcal{G}. \tag{5.8}$$

Note that Existence of Best and Worst implies that our axiologies cannot be fully incomparable. Sequence-Continuity implies that they cannot compare in a lexical way. So as I indicated in section 3.4.2, this theorem cannot apply to axiologies that are comparable or incomparable in such ways. But again, I take it that such theories are extreme and comparatively very implausible; and to that extent at least, this is a much less severe restriction than that imposed by Completeness.

What is important is that the theory expressed in (5.8) can range over axiologies that are only roughly comparable. In fact, it allows us to represent both (non-radical) intra- and intertheoretic incomparability. If a theory T_i features some intratheoretic incomparability, then some two functions in $\mathcal{G}_i$

are not positive affine transformations of each other. If no axiology under consideration features intratheoretic incomparability, then for all i in I, all functions in $\mathcal{G}_i$ are positive affine transformations of each other. There may then still be *inter*theoretic incomparability. In that case, for at least some theory T_i, not all functions in $\mathcal{G}_i$ are the same – that is, some of the functions in $\mathcal{G}_i$ are *nontrivial* positive affine transformations of each other. This is a significant advantage over the Expected Value Theorem.

5.1.3 The Subjectivist Expected Value Theorem for Incompleteness

Let me now make some first steps towards axiomatising incomplete u-value relations without given probabilities. First, I have to reemphasise a point I noted on page 188. There are two reasons for why the u-value relation relative to your credences may be incomplete. You may have credence in theories that give rise to incomparability in values (either intra- or intertheoretically); or you may have fuzzy credences. As I mentioned there, I shall ignore the latter case. Fuzzy credences present a general problem that has nothing to do with axiological credences specifically. So for simplicity, I shall simply be assuming that your credences are sharp, and that the incompleteness of your u-value relation is entirely due to an incomparability in your values.

To model this, say that a probability distribution P over I is *positive* if $P(i) > 0$ for all i in I. Using the function $L : \mathcal{Q}^+ \to \mathcal{K}$ (from page 184), I shall assume the

Reduction Axiom: there is a positive probability distribution P over I

such that, if $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^P$, $\boldsymbol{a} \succeq \boldsymbol{b}$ if and only if $L(\boldsymbol{a}) \dot{\succeq} L(\boldsymbol{b})$.

If your preferences on $\mathcal{K}$ satisfy the Reduction Axiom, they are exactly equal to your preferences on options conditional upon the probability distribution P. I am not aware that this axiom has been used in the literature for reducing incompleteness in preferences to incompleteness in values. Similar axioms have been used, but the ones I know of all depend on the assumption of state-*in*dependent preferences.[4] It is undoubtedly very strong. It not only rules out fuzzy credences, but also basically gives us the required probabilities. But at least as a first step towards bringing together the two branches of expected utility theory, it will nonetheless be interesting to see what this assumptions implies.

To state the relevant uniqueness condition, say that i is *strictly non-uniform under* $\succeq$ if there are $\overline{x}_i, \tilde{x}_i, \underset{\sim}{x}_i, \underline{x}_i$ in X such that $\boldsymbol{a}_{(i,\overline{x}_i)} \succ \boldsymbol{a}_{(i,\tilde{x}_i)} \succ \boldsymbol{a}_{(i,\underset{\sim}{x}_i)} \succ \boldsymbol{a}_{(i,\underline{x}_i)}$. Given this definition, we can state

Theorem 5.2: Suppose a reflexive binary relation $\succeq$ on $\mathcal{Q}$ is N-conformable, and there is a reflexive binary relation $\dot{\succeq}$ on $\mathcal{K}$, such that $\succeq$ and $\dot{\succeq}$ jointly satisfy the Reduction Axiom for some positive probability distribution P on I. Then (i) there exists a nonempty closed convex set $U \subset U^*$ of state-dependent utility functions such that for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$,

$$\mathfrak{a} \dot{\succeq} \mathfrak{b} \text{ iff } \sum_{(i,x)\in Z} P(i)u(i,x)\mathfrak{a}(i,x) \geq \sum_{(i,x)\in Z} P(i)u(i,x)\mathfrak{b}(i,x) \quad \forall u \in U, \tag{5.9}$$

[4] Cf. e.g. the eponymic 'Reduction Axiom' in Ok et al. (2012).

and for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$,

$$\boldsymbol{a} \succeq \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} u(i,x)\boldsymbol{a}(i,x) \geq \sum_{(i,x)\in Z} u(i,x)\boldsymbol{b}(i,x) \quad \forall u \in U. \tag{5.10}$$

(ii) If $\{\boldsymbol{a}_n \succeq \boldsymbol{b}_n\}$ is a basis for $\succeq$ under these axioms, then U is the set of $u \in U^*$ satisfying $\{\mathbf{U}_u(\boldsymbol{a}_n) \geq \mathbf{U}_u(\boldsymbol{b}_n)\}$. And (iii) if each i is strictly non-uniform under $\succeq$, then there is no other probability distribution $Q \neq P$ for which (i) is true.

To apply this theorem to our context, we can expand my explication from chapter 4. Let U be a nonempty closed convex set of theory-dependent utility functions, and P a probability distribution over theories. Suppose you have the credence distribution P over theories that are jointly represented cardinally by U. I shall then say that the pair (U, P) *represents your axiological beliefs cardinally.* My assumption is that, if there is a unique pair $(U \in U^*, P)$, such that U represents your u-value judgments about $\mathcal{Q}$ ordinally, the pair (U, P) represents your u-value judgments about $\mathcal{K}$ ordinally, and U represents each of your axiologies cardinally, then (U, P) represents your axiological beliefs cardinally. Hence if that is so, we can assume, say, that $P(i)$ represents your credence p_i in the theory that is represented by the functions $\mathcal{G}_i = \{G_i : X \to \mathbb{R} \mid G_i = u(i, \cdot), u \in U\}$.

So given our explications, and the assumption that each of your $\succeq_i$ is N*-conformable and strictly non-uniform,[5] the following theorem holds:

Subjectivist Expected Value Theorem for Incompleteness: If your $\succeq_U$ is N-conformable and satisfies the Strong Pareto Condition with re-

[5]Say that $\succeq_i$ is strictly non-uniform if for some $\overline{x}_i, \tilde{x}_i, \underset{\sim}{x}_i, \underline{x}_i$ in X, $a_{\overline{x}_i} \succ_i a_{\tilde{x}_i} \succ_i a_{\underset{\sim}{x}_i} \succ_i a_{\underline{x}_i}$.

spect to your $\succeq_i$, and if your $\succeq_U$ and $\dot{\succeq}_U$ jointly satisfy the Reduction Axiom for some positive probability distribution P on I, then for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$,

$$\mathfrak{a} \dot{\succeq}_U \mathfrak{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \mathfrak{a}(i,x) p_i G_i(x) \geq \sum_{(i,x)\in Z} \mathfrak{b}(i,x) p_i G_i(x) \quad \forall G \in \mathcal{G}. \quad (5.11)$$

This is the most encompassing theorem of this thesis. Again, it does not allow us to represent fully incomparable theories, or theories that compare in a lexical way. Moreover, that your $\succeq_U$ and $\dot{\succeq}_U$ should satisfy the Reduction Axiom is only plausible if you do not rule out the truth of any axiology under consideration completely – or intuitively, if you assign nonzero credence to all axiologies. So to ground a normative argument on this theorem, we would have to restrict the set of axiologies under consideration accordingly. And even apart from that, the Reduction Axiom is ultimately unduly strong. But I take this to be an interesting first result in the relevant direction. It could again be extended to incorporate different accounts of intertheoretic comparisons, but I shall not pursue this here.

5.2 Further Explorations: General Moral Uncertainty

Now that we have a theory of axiological uncertainty that can allow for incompleteness, we are in a better position to explore a final major question that goes beyond the theory I have defended so far. More precisely, we are

now well-equipped to ask whether the account I have been outlining could serve as a general theory of *moral* uncertainty, rather than simply a theory of axiological uncertainty. So this is what I shall examine in the present section. For simplicity, I shall not introduce new labels. So in this section, when I speak of 'EVM', I shall not mean the theory of axiological uncertainty that I have been calling EVM so far. Instead, I shall mean the rough idea of expected value maximisation. My question is whether this idea, EVM, can be applied to moral uncertainty generally.

If EVM should function as a general theory of moral uncertainty, it cannot range merely over *betterness*-relations. Instead, we have to assume that there is some more general relation between options that all moral theories are concerned with. I shall take '(weakly) morally preferable' to be such a relation. So in this section, I shall understand '$\succeq$' as referring to the reflexive part of this relation, and the labels 'T_1', 'T_2', 'T_3',... accordingly as referring to theories about moral preferability. '$a \succeq_i b$' thus means that a is weakly morally preferable to b, according to the moral theory T_i; '$a \succ_i b$' and '$a \sim_i b$' are understood as usual, and denote that according to theory T_i, a is strictly morally preferable to b, or that a and b are equally morally preferable.[6]

There are various problems and questions with applying EVM to moral uncertainty generally.[7] In what follows, I shall focus on one issue only – viz.,

[6]In applying EVM to moral uncertainty, some authors have taken moral theories to imply 'moral choice-worthiness' relations (cf. MacAskill (2014)), or moral 'value' relations (where that is somehow understood more broadly than axiological value; cf. Sepielli (2010) and Ross (2006b)). However, both 'more choice-worthy than' and 'more valuable than' are, as a matter of meaning, transitive. So these interpretations rule out intransitive theories from the outset.

[7]One important problem is the possibility of supererogation, which most standard deontological theories allow. According to these theories, there are options a and b such that a is morally preferable to b, but it is not the case that you *ought* to choose a.

the question whether all, or at least most moral theories satisfy the relevant axioms of decision theory. If moral theories do not satisfy the axioms of decision theory, that raises two problems for EVM, which are both familiar by now. First, a theory may fail to satisfy these axioms because according to that theory, it is not the case that we ought to maximise expected value with regards to ordinary non-normative uncertainty. But then, if we have at least some credence in such theories, it is implausible that EVM can function as a general theory of moral *and* non-normative uncertainty. So such theories raise the question whether EVM is plausible. Secondly and more fundamentally, in order to explain what 'EVM' means and represent our theories with (something like) value-functions, we have to explicate what cardinal intratheoretic comparisons of value mean. Decision theory provides such an explication, but this explication presupposes that the relevant theory satisfies the axioms of decision theory. So if moral theories fail to satisfy the relevant axioms, this raises the question whether EVM can even be formulated as a general theory of moral uncerainty.

This second problem at least is quite generally acknowledged, even by people who endorse EVM as a general theory of moral uncertainty. However, my sense is that the limitations of the decision-theoretic explication are greatly underestimated. For example, in response to the worry that deontological theories do not assign cardinal values to outcomes, Jacob Ross says:

> Any theory that can serve as an adequate basis for action must tell us what to do in cases in which the outcomes of our actions are uncertain. [...] And it follows from Ramsey's representation theorem that for any theory that tells

This raises the problem how moral preferability and the moral 'ought' are weighed under uncertainty. Cf. e.g. Sepielli (2010, ch.6) for a discussion.

> us what to do in such cases of uncertainty and that satisfies certain minimal coherence conditions, we can construct a value function that indicates not just the ordinal values of one's options, but also ratios among the value intervals between them. (2006a, 25; cf. 2006b, 754f.)

By 'minimal coherence conditions', Ross must mean some axioms of decision theory. So he suggests that any minimally coherent theory that can serve as an 'adequate basis for action' will be representable by a value-function.[8]

But unfortunately, when it comes to moral theories, each of these axioms is extremely problematic. Moral theories can be perfectly coherent, and provide a perfectly 'adequate basis for action' even if they violate them. At least, this is what I shall argue in what follows. In my discussion, I shall focus on the von Neumann-Morgenstern axioms, as introduced in chapter 2; section 5.2.1 discusses Transitivity, section 5.2.2 addresses Continuity, and section 5.2.3 elaborates on Independence. However, my general arguments do not depend on this precise formulation of the axioms. As far as I see, they equally apply to all standard formulations of transitivity, independence and continuity constraints. The conclusion of my discussion will be that, strictly speaking, very many standard moral theories do not satisfy these axioms. I shall show that we can *extend*, or *revise* these theories such that they do accord with our constraints; and in some cases at least, that may be the best we can do for now in practice. But, as I shall argue, that does not present a convincing theoretical solution to the core problem. Strictly speaking,

[8]Sepielli (2010, ch.5) also suggests the same method for cardinalisation. However, he does not even mention that a theory has to satisfy certain conditions for this method to be applicable. MacAskill (2014) also suggests the same method, and does mention that moral theories have to satisfy certain axioms for that method to work, but he does not discuss whether they generally do so.

many standard moral theories do not satisfy the axioms, and understanding EVM as a general theory of moral uncertainty is more problematic than its proponents have assumed.

Since there are representation theorems that do not assume Completeness, the incompleteness of many moral theories does not present a problem. It may be worth mentioning, however, that dropping Completeness – even at the level of first-order theories – is absolutely vital if EVM should be anything like a general theory of moral uncertainty. Among non-consequentialist moral theories, incompleteness abounds. For example, on one plausible interpretation at least, the classic deontological view of Ross (1930) is highly incomplete, in cases where '*prima facie* duties' conflict. More generally, many pluralist views of ethics are incomplete. On such views, what is morally most preferable depends on a range of different considerations – such as special obligations, rights as side constraints, impersonal goodness, perfectionist values and commitments to personal projects, say. On many such views, there may not be precise facts about how these considerations weigh against each other, and thus they give rise to incompleteness.[9] Views that allow for moral dilemmas are also plausibly construed as incomplete.[10] And there are numerous other deontological views with similar incompleteness. So it is very important that we have a theory that does not presuppose Completeness.

[9]Such a pluralist theory, explicitly implying incompleteness, is defended in Nagel (1979).
[10]Cf. e.g. Richardson (1994, 115ff.).

5.2.1 Transitivity

How about Transitivity? Unfortunately, there are many moral theories that *prima facie* violate that axiom. (I shall explain the '*prima facie*'-caveat shortly.)

Consider, for example, the following person-affecting view of population ethics: if one has a choice among bringing about two worlds, it is preferable to bring about the world in which the total wellbeing of all the people that exist in both worlds is greater; if this total wellbeing is equal in both worlds, or there is no one that exists in both worlds, the worlds are equally preferable.[11] To see the intransitivity this view implies, suppose that we have some cardinal concept of wellbeing, and consider the following five worlds – where the first number in brackets refers to Brown's level of wellbeing in the respective world, the second number refers to White's level of wellbeing in that world, and 'Ω' indicates that the person does not exist in the world:

$a : (2, \Omega)$, $b : (1, 3)$, $c : (\Omega, 2)$, $d : (3, 1)$, and $e : (2, \Omega)$.

Let a to e represent the options of bringing about these worlds. Then according to the person-affecting view, $a \succ b$, $b \succ c$, $c \succ d$ and $d \succ e$. Since a and e seem to be the same options, this set of judgments is – at least *prima facie* – intransitive.[12]

Note that the intransitivity arises because, according to our view, whether the wellbeing of a person in a world matters depends on whether that person exists in the world we *compare* it with. More generally, according to this view, the moral worth of an option depends on its alternative. I shall say

[11] For a defence of such a view, cf. e.g. Roberts (2003).

[12] I thank John Broome for this example.

that this view features *alternative-dependency*. Alternative-dependency very easily leads to intransitivities of the above sort.

Unfortunately, as Tim Willenken (2012) has shown, common sense morality is full of such dependency. Consider, for example, the following three principles:

> **Numbers**: If faced with a pairwise choice between saving a lesser number of people from some harm and a greater number of people from that same harm, it is morally preferable to save the greater number.
>
> **Dominance**: If faced with a pairwise choice between two options a and b, where each individual is at least as well off if you choose a rather than b and someone is much better off, it is morally preferable to choose a.
>
> **No Pushing**: If faced with a pairwise choice between pushing one person off a bridge to his death in order to block a trolley and letting several other people get killed by that trolley, it is morally preferable to let the greater number get killed.

As Willenken has shown, these three principles generate a *prima facie* deontic cycle.[13] Or consider

> **Promises**: If faced with a pairwise choice between saving someone you have promised to save from death and saving one stranger from death and a second stranger from moderate injury, it is morally preferable to save the person that you have promised to save.

[13]Cf. Willenken (2012, 546) for an example.

No Killing for Promises: If faced with a pairwise choice between letting die someone you have promised to save from death and killing one stranger, it is morally preferable to let the person die.

Together with Dominance, these principles again yield a *prima facie* deontic cycle.[14] And such examples could be multiplied with ease. Since common sense morality is full of alternative-dependency, it is full of such *prima facie* violations of Transitivity.

However, the case is not as simple. All these examples raise the question how options, or outcomes, should be individuated. Suppose that a theory seems to imply that $a \succ b$, $b \succ c$, and $c \succ a$, on grounds of alternative-dependency. One might argue that, in that case, this theory treats a-when-b-was-the-alternative (a_b) as different from a-when-c-was-the-alternative (a_c); after all, our theory features alternative-dependency and thus takes the relevant alternatives really to matter. More concretely, consider again the person-affecting view of population ethics. It does indeed seem plausible that this theory treats $(2, \Omega)_{(3,1)}$ as distinct from $(2, \Omega)_{(1,3)}$. If we bring about $(2, \Omega)$ by rejecting $(3, 1)$, we made Brown worse off than she could have been. But there is no one whom we made worse off than she could have been if we bring about $(2, \Omega)$ by rejecting $(1, 3)$. Since this theory is particularly concerned with these kinds of harms, it may be natural for it to individuate these outcomes more fine-grainedly than I first did. And if we do individuate outcomes more fine-grainedly, our theories no longer violate Transitivity. If a theory says that $a \succ b$, $b \succ c$, and $c \succ a$, and we reindividuate outcomes via their alternatives, our theory implies that $a_b \succ b_a$, $b_c \succ c_b$, and $c_a \succ a_c$.

[14] Cf. Willenken (2012, 551) for an example.

This is perfectly consistent with Transitivity. So there is a general strategy by which we can render theories that seem to violate that axiom consistent with it.

However, there is a downside to this strategy. Many theories that are complete under a coarse-grained individuation of outcomes will become very incomplete under a more fine-grained one. Consider the person-affecting view of population ethics. If we individuate outcomes only by the people who exist in them and their level of wellbeing, the theory explicitly tells us, for any two outcomes, which of them is preferable (or that they are equally preferable). This is not the case if we individuate outcomes more finely. For example, the theory does not imply any ordering of the worlds $(2, \Omega)_{(3,1)}$ and $(1, 3)_{(2,\Omega)}$. Brown could have been better off in both $(2, \Omega)_{(3,1)}$ and $(1, 3)_{(2,\Omega)}$. But are these worlds equally preferable because the harm is the same (1 unit of wellbeing); or is $(2, \Omega)_{(3,1)}$ preferable because in this world Brown is better off than in $(1, 3)_{(2,\Omega)}$? As it stands, the theory is simply silent on this. It is not designed to order outcomes of this more complex kind.

And it is clear *why* the theory is silent. It is logically impossible that you may face a choice between $(2, \Omega)_{(3,1)}$ and $(1, 3)_{(2,\Omega)}$. In such a choice, you could either choose $(1, 3)$ by rejecting world $(2, \Omega)$, or choose $(2, \Omega)$ by rejecting the different world $(3, 1)$. But you can never face these two options at once. I shall thus call this an *impractical choice.* Individuation of outcomes in terms of their alternatives will always lead to impractical choices. But deontological theories are designed to guide your decision-making – and they can perfectly well guide your decision-making while being silent on all impractical choices. To use Ross's phrase, a theory can be a perfectly 'ade-

quate basis for action' even if it does not order options in choices that you can never face. We should not expect deontological views to give advice in impractical choices; there is no reason why they should.

In fact there is a difference here between axiologies and non-consequentialist moral theories, which will become important later on. They have a fundamentally different *structure*. Axiologies are claims about what is valuable in the world. But actual outcomes and prospects are, so to speak, indefinitely fine-grained: for each actual outcome or prospect, we can provide an infinitely long description of it, which distinguishes it from any other one. And for each aspect or property of outcomes, an axiology should imply whether that property is relevant in determining value. So in principle, an axiology is *fragmentary*, not fully specified or well-defined, if it is silent on how certain very fine-grained outcomes or prospects are ordered. This is not to say that a fully specified axiology should be 'complete' in my technical sense – that it should imply that any two outcomes or prospects are *comparable*. But – however fine-grainedly we individuate outcomes or prospects – it should imply for any two of them, either that they are equally good, or that one is better, or that they are incomparable. If it simply remains *silent* on how, or whether, two options compare, then to at least one of them it has not assigned a (precise or rough) value, which by its nature it should. The verdict of an absence (of precise value-comparisons) does not make it underspecified; but the absence of a verdict does.

Accordingly, axiologies should imply a verdict on impractical choices. Like any other, an impractical choice is simply a comparison among two outcomes or prospects, and an axiology should assign each of them a value. Whether we

can face a *choice* between them is completely irrelevant from an axiological point of view.

But again, non-consequentialist theories are designed to guide your decision-making – they tell you not to lie and steal, or not to behave cowardly, and so on. They are simply not – at least not ultimately – in the business of saying what is valuable about outcomes, and how valuable it is. So there is no reason why they should imply a verdict on impractical choices. By a non-consequentialist standard, there is nothing fragmentary or underspecified if a theory is silent on them.

As far as I see, this difference has so far been ignored in the recent debate about whether all moral theories can be 'consequentialised' – i.e. understood as identical or equivalent to a form of consequentialism. Some authors have endorsed the view that 'every moral view is consequentialist' (Dreier (1993, 24)), or that 'every plausible moral view is a mere notational variant of a consequentialist view' (Dreier (2011, 98)).[15] For the reason I have outlined, I think this is strictly speaking false. I take it that consequentialist theories necessarily satisfy Transitivity. If we are to guarantee that non-consequentialist theories do so as well, we have to individuate outcomes in a sufficiently fine-grained way. But then, a non-consequentialist theory, unlike a consequentialist one, may be fully specified and yet completely silent on certain choices.

What exactly does that mean for our present purposes? Fortunately, from the point of view of Completeness and Transitivity, I think this difference does not matter much. There are representation theorems that allow for incom-

[15]Cf. also Louise (2004, 519).

pleteness, and *if* the other conditions of these theorems could be satisfied even under a fine-grained individuation of outcomes, I think these theorems would serve their purpose. If for some two options a and b, a theory neither implies that $a \succeq b$ nor that $b \succeq a$, nothing in these formal theorems requires that this must be because of an explicit verdict of incomparability. So for the purposes of representation, we could treat the absence of a verdict and the verdict of absence in the same way. To be precise, we would have to bear this difference in mind. If the expectation with the relevant set of utility functions yields an incompleteness, in some cases (of non-consequentialist theories) that would indicate the absence of a verdict; in other cases (of consequantialist *or* perhaps non-consequentialist theories) it would indicate the verdict of absence. On a formal level, that difference would be lost. But we could bear it in mind, and no great harm would be done.

So given that we have theorems allowing for incompleteness, Transitivity and the incompleteness emerging from reindividuation would not *in themselves* present a problem. Let me thus turn to Continuity and Independence.

5.2.2 Continuity

As far as I see, questions of continuity are unfortunately rarely discussed in deontological ethics. But as with Transitivity, it seems that many standard deontological theories *prima facie* violate Continuity. This is because, *prima facie*, many deontological constraints are best captured in terms of probability-thresholds.

For example, consider the question whether or not a certain action violates

someone's rights and is therefore ruled out by a constraint. Presumably, you can respect a person's rights even if you take *some* risk of killing her for the sake of a minor pleasure of yours – driving past her on your way to a restaurant, say. However, you presumably violate a person's rights if you take a *considerable* risk of killing her for the sake of that minor pleasure. When does the constraint that we ought to respect people's rights apply? One answer that seems congenial to deontology is that there is a probability threshold. But unsurprisingly, such threshold-views lead to violations of Continuity. To see this, suppose you violate a person's rights if you take a risk of more than 1% of killing her for the sake of a minor pleasure; and suppose that there is thus a constraint against taking such a risk, but not against taking a risk of 1% or less. Furthermore, suppose that it is always morally preferable not to violate anyone's rights than to violate someone's rights, and that if you do not violate anyone's rights it is *ceteris paribus* preferable to risk killing fewer people rather than more. Now consider:

- a killing Brown with 100% probability;
- b killing White and Blue with 1% probability;
- c killing Brown with 1% probability.

Our view implies that $c \succ b \succ a$. However, there is no probability $p \in]0,1[$ such that $pa + (1-p)c \succ b$: for any $p > 0$, $pa + (1-p)c$ involves a risk of more than 1% of killing Brown, thus violates her rights, and b will be morally preferable to it.

Again, it is not difficult to find other examples where such thresholds seem congenial to deontological theorising. For example, consider the ques-

tion whether it is permissible to kill someone for the sake of saving others. Suppose that whether or not it is depends on whether or not that person forfeited her right not to be killed by you – say, by intending to kill the people you could save. Presumably, you do not need to be *absolutely* certain that that person forfeited her rights, otherwise the forfeiture-proviso would be irrelevant in practice. So suppose again that there is some relevant threshold, and you can permissibly kill someone for the sake of saving ten others only if the probability that she is innocent is no more than this threshold. Let that threshold be 5%. And suppose again that it is always preferable not to kill anyone impermissibly rather than to kill someone impermissibly, and that, *ceteris paribus*, it is preferable to permissibly kill fewer people rather than more in order to save others. Consider:

- d killing Blue, where Blue has a 10% probability of being innocent;
- e killing Brown and White, where Brown and White both have a 5% probability of being innocent;
- f killing Blue, where Blue has a 5% probability of being innocent.

Our view implies that $f \succ e \succ d$. However, there is again no probability $p \in]0,1[$ such that $pd + (1-p)f \succ e$: for any $p > 0$, $pd + (1-p)f$ involves killing someone who has a probability of more than 5% of being innocent, is thus impermissible, and e will be morally preferable to it.[16]

Or relatedly, consider the question about when we impermissibly use an innocent person as a means. On standard deontological views, it is impermissible to kill an innocent person as a means to save other people's lives. And

[16]This example is taken from Jackson and Smith (2006); the threshold view is accepted by Aboodi et al. (2008).

on many views, this constraint is 'patient-centred': it has to do with a kind of respect that we owe to that subject, and is independent of the number of lives saved. So it is impermissible to kill someone as as a means to save 50, or 100, or 300 others. Yet presumably, you can permissibly use an innocent person as a means to save someone else's life if doing so does not harm that person. For example, you may arguably push Red into a button if doing so is the only way to save Blue's life, and Red will not be harmed at all by your pushing her. On some views, this may be an infringement of Red's rights, and you might afterwards be under an obligation to apologise to Red for using her. But if it is an infringement of her rights, it is arguably a permissible one.[17] But what if it is uncertain whether the person will be killed by your using her as a means? Again, it seems natural for deontological views to accept a threshold for when we stop treating others as ends in themselves, and impermissibly use them as means. So suppose you impermissibly use an innocent person as a means to save others if you thereby take a risk of more than 2% of killing her; and suppose that there is thus a constraint against taking such a risk, but not against taking a risk of 2% or less. Furthermore, suppose that if you do not use anyone impermissibly as a means it is *ceteris paribus* preferable to save more people rather than fewer. Now consider:

- g using Red as a means to save Blue and Brown, thereby killing Red with 100% probability;
- h using White as a means to save Black, thereby killing White with 2% probability;

[17]For the view that one can sometimes permissibly infringe someone else's rights, cf. e.g. Thomson (1990, ch.6), Fabre (2012, ch.2).

k using Red as a means to save Blue and Brown, thereby killing Red with 2% probability.

Our view implies that $k \succ h \succ g$. However, there is no probability $p \in]0,1[$ such that $pg + (1-p)k \succ h$: for any $p > 0$, $pg + (1-p)k$ involves a risk of more than 2% of killing Red, and thus there will be a constraint against it. So for any $p > 0$, h will be morally preferable to $pg + (1-p)k$.

More examples could be given – for example, concerning the distinction between having and not having promised, between another person's being or not being judicious and autonomous, and many other deontological concerns. *Prima facie*, many standard deontological theories violate Continuity.

Is there a strategy to resist this appearance? At first, it might seem that reindividuating outcomes should again prove helpful. Consider again our example of risking Brown's death for the sake of some minor pleasure; take two outcomes, in both of which you killed Brown, and suppose that in the first one, your actions had a probability of less than 1%, while in the second one they had a probability of more than 1% of killing her. According to our theory, there was an important difference between your actions, in that only one of them constituted a violation of rights. So it seems that our theory treats these outcomes as distinct – implying, say, that Brown is suffering an accidental harm in the first, and a disrespectful rights-violation in the second outcome. More generally, for all views that involve a probability threshold, there seems to be a categorical distinction between an outcome that had more and one that had less than this threshold-probability of coming about. So again, it may be natural for such theories to individuate outcomes more

fine-grainedly than I did.

However, the case of Continuity is importantly different from that of Transitivity. Reindividuating outcomes does *not* help to make our theories satisfy Continuity. Take our example again. Option a consisted of killing Brown with a probability of 100%, option c involved killing her with a probability of 1%. If we individuate outcomes by the probability with which they came about, the outcome of a will be 'Brown is dead, killed as a result of a rights-violation', and one possible outcome of c will be 'Brown is dead, killed as a result of an unlucky accident' (or 'Brown is dead, and the probability of this happening was 1%', or something similar). This will again lead to impracticalities. Consider the option '$pa + (1 - p)c$', with $p \in]0, 1[$. One possible outcome of this option is that Brown was killed as a result of a rights-violation, and another possible outcome is that she was killed as a result of an unlucky accident. But, at least on any standard view of rights, you cannot face such an option in practice: whatever you did, your actions either did or did not *ex ante* violate Brown's rights. Whether or not they did should not differ from outcome to outcome. So if we individuate outcomes via the probability with which they came about, you cannot face an option like '$pa + (1 - p)c$'. In fact, this option not only contributes to an impractical *choice*, together with some other option. It is impractical in itself, all by its own. We might say it is an *impractical option*.

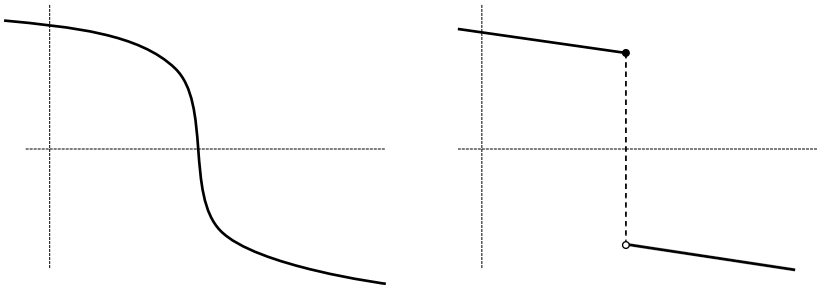
Again, moral theories will generally not give advice in choices involving such impractical options. There is no reason why they should. So they will generally not imply any judgments of the form '$pa + (1 - p)c \succ b$', or '$b \succ pa + (1 - p)c$'. But this is precisely what Continuity would require. So

reindividuating outcomes will *not* help to make theories satisfy that axiom.

Roughly, the relevant difference between Continuity and Transitivity is this. In either case, the pertinent reindividuation will lead to impracticalities, and thus to widespread incompleteness. This is a problem with Continuity, because Continuity itself features what we might call a *conditional minimal completeness constraint*. It requires that, *if* a theory implies that $a \succ b$ and $b \succ c$, it *must* also make some third judgment involving $pa + (1 - p)c$ (viz., that $pa + (1 - p)c \succ b$ for some $p \in]0, 1[$), and thus cannot be *fully* incomplete with respect to $pa + (1 - p)c$. Moreover, even under a fine-grained individuation of outcomes via their probabilities, there will be many (standard, practical) options for which moral theories imply the first two judgments, and thus satisfy the antecedent of the conditional completeness constraint. So then Continuity will require at least minimal completeness with respect to $pa + (1 - p)c$, which the theories will not satisfy with fine-grained outcomes.

Transitivity also features a conditional minimal completeness constraint. It requires that, *if* a theory implies that $a \succ b$ and $b \succ c$, it must also make some third judgment involving a and c (viz., that $a \succ c$), and thus be minimally complete in that respect. However, under a fine-grained individuation of outcomes via their *alternatives*, there will *not* be any options for which our theories imply the first two judgments, and thus satisfy the antecedent of the conditional minimal completeness constraint. So the relevant minimal completeness constraint never becomes effective. Even if theories do not imply judgments of the form $a_b \succeq c_b$, that is no problem, because Transitivity will not require them to do so.

Is there another strategy to resist the *prima facie* violations of Continuity? One might argue that we should simply understand standard deontological theories differently than I have suggested above. Violations of Continuity arise only if there are sharp thresholds – intuitively, if there are sudden leaps in the graph that designates the moral worth of options. But perhaps we can capture the basic deontological ideas in ways that do not involve such leaps. Consider again our view of rights. Instead of saying that a risk of 1% marks a sudden cut-off point at which people's rights are violated, we could assume that, at a risk of 1%, the moral worth of our action decreases very drastically, but continuously. More generally, and graphically illustrated, we could assume that deontological functions of moral worth have drastic continuous drops instead of leaps – as in the following illustration:



Continuous Function Discontinuous Function

Figure 5.1.

One might argue that this still leads to recognisably deontological views. Indeed, one might even argue that this is a more charitable interpretation of these views, since discontinuities of any sort are implausible (one might say).

However, I am sceptical about this move. In many cases, discontinuities and thresholds do seem to be congenial to deontological views, as deontolog-

ical views operate with all-or-nothing concepts. For example, on standard deontological views, there is an important *qualitative* difference in whether or not you violate my rights. Your actions are perfectly permissible as long as you *almost*, but not quite, violate my rights; and they are impermissible if you do. There is no way in which you could violate my rights *a little bit*. Of course, there can be more and less important rights, and so not all violations of rights have to be equally grievous. But even the less important rights will either be violated or not, and that will be a qualitative difference – a matter of all or nothing. All-or-nothing concepts are at the heart of deontology, and in the face of uncertainty many of them will have to be captured by probability thresholds. If we reinterpret such views as continuous, we may produce extensionally similar cousins of them. And working with these cousins may be the best we can do for now, in practice. But that does not present a convincing theoretical solution to our problem.

Moreover, as we shall see in the next section, even if we render our deontological views continuous, this will not help to make them satisfy Independence. So even if we accept this strategy in the face of Continuity, that will not solve the general problem that we are concerned with.

There is a final strategy that one could propose; but I do not think that that strategy will be more promising. It will also be a possible candidate concerning Independence; so let me discuss it in that context, and turn to this final axiom now.

5.2.3 Independence

If what I've said so far was correct, the case of Independence is even more problematic. As with the other axioms, many standard moral theories *prima facie* violate this condition.

Consider, for example, a view according to which a distribution of goods is fair if and only if, in it, people's claims are satisfied in proportion to the strength of these claims. And suppose that, on this view, if a good cannot be divided, the fairest thing to do is to put on a lottery in which people's chances to receive the good are in proportion to the strength of their claims.[18] Now suppose two people, Brown and White, both have a claim to a certain indivisible good G, but Brown's is twice as strong as White's. Suppose you are faced with the options

- a Brown gets G, and
- b White gets G.

On the present theory, $a \succ b$, since this distribution comes closer to a distribution in which everyone's claims are satisfied in proportion to their strength. However, suppose you have a biased coin that lands tails 2/3 of the time, and you can choose among the following lotteries

- c Brown gets G both if the coin lands heads and if it lands tails, and
- d White gets G if the coin lands heads, and Brown if it lands tails.

We now have $d \succ c$; d is actually the fairest option, given that the good is indivisible. However, since c seems to be the same option as $1/3a + 2/3a$,

[18] For a defence of such a view, cf. e.g. Broome (1990).

and d the same as $1/3b + 2/3a$, this theory seems to violate Independence.[19]

It is clear why it does. According to this theory, what matters about an outcome is not only what *actually* happens in that outcome. It is also what *could* have happened, and with what probability, if that outcome did not come about. So according to this view, we cannot evaluate the possible outcomes of our actions independently of one another. This is why it violates the Independence condition.

For the same reason the view from the previous section, on which you violate someone's rights if you take too great a risk of killing her for the sake of a pleasure of yours, will also fail to satisfy Independence. On this view too, what matters is not only what *actually* happened in a particular outcome, but also what could have happened, and with what probability, if that outcome did not come about. Suppose again that you violate someone's rights if you take a risk of more than 1% of killing her for the sake of a minor pleasure, but not if the risk is 1% or less; and consider our options

e killing Brown with 100% probability;

f killing White and Blue with 1% probability;

g killing Brown with 1% probability.

The theory implies that $g \succ f$. But for $p = 0.99$, it implies that $pf+(1-p)e \succ pg + (1-p)e$ – since the latter but not the former option violates someone's rights.

The same holds for the view I considered about impermissibly using someone as a means. As is easy to see, that view will also violate Independence.

[19]Cf. Diamond (1967) for a similar example.

So *prima facie*, violations of Independence are again extremely common in standard deontology. Unfortunately, however, neither of the two strategies we encountered with Transitivity and Continuity works with Independence.

First, violations of Independence do not presuppose any precise probability-thresholds, or sharp cut-off points. Even if we interpret our theories as implying continuous drops instead of discontinuous leaps, the fact remains that what matters, according to them, is not only what *actually* happened in a particular outcome, but also what could have happened if that outcome did not come about. This very general fact causes violations of Independence, and we do not avoid it by rendering our theories continuous. For example, consider our view about rights-violations again. Suppose that the moral worth of options decreases drastically but continuously when we take a risk of more than 1% of killing someone. Then *adding* an additional risk of killing Brown will decrease the worth of the option 'killing White and Blue with 1% probability' less than it will decrease the worth of the option 'killing Brown with 1% probability'. And this phenomenon will lead to violations of Independence. So *even if* we assume that this strategy allows us to avoid violations of Continuity, it does not help with Independence. Whatever we think about interpreting deontological views as continuous, doing so does not help to solve the general problem that we are concerned with.

Secondly, the problem with reindividuating outcomes via the probability with which they came about is exactly the same as with Continuity. Independence also features a conditional minimal completeness constraint. It requires that, if a theory implies that $a \succ b$, it *must* also make judgments involving some compound option $pa + (1-p)c$ (viz., that $pa + (1-p)c \succ pb + (1-p)c$

for all c). Moreover, even under a fine-grained individuation of outcomes via their probabilities, there will be many standard, practical options for which moral theories imply the first judgment, and thus satisfy the antecedent of the conditional completeness constraint. So then Independence will require minimal completeness with respect to $pa + (1 - p)c$. With fine-grained outcomes, however, theories will not even be minimally complete with respect to $pa + (1 - p)c$, because this will be an impractical option, and theories will thus remain silent on it.

As I mentioned briefly on page 260, there is a final move in response to violations of Independence, which could also be made with respect to Continuity. This move builds on the strategy of reindivudating outcomes. As I have shown, in these two cases, reindividuating outcomes will render our theories too incomplete even to satisfy the respective axioms (Continuity and Independence). This is because reindividuating outcomes will produce impractical options, and our theories are not designed to order such options. So what we could do is this. We could *extend* our theories and *make* them imply verdicts on impractical options; and we could do that in ways that accord with the relevant constraints in Continuity and Independence. As far as I see, that would be possible. Moreover, as far as first-order judgments of moral preferability are concerned, it would not alter these views in any way that is *practically* relevant. After all, we would keep all their verdicts on practically possible choices, and only extend them with respect to choices that you cannot possibly face in practice. So one might argue that this kind of extension would be innocuous, and not distort our views in any problematic respect.

However, that would not be a good move, I think. For one thing, by so extending our views, we would turn them into different theories – indeed, theories of a different kind. As I said on page 250, non-consequentialist moral theories have a specific nature, and that is fundamentally different from the nature of axiologies. Only from an axiological point of view is it irrelevant whether options can figure in practical choice. By a non-consequentialist standard, there is nothing fragmentary or underspecified if a theory is silent on impractical pairs of options. In fact, since non-consequentialist theories are not in the business of saying what is valuable about outcomes and how valuable it is, there will often be no *basis* within these views on which they could imply verdicts on impractical choices or options. Hence by making our moral theories order impractical options, we would not somehow finish them, complement them where otherwise they would be fragmentary. We would change their very nature. We would basically turn them into axiologies.

Moreover, even though our extensions would have no practical implications as far as first-order judgments of moral preferability are concerned, they *would* have practical implications under uncertainty. Consider again the options of killing Brown with 100% probability (e), killing White and Blue with 1% probability (f), and killing Brown with 1% probability (g). And suppose we reindividuate outcomes and supplement our theories with verdicts such that, for some p and q in $]0,1[$, $pe+(1-p)g \succ f$ and $f \succ qe+(1-q)g$. The precise verdicts will determine where on the value scale between e and g the value of f lies. And that in turn will determine how important it is to choose f rather than e, according to our theory. And while this may be irrelevant if we are certain of this theory and face a choice between e and f, it is not irrel-

evant if we have to weigh it against a theory according to which e is morally preferable to f. So there would be no basis within the deontological views to make these extensions, and yet they would have significant implications under uncertainty. Hence even if these extensions would for now be the best we could work with in practice, they would not, I think, present a convincing theoretical solution to our problem. If we present a theory of uncertainty about these extended cousins of our deontological views, we would not be presenting a theory of uncertainty about *these* deontological views.

Finally, this last strategy threatens to render our overall theory useless in practice. As I have argued in section 4.2.4, our theory is useful in practice because it allows us to make complicated decisions depend on simpler ones. If we individuate outcomes such that each outcome can, in practice, figure in only one option and choice, that will become very difficult. To check whether our judgments satisfy the axioms, we would have to consider these complex, entirely impractical options. And it is dubious whether we have any firm intuitions or judgments about such options. So by individuating outcomes very fine-grainedly and extending our theories so that they satisfy our axioms, we make it more and more difficult to use EVM in practice.

Conclusion

In conclusion, I think a great number of non-consequentialist views do not satisfy the standard axioms of decision theory, and there is no ultimately successful way in which we can interpret them as doing so. So I am sceptical about whether it is possible to understand the theory I have been developing

as a general theory of moral uncertainty.

I have conceded that it is not even a general theory of axiological uncertainty. I have conceded that there are axiologies that violate the necessary axioms, and that my theory does not apply to them. So I have presented it as a theory of uncertainty about those axiologies that satisfy the relevant conditions – ultimately, the theories that are N*-conformable. One might perhaps do the same with respect to moral theories. Perhaps one might understand my theory as a theory of uncertainty about N*-conformable *moral* theories. However, my sense is that most standard *axiologies* do satisfy our conditions, and that the axiologies I ultimately exclude are comparatively rare and atypical. So the remaining theory can still count as a '(restricted) theory of axiological uncertainty' in some interesting sense. In contrast, as I have shown, it seems that standard deontological views will violate our axioms, and if there are deontological views that satisfy them, then *these* will be comparatively rare and atypical. So the remaining account would not be a 'theory of moral uncertainty' in any interesting sense. Moreover, I have now addressed only one problem with understanding EVM as a general theory of moral uncertainty. There remain other problems – for example, the question of how to accommodate supererogation.[20] Even disregarding the question I have discussed, these other problems would still need to be addressed.

So I think that understanding EVM as a general theory of moral uncertainty is more problematic than some authors have claimed. I have thus restricted the topic of this thesis to axiological uncertainty.

[20]Cf. footnote 7 in this chapter.

5.3 Conclusion

In this thesis, I have tried to explore an axiomatic approach to the problem of axiological uncertainty, and to the idea of Expected Value Maximisation in particular. In chapter 2, I outlined one of the most basic results from state-dependent utility theory, and argued that – at least within certain restrictions, and *modulo* the problem of intertheortic comparisons – that theorem can be applied to axiological uncertainty to formulate and defend EVM. In chapter 3, I then argued that intertheoretic comparisons are possible, and outlined how various accounts of intertheoretic comparisons could be axiomatised by the theorem from chapter 2. In chapter 4, I furnished the overall theory with an account of axiological credences, and thus paved the way for applying it in real life cases. In chapter 5, I extended the theorem so as to make it applicable to incomplete axiologies and to axiologies that are less than fully comparable.

Naturally, there remain many open problems for further research. On the one hand, there are further philosophical questions. For instance, in discussing the axioms, I have restricted myself to comparing their plausibility in our context with their status in other contexts. To make a more complete case for EVM, it would be necessary to say more in their defence. Similarly, the problem of intertheoretic comparions is still very underexplored, and my discussion of it was somewhat speculative in many respects. That problem certainly merits much further attention. And most urgently perhaps, this last chapter raised problems for the theory of *moral* uncertainty, and it would be important to make positive progress on these problems.

On the other hand, there remain many open technical problems. It may be worthwhile to seek theorems for incomplete orderings that yield unique separations of subjective probabilities and state-dependent utilities and that do not rely on the strong Reduction Axiom. It ultimately seems important to explore results for state-dependent utilities that allow for incomparability in values as well as fuzzy credences – so-called multi-prior multi-utility axiomatisations. And since there are infinitely many axiologies, it would certainly be desirable to extend the results of this thesis to a framework that allows for an infinite state-space.

As so often in moral philosophy, I cannot be certain that all my arguments were sound. But I hope that this thesis has at least shed a new light on some problems and difficulties in the theory of normative uncertainty. In particular, I hope I have shown that EVM is a plausible theory, and that the axiomatic approach to axiological uncertainty is worth pursuing. And I hope that all of this was valuable.

Chapter 6

Appendices

6.1 Appendices to Chapter 2

The Expected Value Theorem

To prove the Expected Value Theorem, suppose $\succeq_U$ is vNM-conformable. Karni and Schmeidler's Theorem then immediately implies that there is a function $u : Z \to \mathbb{R}$, unique up to positive affine transformation, such that for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}$,

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \boldsymbol{a}(i,x)u(i,x) \geq \sum_{(i,x)\in Z} \boldsymbol{b}(i,x)u(i,x). \tag{6.1}$$

Now for some theory T_i, consider the set of options in which the probability of T_i is 1, $\mathcal{Q}^i_{p=1} = \{\boldsymbol{a} \in \mathcal{Q} \mid \sum_{x\in X} \boldsymbol{a}(i,x) = 1\}$. According to (6.1), $u(i,\cdot)$ constitutes a utility function on X, which represents the relation $\succeq_U$

restricted to $\mathcal{Q}^i_{p=1}$. That is, for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^i_{p=1}$,

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{x \in X} \boldsymbol{a}(i,x)u(i,x) \geq \sum_{x \in X} \boldsymbol{b}(i,x)u(i,x). \tag{6.2}$$

Since $\succeq_i$ is complete, the Pareto Condition immediately implies that for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^i_{p=1}$, $\boldsymbol{a} \succeq_U \boldsymbol{b}$ if and only if $H_i(\boldsymbol{a}) \succeq_i H_i(\boldsymbol{b})$. Clearly, this is consistent with the assumption that $\succeq_i$ is vNM-conformable. Moreover, $u(i,\cdot)$ thus constitutes a utility function on X that represents T_i ordinally, in the sense of (2.4). Given that goodness is expectational, it also represents T_i cardinally. The same argument applies to all i in I. So u represents the u-value relation ordinally, and represents each axiology cardinally. So given the decision-theoretic explication of intertheoretic comparisons, it also jointly represents the axiologies cardinally. We can thus interpret all functions $u(i,\cdot)$ in (6.1) as value-functions G_i. ■

6.2 Appendices to Chapter 4

The Subjectivist Expected Value Theorem

By basically the same reasoning as in the derivation of the Expected Value Theorem, the Subjectivist Expected Value Theorem follows immediately from Karni and Schmeidler's Theorem 2. ■

The Subjectivist Weighted Value Theorem

Karni and Schmeidler (1980, 7) state the following theorem:

Karni and Schmeidler's Theorem 3: If a reflexive binary relation $\dot{\succeq}$ on $\mathcal{K}$ is vNM-conformable, there is a utility function $u : Z \to \mathbb{R}$, unique up to positive unit-comparable transformation, such that for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$,

$$\mathfrak{a} \dot{\succeq} \mathfrak{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \mathfrak{a}(i,x)u(i,x) \geq \sum_{(i,x)\in Z} \mathfrak{b}(i,x)u(i,x). \tag{6.3}$$

This theorem immediately implies that if the u-value relation $\dot{\succeq}_U$ is vNM-conformable, there is a utility function u such that

$$\mathfrak{a} \dot{\succeq}_U \mathfrak{b} \quad \text{iff} \quad \sum_{(i,x)\in Z} \mathfrak{a}(i,x)u(i,x) \geq \sum_{(i,x)\in Z} \mathfrak{b}(i,x)u(i,x). \tag{6.4}$$

Now for some theory T_i and some outcome y, consider the set $\mathcal{K}^i_y = \{\mathfrak{a} \in \mathcal{K} \,|\, \mathfrak{a}(j,y) = 1 \quad \forall j \neq i\}$. $\mathcal{K}^i_y \subset \mathcal{K}$ is the set of options that if T_i is false certainly lead to y. We then have, for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}^i_y$,

$$\mathfrak{a} \dot{\succeq}_U \mathfrak{b} \quad \text{iff} \quad \sum_{x\in X} \mathfrak{a}(i,x)u(i,x) \geq \sum_{x\in X} \mathfrak{b}(i,x)u(i,x). \tag{6.5}$$

Since all $\succeq_i$ are reflexive and complete, the Pareto Condition immediately implies that for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}^i_y$, $\mathfrak{a} \dot{\succeq}_U \mathfrak{b}$ if and only if $K_i(\mathfrak{a}) \succeq_i K_i(\mathfrak{b})$. So $u(i,\cdot)$ actually constitutes a utility function on X that represents T_i ordinally. Given that goodness is expectational, it also represents T_i cardinally. The same argument applies to all i in I. The uniqueness condition follows immediately from Karni and Schmeidler's Theorem 3. ■

6.3 Appendices to Chapter 5

Theorem 5.1

Let I again be our finite (index) set of theories, and X our a finite set of outcomes, with $Z = \{(i,x) \,|\, i \in I, x \in X\}$ and $\mathcal{Q} = \{\boldsymbol{a} : Z \to \mathbb{R}_+ \mid \sum_{(i,x)\in Z} \boldsymbol{a}(i,x) = 1\}$. Note that Nau's Theorem holds for any finite sets of states and outcomes. So let S' be the singleton $\{k\}$, let X' be a set of $|X| \cdot |I|$ outcomes, and define $Z' = \{(k,x) \,|\, x \in X'\}$, and $\mathcal{K}' = \{\mathfrak{a} : Z' \to \mathbb{R}_+ \mid \sum_{x\in X'} \mathfrak{a}(k,x) = 1\}$. According to Nau's Theorem, if a reflexive relation $\dot{\succeq}$ on $\mathcal{K}'$ is N-conformable, there is a nonempty closed convex set $W \subset W^*$ of functions representing it in the sense of (5.3). Since there is a simple bijection between $\mathcal{Q}$ and $\mathcal{K}'$ (and to each constant act $\boldsymbol{a}_{(i,x)}$ in $\mathcal{Q}$ there corresponds a constant act $\mathfrak{a}_x$ in $\mathcal{K}'$), this implies Theorem 5.1.

I shall spell this out in some detail. To see the bijection between $\mathcal{Q}$ and $\mathcal{K}'$, label the outcomes in X by $X = \{x_1, x_2, ..., x_k\}$, and the outcomes in X' by:

$$\begin{aligned} X' = \{&x_{11}, x_{12}, ..., x_{1k}, \\ &x_{21}, x_{22}, ..., x_{2k}, \\ &..., \\ &x_{n1}, x_{n2}, ..., x_{nk}\} \end{aligned}$$

We have, for all $\boldsymbol{a} \in \mathcal{Q}$ and all $\mathfrak{a} \in \mathcal{K}'$,

$$\sum_{(i,x)\in Z} \boldsymbol{a}(i,x) = \sum_{(k,x)\in Z'} \mathfrak{a}(k,x) = \sum_{x\in X'} \mathfrak{a}(k,x) = 1. \tag{6.6}$$

So with each $\boldsymbol{a} \in \mathcal{Q}$ we can associate an act $\mathfrak{a} \in \mathcal{K}'$ which is such that $\mathfrak{a}(k, x_{ij}) = \boldsymbol{a}(i, x_j)$, and vice versa. More formally, we can define a bijection $M : \mathcal{K}' \to \mathcal{Q}$,

$$\begin{aligned} &M : \mathfrak{a} \mapsto M(\mathfrak{a}),\ M(\mathfrak{a})(i, x_j) = \mathfrak{a}(k, x_{ij}),\ \text{with} \\ &M^{-1} : \boldsymbol{a} \mapsto M^{-1}(\boldsymbol{a}),\ M^{-1}(\boldsymbol{a})(k, x_{ij}) = \boldsymbol{a}(i, x_j). \end{aligned} \tag{6.7}$$

For some best and worst outcomes $\overline{x}$ and $\underline{x}$ in X', we can define

$$\begin{aligned} W^* = \{w : Z' \to \mathbb{R} \mid w(k, \underline{x}) = 0;\ 0 \leq w(k, x) \leq 1\ \ \forall\, x \in X \setminus \{\overline{x}, \underline{x}\}; \\ w(k, \overline{x}) = 1\}. \end{aligned} \tag{6.8}$$

For some reflexive binary relation $\succeq$ on $\mathcal{Q}$, define a binary relation $\dot{\succeq}$ on $\mathcal{K}'$ such that $\mathfrak{a} \,\dot{\succeq}\, \mathfrak{b}$ if and only if $M(\mathfrak{a}) \succeq M(\mathfrak{b})$. Then if $\succeq$ is N-conformable (for $(\overline{i, x}) = (m, x_n)$ and $(\underline{i, x}) = (p, x_q)$), $\dot{\succeq}$ is N-conformable (for $\overline{x} = x_{mn}$ and $\underline{x} = x_{pq}$). From Nau's Theorem, we know that if the latter is true, there is a nonempty closed convex set $W \subset W^*$ of state-dependent utility functions, such that for all $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}'$,

$$\mathfrak{a} \,\dot{\succeq}\, \mathfrak{b} \quad \text{iff} \quad \sum_{x \in X'} \mathfrak{a}(k, x) w(k, x) \geq \sum_{x \in X'} \mathfrak{b}(k, x) w(k, x) \quad \forall w \in W. \tag{6.9}$$

For any $w \in W$, define a corresponding function $u : Z \to \mathbb{R}$, such that $u(i, x_j) = w(k, x_{ij})$, and let U be the set of all such u. Since W is nonempty closed and convex, and $W \subset W^*$, U is nonempty closed and convex, and $U \subset U^*$. Now take any $\boldsymbol{a}$ and $\boldsymbol{b}$, and focus on the corresponding $\mathfrak{a} = M^{-1}(\boldsymbol{a})$

and $\mathfrak{b} = M^{-1}(\boldsymbol{b})$. Given (6.9), we have

$$\begin{aligned}
\boldsymbol{a} \succeq \boldsymbol{b} \;&\Leftrightarrow\; \mathfrak{a} \;\dot{\succeq}\; \mathfrak{b} \\
&\Leftrightarrow\; \sum_{x \in X'} \mathfrak{a}(k,x) w(k,x) \geq \sum_{x \in X'} \mathfrak{b}(k,x) w(k,x) \quad \forall w \in W \\
&\Leftrightarrow\; \sum_{(i,x) \in Z} M(\mathfrak{a})(i,x) u(i,x) \geq \sum_{(i,x) \in Z} M(\mathfrak{b})(i,x) u(i,x) \quad \forall u \in U \\
&\Leftrightarrow\; \sum_{(i,x) \in Z} \boldsymbol{a}(i,x) u(i,x) \geq \sum_{(i,x) \in Z} \boldsymbol{b}(i,x) u(i,x) \quad \forall u \in U. \qquad (6.10)
\end{aligned}$$

So if there is a nonempty closed convex set $W \subset W^*$ of state-dependent utility functions representing $\dot{\succeq}$ in the sense of (6.9), then there is a nonempty closed convex set $U \subset U^*$ of state-dependent utility functions representing $\succeq$ in the sense of (6.10). So this must be true if $\succeq$ is N-conformable.

Similarly, if $\{\boldsymbol{a}_n \succeq \boldsymbol{b}_n\}$ is a basis for $\succeq$ under our axioms, then $\{M^{-1}(\boldsymbol{a}_n) \;\dot{\succeq}\; M^{-1}(\boldsymbol{b}_n)\}$ is a basis for $\dot{\succeq}$; so by Nau's Theorem, W is the set of $w \in W^*$ satisfying $\{\mathbf{U}_w(M^{-1}(\boldsymbol{a}_n)) \geq \mathbf{U}_w(M^{-1}(\boldsymbol{b}_n))\}$. So U must be the set of functions satisfying $\{\mathbf{U}_u(\boldsymbol{a}_n) \geq \mathbf{U}_u(\boldsymbol{b}_n)\}$. ■

The Expected Value Theorem for Incompleteness

Theorem 5.1 implies that if $\succeq_U$ is N-conformable, there exists a nonempty closed convex set $U \subset U^*$ of theory-dependent utility functions, such that for all $\boldsymbol{a}$ and $\boldsymbol{b} \in \mathcal{Q}$,

$$\boldsymbol{a} \succeq_U \boldsymbol{b} \quad \text{iff} \quad \sum_{(i,x) \in Z} u(i,x) \boldsymbol{a}(i,x) \geq \sum_{(i,x) \in Z} u(i,x) \boldsymbol{b}(i,x) \quad \forall u \in U. \qquad (6.11)$$

Now for some theory T_i, consider the set $\mathcal{Q}^i_{p=1} = \{\boldsymbol{a} \in \mathcal{Q} \mid \sum_{x \in X} \boldsymbol{a}(i,x) = 1\}$.

The Strong Pareto Condition immediately implies that for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathcal{Q}^i_{p=1}$, $\boldsymbol{a} \succeq_U \boldsymbol{b}$ if and only if $H_i(\boldsymbol{a}) \succeq_i H_i(\boldsymbol{b})$. So $U_i = \{u(i,\cdot) \mid u \in U\}$ actually constitutes a closed convex set of utility function on X that represents T_i ordinally. Given my explication of intratheoretic comparisons, it also represents T_i cardinally. The same argument applies to all i in I. So given my explication of intertheoretic comparisons, U jointly represents all axiologies cardinally. ■

Theorem 5.2

(i). From Theorem 5.1, we know that there must be a nonempty closed convex set $U \subset U^*$ of theory-dependent utility functions that represents $\succeq$ in the sense of (5.10). To see that $\dot{\succeq}$ can be represented as an expectation of P and the same set of functions U, take some $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$, and focus on $\boldsymbol{a}_P$ and $\boldsymbol{b}_P$ in $\mathcal{Q}$, with $\boldsymbol{a}_P(i,x) = P(i)\mathfrak{a}(i,x)$, and $\boldsymbol{b}_P(i,x) = P(i)\mathfrak{b}(i,x)$ for all $(i,x) \in Z$. We know that

$$\boldsymbol{a}_P \succeq \boldsymbol{b}_P \quad \text{iff} \quad \sum_{(i,x)\in Z} u(i,x)\boldsymbol{a}_P(i,x) \geq \sum_{(i,x)\in Z} u(i,x)\boldsymbol{b}_P(i,x) \quad \forall u \in U. \tag{6.12}$$

Since $\boldsymbol{a}_P$ and $\boldsymbol{b}_P \in \mathcal{Q}^P$, and $\mathfrak{a} = L(\boldsymbol{a}_P), \mathfrak{b} = L(\boldsymbol{b}_P)$, the Reduction Axiom

implies that

$$\begin{aligned}\mathfrak{a} \dot{\succeq} \mathfrak{b} &\Leftrightarrow \boldsymbol{a}_P \succeq \boldsymbol{b}_P \\ &\Leftrightarrow \sum_{(i,x)\in Z} u(i,x)\boldsymbol{a}_P(i,x) \geq \sum_{(i,x)\in Z} u(i,x)\boldsymbol{b}_P(i,x) \quad \forall u \in U \\ &\Leftrightarrow \sum_{(i,x)\in Z} P(i)u(i,x)\mathfrak{a}(i,x) \geq \sum_{(i,x)\in Z} P(i)u(i,x)\mathfrak{b}(i,x) \quad \forall u \in U.\end{aligned} \tag{6.13}$$

So this relation must hold for any $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathcal{K}$.

(ii). This claim follows immediately from the relevant claim in Theorem 5.1.

(iii). By way of negation, suppose we had a different probability distribution $Q \neq P$ for which (i) is true. Then there must be some h and k in I with $P(h) > Q(h)$ and $P(k) < Q(k)$. We define lotteries in $\mathcal{Q}^P$ and $\mathcal{K}$ that contradict this assumption. So suppose that each i is strictly non-uniform under $\succeq$, i.e. that for each i, there are $\overline{x}_i, \tilde{x}_i, \underline{\tilde{x}}_i, \underline{x}_i$ in X such that $\boldsymbol{a}_{(i,\overline{x}_i)} \succ \boldsymbol{a}_{(i,\tilde{x}_i)} \succ \boldsymbol{a}_{(i,\underline{\tilde{x}}_i)} \succ \boldsymbol{a}_{(i,\underline{x}_i)}$. Now for any $r \in [0,1]$, define $\boldsymbol{a}_r$ and $\boldsymbol{b}_r \in \mathcal{Q}^P$ by

$$\begin{aligned}
&\boldsymbol{a}_r(h,\overline{x}_h) = r \cdot P(h), && \boldsymbol{b}_r(k,\underline{x}_k) = r \cdot P(k), \\
&\boldsymbol{a}_r(h,\underline{x}_h) = (1-r) \cdot P(h), && \boldsymbol{b}_r(k,\overline{x}_k) = (1-r) \cdot P(k), \\
&\boldsymbol{a}_r(k,\underline{x}_k) = P(k), && \boldsymbol{b}_r(h,\underline{x}_h) = P(h),
\end{aligned}$$

and assume that $\boldsymbol{a}_r$ and $\boldsymbol{b}_r$ agree outside h and k (i.e. that $\boldsymbol{a}_r(i,x) = \boldsymbol{b}_r(i,x)$ for all $x \in X, i \in I \setminus \{h,k\}$). Similarly, define $\mathfrak{a}_r$ and $\mathfrak{b}_r \in \mathcal{K}$ by

$$\mathfrak{a}_r(h,\overline{x}_h) = r, \qquad \mathfrak{b}_r(k,\underline{x}_k) = r,$$

$$\mathfrak{a}_r(h, \underline{x}_h) = (1-r), \quad \mathfrak{b}_r(k, \overline{x}_k) = (1-r),$$
$$\mathfrak{a}_r(k, \underline{x}_k) = 1, \qquad\qquad \mathfrak{b}_r(h, \underline{x}_h) = 1,$$

and assume that $\mathfrak{a}_r$ and $\mathfrak{b}_r$ agree outside h and k.

Since $\mathfrak{a}_r = L(\boldsymbol{a}_r)$, and $\mathfrak{b}_r = L(\boldsymbol{b}_r)$, the Reduction Axiom implies that $\boldsymbol{a}_r \succeq \boldsymbol{b}_r$ if and only if $\mathfrak{a}_r \dot{\succeq} \mathfrak{b}_r$ for any $r \in [0,1]$. So if (i) is true for Q, then for any $r \in [0,1]$,

$$\sum_{(i,x)\in Z} u(i,x)\boldsymbol{a}_r(i,x) \geq \sum_{(i,x)\in Z} u(i,x)\boldsymbol{b}_r(i,x) \quad \forall u \in U \Leftrightarrow$$
$$\sum_{(i,x)\in Z} Q(i)u(i,x)\mathfrak{a}_r(i,x) \geq \sum_{(i,x)\in Z} Q(i)u(i,x)\mathfrak{b}_r(i,x) \quad \forall u \in U. \tag{6.14}$$

The left-hand side of the biconditional (6.14) is equivalent to

$$\begin{aligned} &rP(h)u(h,\overline{x}_h) + (1-r)P(h)u(h,\underline{x}_h) + P(k)u(k,\underline{x}_k) \geq \\ &\quad rP(k)u(k,\underline{x}_k) + (1-r)P(k)u(k,\overline{x}_k) + P(h)u(h,\underline{x}_h) \quad \forall u \in U \Leftrightarrow \\ &rP(h)[u(h,\overline{x}_h) - u(h,\underline{x}_h)] \geq (1-r)P(k)[u(k,\overline{x}_k) - u(k,\underline{x}_k)] \quad \forall u \in U. \end{aligned} \tag{6.15}$$

Similarly, the right-hand side of (6.14) is equivalent to

$$rQ(h)[u(h,\overline{x}_h) - u(h,\underline{x}_h)] \geq (1-r)Q(k)[u(k,\overline{x}_k) - u(k,\underline{x}_k)] \quad \forall u \in U. \tag{6.16}$$

Now, define $\tilde{r} = \inf\{r \in [0,1] \mid (6.15) \text{ holds}\}$. Such an infimum must exist, since the set $\{r \in [0,1] \mid (6.15) \text{ holds}\}$ is nonempty and bounded. Suppose $\tilde{r} = 0$. Then $\{\boldsymbol{a}_{1/n}\}$ and $\{\boldsymbol{b}_{1/n}\}$ would be two sequences with $\boldsymbol{a}_{1/n} \succeq \boldsymbol{b}_{1/n}$

for all $n \in \mathbb{N}$; hence Sequence-Continuity would imply that $\boldsymbol{a}_0 \succeq \boldsymbol{b}_0$, which (given that P is positive) contradicts our assumptions about $\overline{x}_k$ and $\underline{x}_k$. So $\tilde{r} > 0$. Similarly, suppose $\tilde{r} = 1$. Then

$$\sup\left\{ \frac{u(k, \overline{x}_k) - u(k, \underline{x}_k)}{u(h, \overline{x}_h) - u(h, \underline{x}_h)} \mid u \in U \right\} = \infty. \tag{6.17}$$

Since $U \in U^*$, $1 \geq u(k, \overline{x}_k) - u(k, \underline{x}_k)$ for all u in U. Hence

$$\inf\{u(h, \overline{x}_h) - u(h, \underline{x}_h) \mid u \in U\} = 0, \tag{6.18}$$

which contradicts our assumptions about $\overline{x}_h$ and $\underline{x}_h$. So $1 > \tilde{r}$.

Moreover, given Sequence-Continuity we have $\boldsymbol{a}_{\tilde{r}} \succeq \boldsymbol{b}_{\tilde{r}}$. However, since $P(h) > Q(h)$ and $P(k) < Q(k)$, we know that (6.16), and the right-hand side of (6.14), cannot hold for $r = \tilde{r}$. So (i) cannot be true for Q.[1] ■

The Subjectivist Expected Value Theorem for Incompleteness

By the same reasoning as in the derivation of the Expected Value Theorem for Incompleteness, the Subjectivist Expected Value Theorem for Incompleteness follows immediately from Theorem 5.2. ■

[1] A similar proof is given by Karni and Schmeidler (1980, 12f.) for the uniqueness of the probability distribution in Karni and Schmeidler's Theorem 2. Since they are concerned with complete relations, their proof features simple utility functions instead of sets. Apart from that, the above proof mirrors theirs.

Bibliography

Aboodi, R., Borer, A., and Enoch, D. (2008). Deontology, individualism, and uncertainty, a reply to Jackson and Smith. *Journal of Philosophy*, 105(5):259–272.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 21(4):503–546.

Anscombe, F. and Aumann, R. (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, 34:199–205.

Arrhenius, G. (forthcoming). *Population Ethics. The Challenge of Future Generations*. Oxford University Press, Oxford.

Arrow, K. (1951, 2nd ed. 1963). *Social Choice and Individual Values*. Wiley, New York.

Bales, R. E. (1971). Act-utilitarianism: Account of right-making characteristics or decision-making procedure? *American Philosophical Quarterly*, 8(3):257–265.

Basu, K. (1983). Cardinal utility, utilitarianism, and a class of invariance axioms in welfare analysis. *Journal of Mathematical Economies*, 12:193–206.

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future*. PhD thesis, Rutgers University.

Binmore, K. and Voorhoeve, A. (2003). Defending transitivity against zeno's paradox. *Philosophy and Public Affairs*, 31:272–279.

Blackorby, C., Donaldson, D., and Weymark, J. A. (1984). Social choice with interpersonal utility comparisons: a diagrammatic introduction. *International Economic Review*, 25(2):327–356.

Bossert, W. (1991). On intra- and interpersonal utility comparisons. *Social Choice and Welfare*, 8:207–219.

Bossert, W., Blackorby, C., and Donaldson, D. (2005). *Population issues in social choice theory, welfare economics, and ethics*. Cambridge University Press, Cambridge.

Bossert, W. and Stehling, F. (1994). On the uniqueness of cardinally interpreted utility functions. In Eichhorn, W., editor, *Models and Measurement of Welfare and Inequality*. Springer, Berlin.

Bossert, W. and Weymark, J. A. (2004). Utility in social choice. In Barbera, S., Hammond, P., and Seidl, C., editors, *Handbook of Utility Theory. Volume 2: Extensions*. Kluver, Dordrecht.

Brentano, F. (1969 [1889]). *The Origin of Our Knowledge of Right and Wrong*. trans. R. Chisholm. Routledge and Kegan Paul, London.

Broad, C. D. (1930). *Five Types of Ethical Theory*. Routledge and Kegan Paul, London.

Broome, J. (1990). Fairness. *Proceedings of the Aristotelian Society*, 91:87–102.

Broome, J. (1991). *Weighing Goods*. Oxford University Press, Oxford.

Broome, J. (2004). *Weighing Lives*. Oxford University Press, Oxford.

Broome, J. (2012). *Climate Matters. Ethics in a Warming World*. W.W. Norton and Company, New York.

Broome, J. (2013). *Rationality Through Reasoning*. Oxford University Press, Oxford.

Buchak, L. (2013). *Risk and Rationality*. Oxford University Press, Oxford.

Bykvist, K. (2009a). No good fit: Why the fitting attitude analysis of value fails. *Mind*, 118(1-30).

Bykvist, K. (2009b). Objective versus subjective moral oughts. *Logic, Ethics, and All That Jazz. Essays in Honour of Jordan Howard Sobel. Uppsala Philosophical Studies*, 57.

Bykvist, K. and Olson, J. (2009). Expressivism and moral certitude. *Philosophical Quarterly*, 59(235):202–215.

Christensen, D. (2001). Preference-based arguments for probabillism. *Philosophy of Science*, 68:356–376.

Cotton-Barratt, O., MacAskill, W., and Ord, T. (ms). Normative uncertainty, intertheoretic comparisons, and variance normalisation. Manuscript.

Crisp, R. (2000). Review of 'value ... and what follows', by joel kupperman. *Philosophy*, 75:458–462.

Crisp, R. (2006). Hedonism reconsidered. *Philosophy and Phenomenological Research*, 73(3):619–645.

D'Arms, J. and Jacobson, D. (2000). The moralistic fallacy: On the 'appropriateness' of emotions. *Philosophy and Phenomenological Research*, 61(1):65–90.

D'Aspremont, C. and Gevers, L. (1977). Equity and the informational basis of collective choice. *The Review of Economic Studies*, 44(2):199–209.

de Finetti, B. (1980). Foresight. its logical laws, its subjective sources. In Jr., H. E. K. and Smokler, H. E., editors, *Studies in Subjective Probability*. Robert E. Krieger Publishing Company, Malabar.

Diamond, P. (1967). Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: A comment. *Journal of Political Economy*, 75(5):765–766.

Dovidio, J. F. and Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, 11(4):315–319.

Dreier, J. (1993). Structures of normative theories. *The Monist*, 76(1):22–40.

Dreier, J. (2011). In defense of consequentializing. In Timmons, M., editor, *Oxford Studies in Normative Ethics*, volume 1, pages 97–118. Oxford University Press, Oxford.

Drèze, J. (1987). Decision theory with moral hazard and statedependent preferences. In *Essays on Economic Decisions under Uncertainty*. Cambridge University Press, Cambridge.

Elga, A. (2010). Subjective probabilities should be sharp. *Philosophers' Imprint*, 10(05).

Eriksson, L. and Hájek, A. (2007). What are degrees of belief? *Studia Logica*, 86(2):183–213.

Ewing, A. C. (1947). *The Definition of Good*. Macmillan, London.

Fabre, C. (2012). *Cosmopolitan War*. Oxford University Press, Oxford.

Feldman, F. (2006). Actual utility, the objection from impracticality, and the move to expected utility. *Philosophical Studies*, 129(1):49–79.

Fishburn, P. (1970). *Utility Theory for Decision Making*. John Wiley and Sons Inc., New York.

Fleurbaey, M. and Voorhoeve, A. (2013). Decide as you would with full information! an argument against ex ante pareto. In Eyal, N., Hurst, S., Norheim, O., and Wikler., D., editors, *Inequalities in Health: Concepts, Measures, and Ethics*. Oxford University Press, Oxford.

Gaertner, S. L. and McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, 46(1):23–30.

Gibbard, A. (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Clarendon, Oxford.

Gilboa, I., Samuelson, L., and Schmeidler, D. (2014). No-betting-pareto dominance. *Econometrica*, 82(4):1405–1442.

Good, I. (1967). On the principle of total evidence. *British Journal for the Philosophy of Science*, 17(4):319–321.

Gracely, E. (1996). On the noncomparability of judgments made by different ethical theories. *Metaphilosophy*, 27(3):327–332.

Greaves, H. (forthcoming). Antiprioritarianism. *Utilitas*.

Greaves, H. (ms). A reconsideration of the harsanyi-sen-weymark debate on utilitarianism. Manuscript.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.

Greenwald, A. G. and Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27.

Guerrero, A. A. (2007). Don't know, don't kill: Moral ignorance, culpability, and caution. *Philosophical Studies*, 136(1):59–97.

Gustafsson, J. E. and Torpman, T. O. (2014). In defence of my favourite theory. *Pacific Philosophical Quarterly*, 95(2):159–174.

Hájek, A. (2008). Arguments for – or against – probabilism? *British Journal for the Philosophy of Science*, 59(4):793–819.

Hájek, A. (2012). Interpretations of probability. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/probability-interpret/.

Hammond, P. (1979). Equity in two person situations: Some consequences. *Econometrica*, 47(5):1127–1135.

Harman, E. (forthcoming). The irrelevance of moral uncertainty. In Shafer-Landau, R., editor, *Oxford Studies in Metaethics*. Oxford University Press, Oxford.

Harsanyi, J. (1955). Cardinal welfare, individualistic ethics and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309–321.

Hedden, B. (forthcoming). Does mite make right? on decision-making under normative uncertainty. In Shafer-Landau, R., editor, *Oxford Studies in Metaethics*. Oxford University Press, Oxford.

Howard-Synder, F. (1997). The rejection of objective consequentialism. *Utilitas*, 9(2):241–248.

Hudson, J. (1989). Subjectivization in ethics. *American Philosophical Quarterly*, 26:221–229.

Huemer, M. (2008). In defence of repugnance. *Mind*, 117(468):899–933.

Jackson, F. (1991). Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics*, 101(3):461–482.

Jackson, F. and Smith, M. (2006). Absolutist moral theories and uncertainty. *Journal of Philosophy*, 103(6):267–283.

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

Joyce, J. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, 19(1):153–178.

Joyce, J. (2010). A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*, 24(1):281–323.

Karni, E. (1985). *Decision Making under Uncertainty: The Case of State-Dependent Preferences*. Harvard University Press, Cambridge, MA.

Karni, E. and Mongin, P. (2000). On the determination of subjective probability by choices. *Management Science*, 46(2):233–248.

Karni, E. and Schmeidler, D. (1980). An expected utility theory for state-dependent preferences. *Working Paper 48-80*, The Foerder Institute of Economic Research, Tel Aviv University, Tel Aviv.

Karni, E., Schmeidler, D., and Vind, K. (1983). On state-dependent preferences and subjective probabilities. *Econometrica*, 51(4):1021–1031.

Kolodny, N. and MacFarlane, J. (ms). Ought: Between subjective and objective. Manuscript.

Lewis, D. (1980). A subjectivist's guide to objective chance. In Jeffrey, R., editor, *Studies in Inductive Logic and Probability*, volume 2. University of Berkeley Press, Berkeley.

List, C. (2003). Are interpersonal comparisons of utility indeterminate? *Erkenntnis*, 58:229–260.

Lockhart, T. (2000). *Moral Uncertainty and Its Consequences*. Oxford University Press, Oxford.

Louise, J. (2004). Relativity of value and the consequentialist umbrella. *Philosophical Quarterly*, 54(217):518–536.

MacAskill, W. (2013). The infectiousness of nihilism. *Ethics*, 123(3):508–520.

MacAskill, W. (2014). *Decision-Making under Normative Uncertainty*. PhD thesis, University of Oxford.

Maher, P. (1993). *Betting on Theories*. Cambridge University Press, New York.

McClennen, E. F. (2009). The normative status of the independence principle. In Anand, P., Pattanaik, P., and Puppe, C., editors, *The Handbook of Rational and Social Choice*. Oxford University Press.

McMahan, J. (2010). The meat eaters. *The New York Times*, September 19, 2010.

Meacham, C. J. G. and Weisberg, J. (2011). Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, 89(4):641–663.

Mill, J. S. (1998 [1861]). *Utilitarianism*. R. Crisp, editor. Oxford University Press, New York.

Moller, D. (2011). Abortion and moral risk. *Philosophy*, 86:425–443.

Mongin, P. and D'Aspremont, C. (1998). Utility theory and ethics. In Barbera, S., Hammond, P., and Seidl, C., editors, *Handbook of Utility Theory. Volume 1: Principles*. Kluver, Dordrecht.

Nagel, T. (1979). The fragmentation of value. In *Mortal Questions*. Cambridge University Press, Cambridge.

Nau, R. (2006). The shape of incomplete preferences. *The Annals of Statistics*, 34(5):2430–2448.

Ng, Y.-K. (1997). A case for happiness, cardinalism, and interpersonal comparability. *The Economic Journal*, 107:1848–1858.

Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books, New York.

Oddie, G. (1994). Moral uncertainty and human embryo experimentation. In Fulford, K. W. M., Gillet, G., and Soskice, J. M., editors, *Medicine and Moral Reasoning*. Cambridge University Press, Cambridge.

Ok, E. A., Ortoleva, P., and Riella, G. (2012). Incomplete preferences under uncertainty: Indecisiveness in beliefs vs. tastes. *Econometrica*, 80(4):1791–1808.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press, Oxford.

Parfit, D. (ms). How we can avoid the repugnant conclusion. Manuscript.

Paseau, A. (ms). Sketch of a cardinal utility foundation for economics. Manuscript.

Pfeiffer, R. S. (1985). Abortion policy and the argument from uncertainty. *Social Theory and Practice*, 11:371–386.

Pyke, S. (1993). *Philosophers*. Cornerhouse Publications, Manchester.

Rabinowicz, W. (2008). Value relations. *Theoria*, 74(1):18–49.

Rabinowicz, W. (2012). Value relations revisited. *Economics and Philosophy*, 28(2):133–164.

Rachels, S. (1998). Counterexamples to the transitivity of better than. *Australasian Journal of Philosophy*, 76(1):71–83.

Ramsey, F. P. (1990 [1926]). Truth and probability. In Mellor, D., editor, *F. P. Ramsey: Philosophical Papers*. Cambridge University Press.

Raz, J. (1986). *The Morality of Freedom*. Clarendon, Oxford.

Richardson, H. (1994). *Practical Reasoning about Final Ends*. Cambridge University Press, Cambridge.

Roberts, K. W. (1980a). Interpersonal comparability and social choice theory. *The Review of Economic Studies*, 47(2):421–439.

Roberts, K. W. (1980b). Social choice theory: The single-profile and multi-profile approaches. *The Review of Economic Studies*, 47(2):441–450.

Roberts, M. (2003). Is the person-affecting intuition paradoxical? *Theory and Decision*, 55(1):1–44.

Rosen, G. (2003). Culpability and ignorance. *Proceedings of the Aristotelian Society*, 103(1):61–84.

Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives*, 18(1):295–313.

Ross, J. (2006a). *Acceptance and Practical Reason*. PhD thesis, Rutgers University.

Ross, J. (2006b). Rejecting ethical deflationism. *Ethics*, 116:742–768.

Ross, W. D. (2002 [1930]). *The Right and the Good*. Oxford University Press, Oxford.

Savage, L. (1954). *The Foundations of Statistics*. Dover, New York.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Harvard University Press, Cambridge, MA.

Schroeder, M. (2010). Value and the right kind of reason. In Shafer-Landau, R., editor, *Oxford Studies in Metaethics*, volume 5, pages 25–55. Oxford University Press.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf

between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, 91(4):531–553.

Schwitzgebel, E. (2014). The moral behavior of ethicists and the role of the philosopher. In Luetge, C., Rusch, H., and Uhl, M., editors, *Experimental Ethics. Toward an Empirical Moral Philosophy*. Palgrave Macmillan.

Schwitzgebel, E. and Rust, J. (2014). The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude, and directly observed behavior. *Philosophical Psychology*, 27(3):293–327.

Sepielli, A. (2006). Review of 'moral uncertainty and its consequences', by ted lockhart. *Ethics*, 116(3):599–604.

Sepielli, A. (2009). What to do when you don't know what to do. In Shafer-Landau, R., editor, *Oxford Studies in Metaethics*, volume 4, pages 5–28. Oxford University Press, Oxford.

Sepielli, A. (2010). *'Along an Imperfectly Lighted Path': Practical Rationality and Normative Uncertainty*. PhD thesis, Rutgers University.

Sepielli, A. (2012). Normative uncertainty for non-cognitivists. *Philosophical Studies*, 160(2):191–207.

Sepielli, A. (2013a). What to do when you don't know what to do when you don't know what to do.... *Noûs*, 47(1):521–544.

Sepielli, A. (2013b). Moral uncertainty and the principle of equity among moral theories. *Philosophy and Phenomenological Research*, 86(3):580–589.

Sepielli, A. (2014). Should you look before you leap? *The Philosophers' Magazine*, 66:89–93.

Sidgwick, H. (1907). *The Methods of Ethics*. Macmillan, London, 7 edition.

Smith, M. (1994). *The Moral Problem*. Basil Blackwell, Oxford.

Smith, M. (2002). Evaluation, uncertainty and motivation. *Ethical Theory and Moral Practice*, 5(3):305–320.

Strohminger, N., Caldwell, B., Cameron, D., Borg, J. S., and Sinnott-Armstrong, W. (2014). Implicit morality: A methodological survey. In Luetge, C., Rusch, H., and Uhl, M., editors, *Experimental Ethics. Toward an Empirical Moral Philosophy*. Palgrave Macmillan.

Temkin, L. (2012). *Rethinking The Good. Moral Ideals and the Nature of Practical Reasoning*. Oxford University Press, Oxford.

Thomson, J. J. (1990). *The Realm of Rights*. Harvard University Press, Cambridge, MA.

Vallentyne, P. (1993). The connection between prudential and moral goodness. *Journal of Social Philosophy*, 24(2):105–128.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

Voorhoeve, A. (2013). Vaulting intuition: Temkin's critique of transitivity. *Philosophy and Economics*, 29(3):409–425.

Weatherson, B. (2002). Review of 'moral uncertainty and its consequences', by ted lockhart. *Mind*, 111:693–696.

Weatherson, B. (2014). Running risks morally. *Philosophical Studies*, 167(1):141–163.

White, R. (2010). Evidential symmetry and mushy credence. In Gendler, T. S. and Hawthorne, J., editors, *Oxford Studies in Epistemology*, volume 3, pages 161–186. Oxford University Press, Oxford.

Wiggins, D. (1987). *Needs, Values, Truth: Essays in the Philosophy of Value*. Blackwell, Oxford.

Wiland, E. (2005). Monkeys, typewriters, and objective consequentialism. *Ratio*, 18(3):352–360.

Willenken, T. (2012). Deontic cycling and the structure of commonsense morality. *Ethics*, 122(3):545–561.

Zimmerman, M. (2008). *Living with Uncertainty*. Cambridge University Press, Cambridge.

Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science*, 67:45–69.