Normative Uncertainty

William MacAskill
St Anne's College
University of Oxford
February 2014

Thesis for the degree of Doctor of Philosophy

Word Count: 66,627

Abstract

Very often, we are unsure about what we ought to do. Under what conditions should we help to improve the lives of distant strangers rather than those of our family members? At what point does an embryo or foetus become a person, with all the rights that that entails? Is it ever permissible to raise and kill non-human animals in order to use their meat for food?

Sometimes, this uncertainty arises out of *empirical* uncertainty: we might not know to what extent non-human animals feel pain, or how much we are really able to improve the lives of distant strangers compared to our family members. But this uncertainty can also arise out of fundamental normative uncertainty: out of not knowing, for example, what moral weight the wellbeing of distant strangers has compared to the wellbeing of our family; or whether non-human animals are worthy of moral concern even given knowledge of all the facts about their biology and psychology.

In fact, for even moderately reflective agents, decision-making under normative uncertainty is ubiquitous. Given this, one might have expected philosophers to have devoted considerable research time to the question of how one ought to take one's normative uncertainty into account in one's decisions. But the issue has been largely neglected.

This thesis attempts to begin to fill this gap. It addresses the question: what ought one to do when one is uncertain about what one ought to do? It develops a view that I call *metanormativism*: the view that there are second-order norms that govern action that are relative to a decision-maker's uncertainty about first-order normative claims. In consists of two distinct parts.

The first part (Chapters 1-4) develops a general metanormative theory. I argue in favour of the view that decision-makers should *maximise expected choice-worthiness*, treating normative uncertainty analogously with how they treat empirical uncertainty. I defend this view at length in response to two key problems, which I call the problems of *merely ordinal theories* and the problem of *intertheoretic comparisons*.

The second part (Chapters 5-7) explores the implications of metanormativism for other philosophical issues. I suggest that it has important implications for the theory of rational action in the face of incomparable values, for the causal/evidential debate in decision-theory, and for the value we should ascribe to research into moral philosophy.

Acknowledgements

I am indebted to a very large number of people.

For the generous financial support that enabled me to pursue the BPhil and the DPhil, and to spend time as a visiting student at both NYU and Princeton, I thank the Pirie-Reid Fund, St Edmund Hall, St Anne's College, the Arts and Humanities Research Council, The St Andrew's Society of Washington D.C., the Fulbright Commission, and the Society for Applied Philosophy.

This thesis was supervised by Krister Bykvist and John Broome. They have provided an incredible number of insightful comments on every aspect of my work, and have provided vital guidance on the most central components of this thesis. It's been an honour to work with them.

For their comments on various aspects of this work, I thank: Arif Ahmed, Gustaf Arrhenius, Frank Arntzenius, Ralf Bader, Nick Beckstead, Rachael Briggs, Patrick Butlin, Owen Cotton-Barratt, Daniel Deasy, Max Edwards, Andy Egan, Ben Eidelson, Adam Elga, Hilary Greaves, Johan Gustafsson, Alan Hajek, Elizabeth Harman, Michelle Hutchinson, Harvey Lederman, Amanda MacAskill, George Marshall, Alexander Mascolo, Chris Maughan, Andreas Mogensen, Graham Oddie, Mike Otsuka, Zee Perry, Stefan Riedener, Erica Shumener, Robert Simpson, Peter Singer, Ben Plommer, Michael Smith, Bastian Stern, Trevor Teitel, James Ting-Edwards, Tom Olle Torpman, Aron Vallinder, Gerard Vong, Ralph Wedgwood, Alex Worsnip, and Charlotte Wright. I also thank audiences at the Ockham Society, the Oxford Jurisprudence Discussion Group, the Oxford Dissertation Seminar, the Stockholm Centre for Healthcare Ethics, the Rocky Mountain Ethics Congress, Princeton's Economics and Philosophy workshop, Princeton's Moral Epistemology workshop, Oxford's CRNAP workshop, Princeton's DeCamp Bioethics Seminar, and the LSE. I owe a particular debt of gratitude to Toby Ord. If it were not for our five-hour long first meeting in a graveyard in St Edmund Hall (and many subsequent multi-hour meetings after that) I would not have had the confidence to pursue this topic, and would not have been exposed to the quite dazzling number of insights he has had on these issues.

For their acceptance and even encouragement of my choice of career as a philosopher, I thank my parents, Mair and Robin Crouch, and my brothers, Iain and Tom. I thank my many friends (Scottish, Cantabrigian, Oxonian and American) for keeping me sane.

Finally, and most importantly, for her support and encouragement, her keen intellect and her remarkable creativity, I thank my wife, Amanda MacAskill.

Table of Contents

Introduction

Normative uncertainty is a fact of life.

Suppose that I have £20 to spend. With that money, I could eat out at a delightful Indian restaurant. Or I could pay for four long-lasting insecticide-treated bednets that would protect eight children against malaria. In comparing these two options, let us suppose that I know all the morally relevant facts about what that £20 could do. Even so, I still don't know whether I'm obligated to donate that money or whether it's permissible for me to pay for the meal out, because I just don't know how strong my moral obligations to distant strangers are. So I don't ultimately know what I ought to do.

For an example of normative uncertainty on a larger scale, suppose that the members of a government are making a decision about whether to tax carbon emissions. They know, let us suppose, all the relevant facts about what would happen as a result of the tax: it would make presently existing people worse off, as they would consume less oil and coal, and therefore be less economically productive; but it would slow the onset of climate change, thereby increasing the welfare of people living in the future. But the members of the government don't know how to weigh the interest of future people against the interests of presently existing people. So, again, those in this government don't ultimately know what they ought to do.

In both of these cases, the uncertainty in question is not uncertainty about what will happen, but rather is fundamental normative uncertainty. Recently, some philosophers have suggested that there are norms that govern how one ought to act that take into

account one's fundamental normative uncertainty. [1] I call this suggestion *metanormativism*.[2] In this thesis, I first argue in favour of a particular metanormative theory, and then discuss some implications of metanormativism for other philosophical issues.

In this introduction I clarify the nature of my project. In section I, I give the basic motivation for metanormativism. In section II, I provide the framework within which this thesis proceeds. In section III, I give a synopsis of the rest of this thesis.

I. The motivation for metanormativism

There are two main motivations for metanormativism. The first is simply an appeal to intuitions about cases. Consider the following example:

> *Moral Dominance*
>
> Jane is at dinner, and she can either choose foie gras, or the vegetarian risotto. Let's suppose that, according to the true moral theory, both of these options are equally choice-worthy: animal welfare is not of moral value so there is no moral reason for choosing one meal over another, and Jane would find either meal

[1] For example: (Guerrero 2007; Lockhart 2000; Oddie 1995; Ross 2006; Sepielli 2009). John Broome

[2] A note on terminology: Metanormativism isn't *about* normativity, in the way that meta-ethics is about ethics, or that a meta-language is about a language. Rather, 'meta' is used in the sense of 'over' or 'beyond': that is, in the sense used in the word 'metacarpal', where, the metacarpal bones in the hand are located beyond the carpal bones. Regarding metanormativism, there is a clear analogy with the debate about the subjective or objective ought in moral theory (that is, whether moral norms are evidence-relative or belief-relative in some way). However, using the term 'normative subjectivism' instead of 'metanormativism' would have had misleading associations with subjectivism in meta-ethics. So I went with 'metanormativism' – with the caveat that this shouldn't be confused with the study *of* normativity.

> equally tasty, and so she has no prudential reason for preferring one over the other. Let's suppose that Jane has high credence in that view. But she also finds plausible the view that animal welfare is of moral value, according to which the risotto is the more choice-worthy option.

In this situation, choosing the risotto over the foie gras is more choice-worthy according to some moral views in which she has credence, and less choice-worthy according to none. In the language of decision-theory, the risotto *dominates* the foie gras. So it seems very clear that, in some sense of 'ought', Jane ought to choose the risotto, and ought not to buy the foie gras.[3] But, if so, then there must be a sense of 'ought' that takes into account Jane's first-order normative uncertainty.

A second motivation for metanormativism is based on the idea of action-guidingness. There has been a debate concerning whether there is a sense of 'ought' that is relative to the decision-maker's beliefs or credences (a 'subjective' sense of ought), in addition to a sense of 'ought' that is not relative to the decision-maker's beliefs or credences (an 'objective' sense of ought).

The principal argument for thinking that there must be a subjective sense of 'ought' is because the objective sense of 'ought' is not sufficiently action-guiding. Consider the following case:[4]

[3] In what follows I will talk about different 'senses' of the term 'ought'. There is an extensive debate about the semantics of 'ought', which I do not want to enter into here. My own view is that 'ought' is context-sensitive: (Wedgwood ms) gives what seems to be a plausible semantics. But nothing turns on the precise account of the semantics of 'ought'.

[4] This is a 'Jackson case', from (Jackson 1991, 462–3). (Regan 1980, 264–5) gives a similar example.

Susan, and the Medicine - I

Susan is a doctor, who has a sick patient, Greg. Susan is unsure whether Greg has condition X or condition Y: she thinks each possibility is equally likely. And it is impossible for her to gain any evidence that will help her improve her state of knowledge any further. She has a choice of three drugs that she can give Greg: drugs A, B, and C. If she gives him drug A, and he has condition X, then he will be completely cured; but if she gives him drug A, and he has condition Y, then he will die. If she gives him drug C, and he has condition Y, then he will be completely cured; but if she gives him drug C, and he has condition X, then he will die. If she gives him drug B, then he will be almost completely cured, whichever condition he has, but not completely cured.

Her decision can be represented in the following table, using numbers to represent how good each outcome would be:

	Greg has condition X – 50%	Greg has condition Y – 50%
A	100	0
B	99	99
C	0	100

Finally, suppose that, as a matter of fact, Greg has condition Y. So giving Greg drug C would completely cure him. What should Susan do?

In *some* sense, it seems that Susan ought to give Greg drug C: doing so is what will actually cure Greg. But given that she doesn't *know* that Greg has condition Y, it seems that it would be reckless for Susan to administer drug C. As far as she knows, in doing

so she'd be taking a 50% risk of Greg's death. Ands so it also seems that there's a sense of 'ought' according to which she ought to administer drug B.

In this case, the objective consequentialist's recommendation — "do what actually has the best consequences" — is not useful advice for Susan. It is not a piece of advice that the she can act on, because she does not know, and is not able to come to know, what action actually has the best consequences. So one might worry that the objective consequentialist's recommendation is not sufficiently action-guiding: it's very rare that a decision-maker will be in a position to know what she ought to do. In contrast, so the argument goes, if there is a subjective sense of 'ought' then the decision-maker will very often know what she ought to do. So the thought that there should be at least some sense of 'ought' that is sufficiently action-guiding motivates the idea that there is a subjective sense of 'ought'.

Similar considerations motivate metanormativism. Just as one is very often not in a position to know what the consequences of one's actions are, one is very often not in a position to know which moral norms are true; in which case a sufficiently action-guiding sense of 'ought' must take into account normative uncertainty as well. Consider the following variant of the case:[5]

> *Susan and the Medicine - II*
>
> Susan is a doctor, who faces three sick individuals, Greg, Harold and Harry. Greg is a human patient, whereas Harold and Harry are chimpanzees. They all suffer from the same condition. She has a vial of a drug, D. If she administers all

[5] This is a Jackson case under moral uncertainty. A similar case is given by (Zimmerman 2008, 35).

of drug D to Greg, he will be completely cured, and if she administers all of drug to the chimpanzees, they will both be completely cured (health 100%). If she splits the drug between the three, then Greg will be almost completely cured (health 99%), and Harold and Harry will be partially cured (health 50%). She is unsure about the value of the welfare of non-human animals: she thinks it is equally likely that chimpanzees' welfare has no moral value and that chimpanzees' welfare has the same moral value as human welfare. And, let us suppose, there is no way that she can improve her epistemic state with respect to the relative value of humans and chimpanzees.

Using numbers to represent how good each outcome is: Sophie is certain that completely curing Greg is of value 100 and that partially curing Greg is of value 99. If chimpanzee welfare is of moral value, then curing one of the chimpanzees is of value 100, and partially curing one of the chimpanzees is of value 50.

Her three options are as follows:

A: Give all of the drug to Greg

B: Split the drug

C: Give all of the drug to Harold and Harry

Her decision can be represented in the following table, using numbers to represent how good each outcome would be.

	Chimpanzee welfare is of no moral value – 50%	Chimpanzee welfare is of significant moral value – 50%
A	100	0
B	99	199
C	0	200

Finally, suppose that, according to the true moral theory, chimpanzee welfare is of the same moral value as human welfare and that therefore, she should give all of the drug to Harold and Harry. What should she do?

In the first variant of this case, intuitively Susan would be reckless not to administer drug B. Analogously, in the case above, it seems it would be morally reckless for Susan not to choose option B: given what she knows, she would be risking severe wrongdoing by choosing either option A or option C. Moreover, 'do what's right given your empirical credences only' isn't useful advice for Susan, because it doesn't take into account her uncertainty over firs-order normative views. So, it seems that, in this case, the subjective sense of 'ought' isn't sufficiently action-guiding. And this motivates the idea that there is a sense of 'ought' that takes into account both one's empirical uncertainty and one's normative uncertainty as well. That is, it motivates metanormativism.

For these reasons, I think that there is a clear *prima facie* case in favour of the idea that there are norms that take first-order normative uncertainty into account. In this thesis, I do not argue further for this idea. My main intention is to explore this idea, rather than defend it at length.

II. My Framework

In this section, I define terms and introduce the framework within which I work. I will generally try my best to avoid unnecessary formalism in this thesis, but I state my framework formally in this section for precision.

What I will call a *decision-situation* is a quintuple $\langle S, t, \mathcal{A}, \mathcal{T}, C \rangle$, where *S* is a decision-maker, *t* is a time, $\mathcal{A}$ is a set of options that the decision-maker has the power to bring about at that time, $\mathcal{T}$ is the set, assumed to be finite, of first-order normative theories considered by the decision-maker at that time, and C is a credence function representing the decision-maker's beliefs about first-order normative theories. I'll now elaborate on each of the last three members of this quintuple in turn.

First, options. An *option* is a proposition, understood as a set of centred possible worlds, that the decision-maker has the power to make true at a time. A centred possible world is a triple of a world, an agent in that world, and a time in the history of that world: an option therefore can include the action available to the decision-maker, as well as the decision-maker's intention, motive, the outcome of the action, and everything else that could be normatively relevant to the decision-maker's decision. By convention, the options in $\mathcal{A}$ are specified so as to be mutually exclusive and jointly exhaustive. I will use the letters A, B, C etc to refer to specific options, and italicised letters *A, B, C* etc to refer to variables.

Second, normative theories. A *first-order normative theory* is a function from decision situations to an ordering of the options in the set $\mathcal{A}$ in terms of their choice-worthiness, where the function's domain is all possible decision situations. The theories in $\mathcal{T}$ form a

partition. I take choice-worthiness to be defined as the ordering that first-order normative theories produce. I use "$A \succcurlyeq_{T_i} B$" to mean "*A* is at least as choice-worthy as *B*, according to T_i", and I define "*A* is strictly more choice-worthy than *B*, according to T_i" (or "$A \succ_{T_i} B$") as $\left(A \succcurlyeq_{T_i} B\right)$ & $B \not\succcurlyeq_{T_i} A$, and "A is equally as choice-worthy as B, according to T_i" (or "$A \sim_{T_i} B$") as $\left(A \succcurlyeq_{T_i} B\right)$ & $\left(B \succcurlyeq_{T_i} A\right)$. For all of Part I of the thesis I assume that the relation '*A* is at least as choice-worthy as *B*' is reflexive (for all *A*, $A \succcurlyeq A$), transitive (for all *A, B, C,* if $A \succcurlyeq B$ and $B \succcurlyeq C$ then $A \succcurlyeq C$) and complete (for all *A*, *B*, either $A \succcurlyeq B$ or $B \succcurlyeq A$). I discuss problems arising from theories that give incomplete choice-worthiness orderings in chapter 5. Note that by 'choice-worthiness' I do not mean merely *moral* choice-worthiness, in the sense in which the moral choice-worthiness of an act might be weighed against the prudential choice-worthiness of it. I mean *all-things-considered* choice-worthiness: the normative ordering of options after all normatively relevant considerations have been taken into account (according to a particular first-order normative theory). As I use the terms, a theory regards *A* as 'permissible' iff *A* is maximally choice-worthy.[6]

First-order normative theories' choice-worthiness orderings can be represented by a *choice-worthiness function*, which is a function from options to numbers such that $CW_i(A) \geq CW_i(B)$ iff $A \succcurlyeq_{T_i} B$, using "$CW_i(A)$" to represent the numerical value that the choice-worthiness function CW_i assigns to A.

A normative theory (or just 'theory') provides a choice-worthiness function, though does not necessarily merely consist of a choice-worthiness function. (It will become important,

[6] In this thesis, I do not consider how to accommodate theories that involve supererogation. This is an important issue, but one that must be left for another time. So I only discuss theories according to which all permissible options are maximally choice-worthy.

in chapter 4, that normative theories often do not merely consist of a choice-worthiness ordering, but also come along with an accompanying metaphysical account of the nature of value and choice-worthiness.)

Normative theories can vary in the information that they provide. First, a theory can be *merely ordinal*. Merely ordinal theories rank options as 1st, 2nd, 3rd (etc) in terms of choice-worthiness. But they don't give any information about *how much* more choice-worthy the most choice-worthy option is, rather than the second most choice-worthy. More precisely: Let $\mathrm{CW}_i(A)$ represent $\succcurlyeq_{T_i}$. If $\mathrm{T_i}$ is merely ordinal, then $\mathrm{CW}_j(A)$ also represents $\succcurlyeq_{T_i}$ iff $\mathrm{CW}_i(A) = \mathrm{f}(\mathrm{CW}_j(A))$, where f(x) is any increasing transformation.

Second, a theory can be *cardinal*. Cardinal theories not only give us ordinal information about choice-worthiness, they also tell us the ratio of differences in choice-worthiness between options, giving sense to the intuitive idea of *how much* more choice-worthy one option is rather than another. More precisely: Let $\mathrm{CW}_i(A)$ represent $\succcurlyeq_{T_i}$. If $\mathrm{T_i}$ is cardinal, then $\mathrm{CW}_j(A)$ also represents $\succcurlyeq_{T_i}$ iff $\mathrm{CW}_i(A) = \mathrm{a}\mathrm{CW}_j(A) + \mathrm{b}$, where $\mathrm{a} > 0$.

Third, a theory can be *ratio-scale*. Ratio-scale theories give meaning to ratios between the absolute levels of choice-worthiness of options. More precisely: Let $\mathrm{CW}_i(A)$ represent $\succcurlyeq_{T_i}$. If $\mathrm{T_i}$ is ratio-scale, then $\mathrm{CW}_j(A)$ also represents $\succcurlyeq_{T_i}$ iff $\mathrm{CW}_i(A) = \mathrm{a}\mathrm{CW}_j(A)$, where $\mathrm{a} > 0$. In this thesis I do not discuss ratio-scale measurable theories, because I ultimately argue in favour of *maximise expected choice-worthiness*, which does not require ratio-scale

measurable choice-worthiness.[7] If one has credence in ratio-scale theories, they should be treated in the same way as cardinal theories.

The fifth element in a decision-situation is the credence function. The credence function C is a representation of the decision-makers partial beliefs over first-order normative theories. The credence function is a function from every theory $T_i \in \mathcal{T}$ to a real number in the interval [0,1], such that the sum of credences across all theories in $\mathcal{T}$ equals 1. C is a probability function that satisfies the Kolmogorov axioms. With the exception of chapter 6 in both cases, for simplicity I assume empirical certainty, and I assume that the decision-makers' credences across theories are evidentially independent of her actions; nothing will hang on this in what follows. One might worry that, if non-cognitivism true, then one cannot make sense of credences over normative theories. There has been debate on this issue in the literature.[8] However, I wish to sidestep this debate in this thesis, so for the purpose of this thesis I assume that cognitivism is true.

Having elucidated decision-situations, we can move on to what I call *metanormative theory*. A *metanormative theory* is a function from decision situations to an ordering of the options in the set $\mathcal{A}$ in terms of their appropriateness, using "$\succcurlyeq_A$" to represent the relation "is more appropriate than". As long as this relation is reflexive, transitive, and complete, it may be represented with an *appropriateness function*: a function from options to numbers

[7] Later in this thesis I will discuss comparability across theories. However I will only discuss comparability of units of choice-worthiness across theories, rather than comparability of levels of choice-worthiness. Again, this is because, with one important caveat, MEC only requires unit-comparability, and does nor require level-comparability. The caveat is that if we endorsed the evidential version of *maximize expected choice-worthiness*, then, if the decision-maker's credences in theories are sometimes not evidentially independent of her choice of options, then we *would* require level-comparability across normative theories. However, I ultimately (in chapter 6) defend the causal version of *maximize expected choice-worthiness*. So I don't need comparisons of levels across theories.

[8] (Smith 2002; Bykvist and Olson 2009; Sepielli 2012; Bykvist and Olson 2012).

such that $AP(A) \geq AP(B)$ iff $A \succcurlyeq_A B$, where "$AP(A)$" is the numerical value that the appropriateness function assigns to A. Note that a metanormative theory does not take into account uncertainty about which metanormative theory it is true, though I discuss this issue in chapter 6.

I introduce the technical term 'appropriateness' in order to remain neutral on the issue of whether metanormative norms are rational norms, or some other sort of norms (though noting that they can't be first-order norms provided by first-order normative theories, on pain of inconsistency).[9] I use the term 'appropriate' rather than 'ought' because 'ought' is a non-comparative concept: it doesn't make sense to say that an action *x* is more or less 'oughty' than another. But it does make sense to say that an action is more or less appropriate than another, and in this thesis I will sometimes be interested in how a metanormative theory ranks options that are not maximally appropriate. From an appropriateness ranking, with a little bit of extra argument, one can derive oughts: on the most natural view, if *A* is the most appropriate option, then one ought (in the sense of 'ought' that I'm interested in) to do *A*. When I say that *A* is an 'appropriate' option (using the term as a monadic predicate), I mean that it is a maximally appropriate option in the decision-situation under discussion.

The project of the first part of this thesis is to determine what the correct metanormative theory is. In the next chapter of this thesis, I will defend the idea that, when all theories in $\mathcal{T}$ are cardinally measurable and intertheoretically comparable, the

[9] See (Wedgwood 2013) for analysis of what I call appropriateness in terms of *enkrasia*. (Lockhart 2000, 24,26), (Sepielli 2009, 10) and (Ross 2006) all take metanormative norms to be norms of rationality. (Weatherson 2014) and (Harman 2014) both understand metanormative norms as moral norms. So there is an odd situation in the literature where the defenders of metanormavism (Lockhart, Ross, and Sepielli) and the critics of the view (Weatherson and Harman) seem to be talking past one another.

appropriateness of an option is given by its expected choice-worthiness, where the expected choice-worthiness (EC) of an option is as follows:[10]

$$EC(A) = \sum_{i=1}^{n} C(T_i)\, CW_i(A)$$

The appropriate options are those with the highest expected choice-worthiness. In line with the literature, I will refer to this view as *maximise expected choice-worthiness* (MEC).

III. Synopsis

The first part of my thesis is an attempt to work out what the correct metanormative theory is. In chapter 1, I consider what to do in the 'best-case' scenario: when choice-worthiness is cardinal and comparable across all theories in which the decision-maker has a non-zero credence. I argue that the correct metanormative theory in such a situation is MEC. However, we are not always in the best-case scenario. Sometimes, theories are merely ordinal, and, sometimes, even when theories are cardinal, choice-worthiness is not comparable between them. In either of these situations, MEC cannot be applied. In light of this problem, I propose that the correct metanormative theory is sensitive to the different sorts of information that different theories provide. In chapter 2, I consider how to take normative uncertainty into account in conditions where all theories provide merely ordinal choice-worthiness, and where choice-worthiness is non-comparable between theories, arguing in favour of the *Borda Rule*. In chapter 3, I

[10] Remembering that, for the time being, I am assuming that choice of options is evidentially independent of which theories are true.

consider how to take normative uncertainty into account in conditions where theories provide cardinal choice-worthiness, but where choice-worthiness is not comparable across theories, arguing in favour of what I call *Variance Voting*. I also discuss what it's appropriate to do in varying informational conditions. I suggest that my discussion in chapters 2 and 3 leads to a vindication of a type of MEC. In the final chapter of Part I, chapter 4, I discuss the problem of when, if ever, choice-worthiness is comparable across theories, and I give an account of what makes intertheoretic comparisons of choice-worthiness possible.

Part II of my thesis discusses three further issues that arise out of consideration of normative uncertainty. In chapter 5, I introduce a new problem for MEC: what I call the problem of infectious incomparability. I argue that this problem shows one prominent argument that has been given in the literature on normative uncertainty to be mistaken, but that the problem is ultimately surmountable. In chapter 6, I look at the implications of taking *decision-theoretic* uncertainty into account in one's decision-making. I suggest that doing so can provide a novel explanation of our divergent intuitions across Newcomb cases and similar cases that have been given in the literature, and can provide novel arguments against evidential decision theory. In chapter 7, I show how my framework can be used to show the value of studying and researching moral philosophy.

Part I: Metanormative Theory

Chapter 1: Maximising Expected Choice-Worthiness

Introduction

In this chapter I argue that, when all theories in which the decision-maker has non-zero credence are cardinally measurable and intertheoretically comparable, the appropriate options are those with the highest expected choice-worthiness. The structure of the chapter is as follows. In section I, I consider and argue against what I call the "My Favourite Theory" view. In section II, I consider and argue against what I call the "My Favourite Option" view. In section III, I give a general argument in favour of what I call *comparativism*. In section IV, I argue directly for the idea that, under normative uncertainty, it is appropriate to *maximise expected choice-worthiness*. In section V, I consider and respond to a number of objections to this view.

I. Against My Favourite Theory

One might think that, under moral uncertainty, one should simply follow the moral view that one thinks is most likely. This has been suggested as the correct principle by Edward Gracely, in one of the earliest modern papers on moral uncertainty:

> [T]he proper approach to uncertainty about the rightness of ethical theories is to determine the one most likely to be right, and to act in accord with its dictates.[11]

[11] (Gracely 1996, 301)

It has also recently been defended by Gustafsson and Torpman.[12] Making this more precise, we could define it as follows:

> *My Favourite Theory (MFT): A* is an appropriate option iff *A* is a permissible option according to the theory that *S* has highest credence in.

This is an elegant view. But it has major problems. I'll first mention two more minor issues that need to be straightened out, which rob the view of much of its elegance, before moving on to a dilemma that I believe ultimately sinks the view.

The first minor problem is that, sometimes, one will have equal highest credence in more than one moral theory. And what is it appropriate to do then? Picking one theory at random seems arbitrary. So, instead, one could claim that if *A* is permissible according to *any* of the theories in which one has highest credence then *A* is appropriate. But that has odd results too. Suppose that John is 50/50 split on a radical pro-life view and a pro-choice view. In which case, according to this version of MFT, it would be appropriate for John to try to sabotage abortion clinics on Wednesday (because doing so is permissible according to the radical pro-life view) and appropriate for John to punish himself for doing so on Thursday (because doing so is permissible according to the pro-choice view). But that seems bizarre.

The second minor problem is that the view violates the following principle:

> *Dominance*: If *A* is more choice-worthy than *B* according to some theories in which *S* has credence, and equally choice-worthy according to all other theories in which *S* has credence, then *A* is more appropriate than *B*.

[12] (J. Gustafsson and Torpman forthcoming).

MFT violates this in the following case:[13]

	T_1 – 40%	T_2 – 60%
A	permissible	permissible
B	impermissible	permissible

That is, according to MFT it's equally appropriate to choose either A or B, even though A is certainly permissible, whereas B is possibly impermissible. But there's no possible downside to choosing A, whereas there is a possible downside to choosing B. So it seems very plausible that it's appropriate to choose A and inappropriate to choose B.

These problems are bugs for the view, rather than fundamental objections. They can be overcome, at a cost of sacrificing some of the view's elegance, by modifying it slightly. This is what Gustafsson and Torpman do. Translating their proposal into my terminology, the version of MFT that they defend is as follows:

> *My Favourite Theory (Gustafsson and Torpman):* An option *A* is appropriate for *S* if and only if:
>
> 1. *A* is permitted by a moral theory T_i such that
>   a. T_i is in the set $\mathcal{T}'$ of moral theories that are at least as credible as any moral theory for *S* and
>   b. *S* has not violated T_i more recently than any other moral theory in $\mathcal{T}'$, and
> 2. There is no option *B* and no moral theory T_j such that

[13] I will often represent decision-situations using tables. In these tables, options are listed in the left-hand column, theories and the decision-maker's credences in those theories (represented using percentages) are listed in the top column, and all other cells in the table refer to the choice-worthiness (or normative status) of the option in the cell's row, according the theory in the cell's column.

a. T_j requires B and not A and

b. No moral theory that is at least as credible as T_j for S requires A and not B.

The first clause is designed to escape the problem of equal highest-credence theories. Clause 1(b) ensures that one does not perform bizarre courses of action: in the case above, if one sabotages the abortion clinic on Wednesday (following the radical pro-life view, but violating the pro-choice view), then it is not appropriate to prevent the sabotage of the abortion clinic on Thursday (because one has violated the pro-choice view more recently than any other view). The second clause is designed to escape the problem of violating Dominance, generating a lexical version of MFT. If one's favourite theory regards all options as permissible, then one goes with the recommendation of one's second-favourite theory; if that regards all options as permissible, then one goes with the recommendation of one's third-favourite theory; and so on. This version of MFT no longer has the appeal of simplicity. But it avoids the counterintuitive results mentioned so far.

The much deeper issue with any version of MFT is that it's going to run into what I'll call the *problem of theory-individuation*.[14] Consider the following case. Suppose that Sophie has credence in two different theories: a form of non-consequentialism and a form of hedonistic utilitarianism, and she's choosing between two options. A is the option of killing one person in order to save ten people. *B* is the option of refraining from doing so. So her decision situation is as follows:

[14] This problem was first discovered by Toby Ord, and told to me in conversation.

	Non-consequentialism – 40%	Hedonistic Utilitarianism – 60%
A	impermissible	permissible
B	permissible	impermissible

According to any version of MFT, A is the appropriate option. However, suppose that Sophie then learns of a subtle distinction between different forms of hedonistic utilitarianism. So she realizes that the hedonistic theory she had credence in was actually an umbrella for two slightly different forms of hedonistic utilitarianism. So her decision situation instead looks as follows:

	Non-Consequentialism – 40%	Hedonistic Utilitarianism-1 – 30%	Hedonistic Utilitarianism-2 – 30%
A	impermissible	permissible	permissible
B	permissible	impermissible	impermissible

In this new decision situation, according to MFT, B is the appropriate option. So MFT is sensitive to how exactly we choose to individuate moral theories. But that seems crazily arbitrary. If A is the appropriate option for Sophie in the first decision situation, it should still be the appropriate option in the second decision situation.

Gustafsson and Torpman respond to this. They give an account of how to individuate moral theories, as follows:

> Regard moral theories T_i and T_j as versions of the same moral theory if and only if you are certain that you will never face a situation where T_i and T_j yield different prescriptions.[15]

[15] (J. Gustafsson and Torpman forthcoming, 14)

This avoids the arbitrariness problem, but in doing so means that their view faces an even bigger problem, which is that any real-life decision-maker will have vanishingly small credence in their favourite theory. Suppose that Tracy is deciding whether to allocate resources in such a way as to provide a larger total benefit, but with an inegalitarian distribution (option A), or in such a way as to provide a slightly smaller total benefit, but with an egalitarian distribution (option B). She has some credence in utilitarianism (U), but is almost certain in prioritarianism (P). However, she's not sure exactly how concave the prioritarian function should be. This uncertainty doesn't make any difference to the prioritarian recommendation in the case at hand; but in does make a small difference in some very rare cases. So her decision situation looks as follows:

	U – 2%	P_1 – 1%	P_2 – 1%	…	P_{98} – 1%
A	permissible	impermissible	impermissible	…	impermissible
B	impermissible	permissible	permissible	…	permissible

So on Gustaffson and Torpman's version of MFT, the appropriate option for Tracy is A, even though she's almost certain that it's wrong to choose A. This is clearly the wrong result.

IV. Against My Favourite Option

The solution to these last two problems with My Favourite Theory might seem obvious. Rather than focus on what *theory* the decision-maker has most credence in, we should

instead think about what *option* is most likely to be right, in a given decision situation. That is, we should endorse something like the following:

> *My Favourite Option (MFO)*: *A* is an appropriate option for *S* iff *S* thinks that *A* is the option, or one of the options, that is most likely to be permissible.[16]

MFO isn't sensitive to how we individuate theories. And it would get the right answer in the prioritarianism and utilitarianism case above. But it has serious problems of its own.

The first problem is that it is susceptible to money-pumps: that is, it can recommend that the decision-maker makes a series of decisions such that she is predictably left strictly worse off than she was when she started. Consider the following case. Ursula has equal credence in three consequentialist theories. These theories value the options A, A', B and C as follows (with higher numbers representing a more choice-worthy option):

	T_1 – 33%	T_2 – 33%	T_3 – 34%
A	4	2	3
A′	3	1	2
B	2	4	1
C	1	3	4

Suppose that Ursula has recently chosen option A. But now she is given the opportunity to change her mind. She is offered the choice between A and C. This decision-situation is as follows:

[16] (Lockhart 2000, 26) suggests this view, though ultimately rejects it.

	T_1 – 33%	T_2 – 33%	T_3 – 34%
A	4	2	3
C	1	3	4

In this decision-situation, A is permissible according to T_1 and impermissible according to T_2 and T_3; C is permissible according to both T_2 and T_3 and impermissible according to T_1. C is the option with the highest probability of permissibility. So C is the most appropriate option, according to MFO. So Ursula, following MFO, chooses C.

Ursula is then offered the choice between C and B. This second decision-situation looks as follows:

	T_1 – 33%	T_2 – 33%	T_3 – 34%
B	2	4	1
C	1	3	4

In this decision-situation, C is permissible according to T_3, but impermissible according to T_1 and T_2. B is permissible according to T_1 and T_2, but impermissible according to T_3. B is the option with the highest probability of permissibility. So B is the most appropriate option, according to MFO. So Ursula, following MFO, chooses B.

Next, however, let us suppose that Ursula is offered the choice between B and A'. This third decision-situation looks as follows:

	T_1 – 33%	T_2 – 33%	T_3 – 34%
A′	3	1	2
B	2	4	1

In this decision-situation, B is permissible according to T_2, but impermissible according to T_1 and T_3. C is permissible according to T_1 and T_3, but impermissible according to T_2. A' is the option with the highest probability of permissibility. So A' is the most appropriate option, according to MFO. So Ursula, following MFO, chooses A'.

But A' is an option that Ursula is certain is worse than A. By following MFO, Ursula has ended up strictly worse off than when she started. Moreover, the process could be iterated. Ursula could start off with an extremely valuable option A, but then be led to choose worse and worse options, until eventually she ends up with an option that is very bad indeed. This is known as a money-pump (or 'value pump'). And it looks like a problem for MFO.

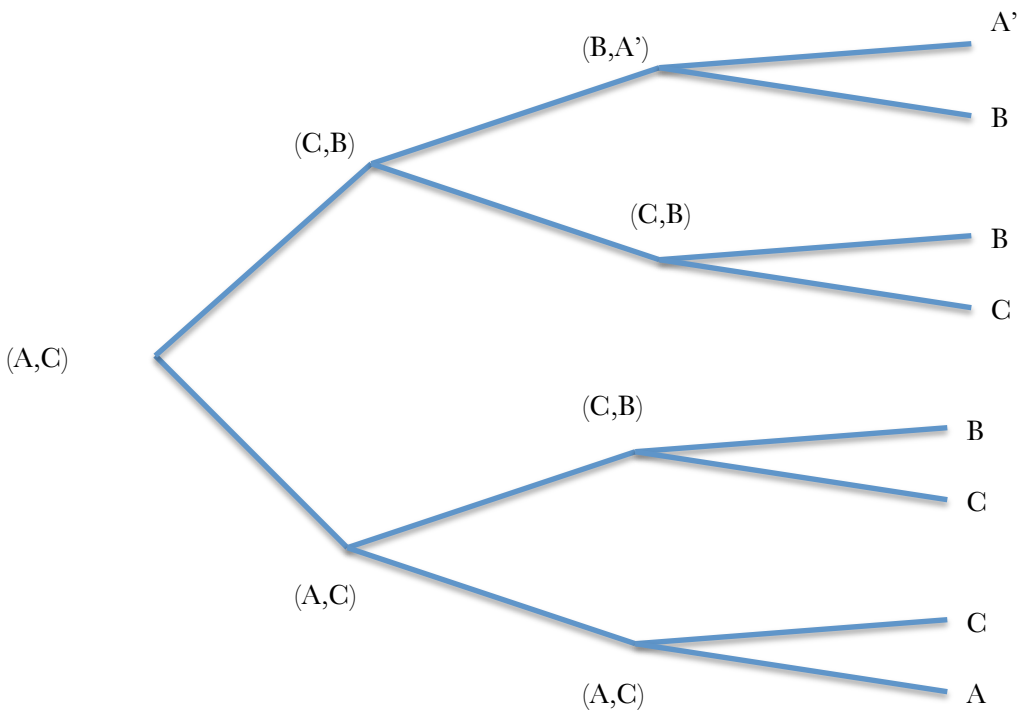
There is a standard response to the money-pump argument, as follows.[17] Either Ursula knows what trades are going to happen or she doesn't. If she doesn't, then the objection doesn't have much force: it's unfortunate that she doesn't know what's going to happen in the future; but it's that lack of knowledge that makes her lose money, and there is no reason to doubt Ursula's rationality just because a lack of knowledge makes her lose money. If in contrast she does know what trades are going to happen, then she can foresee what her choices will be at later stages, and use backwards induction to work out what that means her choices should be at the first stage. In the case above, if Ursula knows what trades are going to be made to her, she can *see* that, at Stage 3, she will

[17] E.g. (Schick 1986; Mongin 2000).

choose A' (because the choice is between B and A'). Given that she knows that she will trade at Stage 3, then she should refuse to trade at Stage 2 (because her choice, at Stage 2, is effectively between C and A'). But she still prefers C to A (which is her choice at Stage 1). So she will trade at the first step, and then stop. She won't end up strictly worse off than when she started.

That's the standard counterargument to money pumps. But it doesn't work in more complex cases.[18] Suppose, now that the money-pumper is persistent. Even if Ursula refuses to trade, the money-pumper then will offer the same trade again, up to three times. If at any point Ursula accepts a trade, then the money-pumper will offer a subsequent trade, in the manner of the last example, again up to three times. So the decision tree can be represented as follows:

[18] (Rabinowicz 2000).



The trades proceed in three stages. Let's consider what happens when Ursula uses foresight and backwards induction. At the third stage, fourth node down, Ursula can foresee she would choose C. At the third stage, third node down, Ursula can foresee she would choose B. This means that Ursula can foresee that at the second stage, second node down, she would choose C (and then subsequently choose B).

At the third stage, second node down, Ursula can foresee she would choose B. At the third stage, first node down, Ursula can foresee she would choose A'. This means that she can foresee that at the second stage, first node down, the agent would choose B (and then subsequently choose A').

This means at the first stage, Ursula's choice is effectively between B (if she takes the lower path) and A' (if she takes the upper path). She prefers A' to B, so she would

choose the upper path, trading A for C, C for B, and then C for A'. But that means that, *even though* Ursula can see what trades will happen in the future, she still ends up in a situation that is strictly worse than the situation she was in before.

This is problematic for MFO, though perhaps not a fatal problem. The deeper problem is given in the next section.

V. In favour of comparativism

Let's define *comparativism* as the view that what it's appropriate to do is a function *both* of the credences that the decision-maker assigns to different moral theories and of the degree of choice-worthiness that those theories assign to different options. MFT and MFO are both non-comparativist: they are sensitive only to the credences that the decision-maker has across moral theories.

But we can construct clear counterexamples to non-comparativism. Consider, again, the second variant of *Susan and the Medicine*:

> *Susan and the Medicine - II*
>
> Susan is a doctor, who faces three sick individuals, Greg, Harold and Harry. Greg is a human patient, whereas Harold and Harry are chimpanzees. They all suffer from the same condition. She has a vial of a drug, D. If she administers all of drug D to Greg, he will be completely cured, and if she administers all of drug to the chimpanzees, they will both be completely cured (health 100%). If she splits the drug between the three, then Greg will be almost completely cured

(health 99%), and Harold and Harry will be partially cured (health 50%). She is unsure about the value of the welfare of non-human animals: she thinks it is equally likely that chimpanzees' welfare has no moral value and that chimpanzees' welfare has the same moral value as human welfare. And, let us suppose, there is no way that she can improve her epistemic state with respect to the relative value of humans and chimpanzees.

Using numbers to represent how good each outcome is: Sophie is certain that completely curing Greg is of value 100 and that partially curing Greg is of value 99. If chimpanzee welfare is of moral value, then curing one of the chimpanzees is of value 100, and partially curing one of the chimpanzees is of value 50.

Her three options are as follows:

A: Give all of the drug to Greg

B: Split the drug

C: Give all of the drug to Harold and Harry

Her decision can be represented in the following table, using numbers to represent how good each outcome would be.

	Chimpanzee welfare is of no moral value – 50%	Chimpanzee welfare is of significant moral value – 50%
A	100	0
B	99	199
C	0	200

> Finally, suppose that, according to the true moral theory, chimpanzee welfare is of the same moral value as human welfare and that therefore, she should give all of the drug to Harold and Harry. What should she do?

According to both MFT and MFO, in this case, both A and C are appropriate options, but B is inappropriate. But that seems wrong. B seems like the appropriate option, because, in choosing either A or C, Susan is risking grave wrongdoing. B seems like the best hedge between the two theories in which she has credence. But if so, then any metanormative theory according to which what it's appropriate to do is always what it's maximally choice-worthy to do according to some theory in which one has credence (including MFT, MFO, and variants thereof) is false.

Moreover, this case shows that one understanding of the central metanormative question that has been given in the literature is wrong. Jacob Ross seems to think that the central metanormative question is "what ethical theories are worthy of acceptance and what ethical theories should be rejected," where Ross defines acceptance as follows:[19]

> to accept a theory is to aim to choose whatever option this theory would recommend, or in other words, to aim to choose the option that one would regard as best on the assumption that this theory is true. For example, to accept

[19] (Ross 2006, 743).

> utilitarianism is to aim to act in such a way as to produce as much total welfare as possible, to accept Kantianism is to aim to act only on maxims that one could will as universal laws, and to accept the Mosaic Code is to aim to perform only actions that conform to its Ten Commandments.

The above case shows that this cannot be the right way of thinking about things. Option B is wrong according to all theories in which Susan has credence: she is certain that it's wrong. The central metanormative question is therefore not about which first-order normative theory to accept: indeed, in cases like Susan's there's *no* moral theory that she should accept. Instead, it's about which *option* it's appropriate to choose.[20]

VI. In favour of treating moral and empirical uncertainty analogously

In the previous section we saw an argument in favour of comparativism: that what it is appropriate to do is some function of both the decision-maker's credences over theories, and of the degrees of choice-worthiness that those theories assign to options. But which form of comparativism is correct? In this section I argue that, when choice-worthiness differences are comparable across theories, we should handle moral uncertainty in just the same way that we should handle empirical uncertainty. Expected utility theory is the standard account of how to handle empirical uncertainty. So MEC should be the standard account of how to handle moral uncertainty. (I therefore am perfectly open to those who wish to follow Lara Buchak[21] and endorse a form of risk-weighted expected utility theory. My primary claim is that they one should endorse maximizing risk-weighted choice-worthiness iff risk-weighted expected utility theory is the correct way to

[20] One could say that, in Susan's case, she should accept a theory that represents a hedge between the two theories in which she has credence. But why should she accept a theory that she knows to be false? This seems to be an unintuitive way of describing the situation, for no additional benefit.
[21] (Buchak 2013).

accommodate empirical uncertainty. I don't wish to enter into this debate, so I assume that the risk-neutral version of expected utility theory is the correct formal framework for accommodating empirical uncertainty.)

The argument for treating empirical and moral uncertainty analogously begins by considering that there are very many ways of distinguishing between proposition-types: we can divide propositions into the a priori and a posteriori, the necessary and contingent, or those that pertain to biology and those that do not. These could all feature into uncertainty over states of nature. Yet, intuitively, in all these cases the nature of the propositions over which one is uncertain does not affect which decision-theory we should use. So it would seem arbitrary to think that *only* in the case of normative propositions does the nature of the propositions believed affect which decision-theory is relevant. So it seems that the default view is the moral and empirical uncertainty should be treated in the same way.

Moreover, the analogy between decision-making under empirical uncertainty and decision-making under moral uncertainty becomes considerably stronger when we consider that the decision-maker might not even *know* the nature of her uncertainty.

Suppose, for example, that Sophie is deciding whether to each chicken. She's certain that she ought not to eat an animal if that animal is a person, but she is uncertain about whether chickens are persons or not. And suppose that she has no idea whether her uncertainty stems from empirical uncertainty, about chickens' capacity for certain experiences, or from moral uncertainty, about what the sorts of attributes qualify one as a person in the morally relevant sense.

It doesn't seem plausible to suppose that the nature of her uncertainty could make a difference as to what she should decide. It seems even less plausible to think that it could be extremely important for Sophie to find out the nature of her uncertainty before making her decision. But if we think that moral and empirical uncertainty should be treated in different ways, then this is what we're committed to. If her uncertainty stems from empirical uncertainty, then that uncertainty should be taken into account, and everyone would agree that she ought not (in the subjective sense of 'ought') to eat the chicken. If her uncertainty stems from moral uncertainty and moral and empirical uncertainty should be treated differently, then it might be that she should eat the chicken. But then, because finding out the nature of her uncertainty could potentially completely change her decision, she should potentially invest significant resources into finding out what the nature of her uncertainty is. But it seems bizarre that she should do this.

So, as well as pointing out the problems with alternative views, as I did in section I-V, there seems to be a strong direct argument for the view that moral and empirical uncertainty should be treated in the same way. Under empirical uncertainty, expected utility theory is the standard formal framework. So I shall take that as the default correct formal framework under moral uncertainty as well, and that maximising expected choice-worthiness is correct.

This concludes my argument in favour of maximizing expected choice-worthiness. In the rest of this chapter I'll consider and respond to three objections to this view.

VII. Objections and responses

i. Disanalogy with Prudential Reasoning

One line of objection to MEC comes from Brian Weatherson[22] (though he uses the argument against metanormativism in general rather than MEC in particular). The objection is that MEC doesn't seem plausible in cases where there are only prudential reasons at stake. Weatherson gives the following case:

> *Bob and the Art Gallery*
>
> Bob has to decide whether to spend some time at an art gallery on his way home. He knows the art there will be beautiful, and he knows it will leave him cold. There isn't any cost to going, but there isn't anything else he'll gain by going either. He thinks it's unlikely that there's any prudential value in appreciating beauty, but he's not certain. As it happens, it really is true that there's no prudential value in appreciating beauty. What should Bob do?

Weatherson thinks that Bob makes no mistake in walking home. But, as is stipulated in the case, there's some chance that Bob will benefit, prudentially, from going to the art gallery, and there's no downside. This example, so the objection goes, shows that Dominance is false, and therefore shows that MEC is false.

I think, however, that the example is poorly chosen. Weatherson stipulates in the case that there's no cost to spending time in the art gallery. But it's difficult to imagine that to be the case: the time that Bob would spend in the art gallery, having an experience

[22] (Weatherson 2014, 148–149)

that 'leaves him cold', could presumably been used to do something else more enjoyable instead.[23] In which case, depending on how exactly the example was specified, MEC would recommend that Bob goes home rather than to the art gallery. In order to correct for this, we could modify the case, and suppose that Bob has the choice of two routes home, A and B. Both will take him exactly the same length of time. But route B passes by great works of architecture than Bob hasn't seen before, whereas route A does not. In this case, where there is some probability that viewing great architecture has prudential value, MEC really would say it's appropriate to choose route B and not appropriate to choose route A. But that seems like the correct answer.

Other cases also suggest that MEC gets the right answer in purely prudential cases. Consider the following case.

> *Charlie and the Experience Machine*
>
> Charlie is a firm believer in hedonism, but he's not certain, and gives some significant credence to the objective list theory of wellbeing. He is offered the chance to plug in to the Experience Machine. If he plugs in, his experienced life will be much the same as it would have been anyway, but just a little bit more pleasant. However, he would be living in a fiction, and so wouldn't achieve the objective goods of achievement and friendship. As it happens, hedonism is true. Is there any sense in which Charlie should not plug in?

In this case, it seems clear that there's a sense in which Charlie should not plug in. Given his uncertainty, it would be too risky for him to plug in. That is: it would be

[23] I thank Amanda MacAskill for this point, and for the following example.

appropriate for him to refrain from plugging in, even if hedonism were true, and even if he was fairly confident, but not sure, that hedonism were true.

Given empirical certainty, we generally have a pretty good grasp of what outcomes are good for us or bad for us; there is far less disagreement, in terms of their recommendations, between the leading theories of wellbeing than there is between the leading theories of morality. That makes it harder to come up with cases where there is a clear distinction between what it's appropriate for someone to do, and what someone prudentially ought to do (in the sense of 'ought' that's relevant to first-order prudential theories). But, other than that factor, I don't see any difference in the force of the arguments in favour of MEC whether we're considering moral reasons or prudential reasons.

ii. Demandingness?

A second objection is that MEC is too *demanding*. It has implications that require too great a personal sacrifice from us.[24] For example, Peter Singer has, for a long time, been arguing that members of affluent countries are obligated to give a large proportion of their income to those living in extreme poverty, and that failing to do so is as wrong, morally, as walking past a drowning child whose life one easily could save.[25] Many people who have heard the argument don't believe it to be sound; but even those who reject the argument should have at least some credence that its conclusion is true. And everyone agrees that it's at least permissible to donate the money. So isn't there a

[24] Weatherson hints at this objection in (2002); it is made at length in (Barry and Tomlin ms).
[25] (Singer 1972).

dominance argument for giving to fight extreme poverty? That is, doesn't the decision situation look like this:

	Singer's conclusion is correct	Singer's conclusion is incorrect
Give	permissible	permissible
Don't Give	impermissible	permissible

And, if so, then it is appropriate for us, as citizens of affluent countries, to give a large proportion of our income to fight poverty in the developing world. But, so the objection goes, that is too much to demand of us. So Dominance, and therefore MEC, should be rejected.

My first response to this objection is that it is guilty of double counting. Considerations relating to demandingness *are* relevant to consideration of what it's appropriate to do under moral uncertainty. But they are relevant because they are relevant to what credences one ought to have across different moral theories. If they were *also* relevant to which metanormative theory is true, then one has given demandingness considerations twice the weight that they should have. As an analogy: it would clearly be incorrect to argue against MEC because, in some cases, it claims that it is appropriate for one to refrain from eating meat, even though (so the objection goes) there's nothing wrong with eating meat. That would be double-counting the arguments against the view that it is impermissible to eat meat. In general it seems illegitimate to move from claims about first-order normative matters to conclusions about which metanormative theory is true. One might respond that metanormative norms are norms about rationality, and the

claim that MEC is too rationally demanding is a separate objection. But it seems odd to me to claim that a theory of rationality, if that's what MEC is, can be too demanding.

However, I do think that it's reasonable to be suspicious of the above 'Dominance' argument for giving a large proportion of one's income to fight global poverty. As I stated in the introduction, a metanormative theory should take into account uncertainty about what the *all-things-considered* choice-worthiness ordering is. And the decision-maker who rejects Singer's argument should have some reasonable credence in the view that, all things considered, she ought to spend the money on herself (or on her friends and family). This would be true on the view according to which there is no moral reason to give, whereas there is a prudential reason to spend the money on herself (and on her friends). So the decision-situation for a typical decision-maker might look as follows:

	Singer's argument is correct	Singer's argument is mistaken + prudential reasons to benefit oneself	Singer's argument is mistaken + no prudential reasons to benefit oneself
Give	permissible	slightly wrong	Permissible
Don't Give	gravely wrong	permissible	Permissible

Given this, what it's wrong to do depends on exactly how likely the decision-maker finds Singer's view. It costs approximately $3400 to save the life of a child living in extreme poverty,[26] and it would clearly be wrong, on the common sense view, for someone living in an affluent country not to save a drowning child even if it were at a personal cost of $3400. It seems to me that this intuition still holds even if it cost $3400

[26] This is the current best-guess estimate from www.GiveWell.com, a leading charity evaluator (though the estimate varies over time).

to prevent a one in ten chance of a child drowning. In which case, the difference in choice-worthiness between giving and not-giving, given that Singer's conclusion is true, is at least 10x as great as the difference in choice-worthiness between giving and not-giving, given that Singer's conclusion is false. So if one has at least 10% credence in Singer's view, then it would be inappropriate not to give. However, the intuition becomes much more shaky if the $3400 only gave the drowning child an additional one in a hundred chance of living. So perhaps the difference in choice-worthiness between giving and not-giving, on the assumption that Singer's conclusion is true, is less than 100x as great as the difference in choice-worthiness between not-giving and giving, on the assumption that Singer's conclusion is false. In which case, it would be appropriate to spend the money on oneself if one has less than 1% credence that Singer's conclusion is true.

The above argument was very rough. But it shows, at least, that there is no two-line knockdown argument from moral uncertainty to the appropriateness of giving. Making that argument requires doing first-order moral philosophy, in order to determine how great a credence one should assign to the conclusion of Singer's view. And that, I think, should make us a lot less suspicious of MEC. The two-line argument seemed too easy to be sound.[27] And that's true. But the error was not with MEC itself: the error was that MEC was being applied in too simple-minded a way.

[27] Weatherson says: "The principle has some rather striking consequences, so striking we might fear for its refutation by a quick modus tollens" (2002, 694) and "I'm arguing against philosophers who, like Pascal, think they can convince us to act as if they are right as soon as we agree there is a non-zero chance that they are right. I'm as a rule deeply sceptical of any such move, whether it be in ethics, theology, or anywhere else" (2014, 145).

iii. Fanaticism

The third objection is that MEC will result in *fanaticism*: that is, the expected choice-worthiness will be dominated by theories according to which moral situations are incredibly high stakes.[28] Consider the following case:

> *Doug's Lie*
>
> Doug in unsure between two moral theories: utilitarianism, and an absolutist form of non-consequentialism. Doug has the option to tell a lie, and in doing so mildly harm another person, in order to save the lives of ten people. For utilitarianism, the difference in choice-worthiness between saving ten people and saving none, all other things being equal, is 10. The difference in choice-worthiness between doing nothing and telling a lie, all other things being equal is 0.1. Absolutism agrees that it is choice-worthy to save lives, and that it's more choice-worthy to save more lives. However, according to the absolutist, telling a lie is absolutely wrong, such that it is never permissible to tell a lie, no matter how great the consequences. Doug is almost certain that utilitarianism is correct, but has a very small credence that the absolutist view is true.

In the above case, it seems very obvious, intuitively, that it's appropriate for Doug to lie: he's almost certain both that it's the right thing to do, and that it's extremely important that he tells the lie. But, so the objection goes, this is not what MEC would recommend.

According to this objection, the most natural way to represent the absolutist theory decision-theoretically is to say that telling a lie is infinitely wrong according to

[28] This problem is first raised by (Ross 2006, 765).

absolutism. If so, then, no matter how small Doug's credence is in absolutism, then the expected choice-worthiness of telling a lie is less than that of refraining from telling a lie. That is, the decision-situation looks as follows:

	Utilitarianism	Absolutism
Lie	+9.9	$-\infty$
Don't lie	0	0

The fanaticism problem arises in a particularly strong form for theories that seem to posit infinite degrees of choice-worthiness. But the same problem arises even for huge but finite amounts of choice-worthiness. For example, we should certainly give non-zero credence to insects having moral value. It seems plausible that we should at least give a one in ten thousand likelihood to the idea that insects have one-millionth the moral value of humans. But if we do this then, given the 10^19 insects in the world, saving insects from early death might well be, in terms of expected choice-worthiness, a more important moral goal than saving humans from early death. And perhaps it would be overconfident to assign less than one in a billion likelihood to bacteria having one billionth of the value of humans. If so, then, given the 10^31 bacteria in the world, preventing the early death of bacteria might be, in expected choice-worthiness terms, the most important moral goal there is. But these conclusions are ridiculous.

Jacob Ross considers this problem, and offers two responses. His first response is to bite the bullet, that is: "to endorse the Pascalian conclusion, however counterintuitive it may seem at first."[29] His second response is to suggest that one should not have a non-

[29] (Ross 2006, 766)

infinitesimal credence in fanatical theories: "If, therefore, one is subject to rational criticism in this case, it is not in choosing to accept [a fanatical theory] but rather in having a positive, noninfinitesimal degree of credence in a theory that is so fanatical that its contribution to the expected values of one's options swamps that of all other theories."

I cannot endorse either of these responses. Regarding the second, it is deeply implausible to claim that one should have zero credence or infinitesimal credence in any fanatical theories. I believe that absolutist theories are incorrect, but they are not so implausible as to warrant credence 0. On the standard understanding of credences,[30] to have credence 0 in a proposition is to be certain that one could never gain any evidence that would change one's view away from credence 0. But we can clearly imagine such evidence. For example, if everyone else in the world came to believe in absolutism after lengthy philosophical reflection, we would have reason to have positive credence in absolutism. Or if we discovered that there is a God, and His booming voice told us that absolutism is true, that would also provide evidence for absolutism. Nor, I think, does the idea of merely infinitesimal credence fare much better. First, doing so requires departing from standard Bayesianism, according to which a credence function to map onto the real numbers (which does not include infinitesimals).[31] But, secondly, even if we allow the possibility of rational infinitesimal credences, it seems overconfident, to say the least, to have such a low credence in absolutist views, despite the testimony of, for example, Kant and Anscombe (on at least some interpretations). And if it's true that

[30] Though see (Hájek 2003b) for arguments against the standard view.
[31] See (Easwaran 2014) for arguments against using hyperreals in our models of credences. See (Hájek 2003a) for discussion of how invoking infinitesimals fails to help with the 'fanaticism' problem within decision theory under empirical uncertainty.

even just some decision-makers should rationally have very small but non-infinitesimal credences in absolutist theories, then the fanaticism problem still looms large.

Regarding Ross's first response, the fanaticism problem does not merely generate grossly counterintuitive result in cases like *Doug's Lie*. Rather, it simply *breaks* MEC. In any real-life variant of *Doug's Lie*, Doug should have some non-zero credence in a view according to which it's absolutely wrong not to save those lives. In which case, the expected choice-worthiness of lying is also negative infinity. And this will be true for any decision a real-life decision maker faces. For any option, there will always be some theory in which the decision-maker has non-zero credence according to which that option is infinitely wrong.[32]

A better response is simply to note that this problem arises under empirical uncertainty as well as under moral uncertainty. One should not give 0 credence to the idea that an infinitely good heaven exists, which one can enter only if one goes to church; or that it will be possible in the future through science to produce infinitely good outcomes. This is a tricky issue within decision theory and in my view no wholly satisfactory solution exists.[33] But it is not a problem that is unique to moral uncertainty. And the primary claim in this chapter is that empirical and moral uncertainty should be treated in the same way. So whatever is the best solution to the fanaticism problem under empirical uncertainty is the best solution to the fanaticism problem under moral uncertainty. So we can put this issue to the side: my goal in this thesis is to explore the *distinctive*

[32] For further discussion of the problems that infinite amounts of value pose for decision-theory, see (Hájek 2003a).

[33] The standard response is to endorse prudential and moral theories whose choice-worthiness functions are bounded above and below. But this idea has severe problems of its own: making the choice-worthiness of decisions oddly dependent on facts about the past, and making bizarre recommendations when the decision-maker is close to the bound. (Beckstead ms) provides discussion.

problems that arise for decision-making under normative uncertainty. This is my primary response to the problem. However, there are, I think, two more moral uncertainty-specific things that one can say on this issue, so I briefly mention them before moving on. They both pertain to how to make comparisons of magnitudes of choice-worthiness across theories.

First, one could argue that, really, the absolutist theory is incomparable with the utilitarian theory. If so, then it does not fall under the remit of this chapter; instead, it should be dealt with along with merely ordinal theories, in the next chapter, or with cardinal but incomparable theories, in chapter 3. Either way, on the views I defend, if absolutism and utilitarianism are regarded as incomparable, then the absolutist theory will not be able to swamp other views.

Second, even if one does suppose that absolutism and utilitarianism are comparable, I think that the fanaticism problem is not as bad as it seems.[34] To see this, consider that the representation of absolutism as assigning infinite un-choice-worthiness to lying seems like a bad representation, and that the objector misled us in claiming that it was the most natural representation. For example, one wants to be able to make claims like: "if absolutism is true, then telling two lies is twice as wrong as telling one lie" or "if absolutism is true, then telling what is certainly is lie is twice as wrong as doing something that has a 50% chance of being telling a lie." We can't make sense of these claims if all lie-tellings are of infinite negative choice-worthiness. Instead, we can represent them lexicographically, using vectors, where they assign a two-dimensional choice-worthiness vector to options, such that one should maximize the first element

[34] With this idea coming out of conversation with Toby Ord.

but then, in case of a tie, maximize the second element. If so, then the decision-situation would look as follows:

	Utilitarianism	Absolutism
Lie	+9.9	(-0.1, 10)
Don't lie	0	(0, 0)

This more accurate way of representing absolutism then allows us to see clearly an implicit assumption that we made in *Doug's Lie*. That is, we implicitly made the intertheoretic comparison between absolutism and utilitarianism in one specific way: we assumed that the difference in choice-worthiness between saving one person and doing nothing was the same according to both theories. We assumed that both theories agree that saving lives is important, but that absolutism thinks that not-lying is vastly more important — in fact, lexically more important. If so, then the correct way to represent the two normalized theories is as follows:

	Utilitarianism	Absolutism
Lie	(0, +9.9)	(-0.1, 10)
Don't lie	(0, 0)	(0, 0)

If we represent the theories this way, the first element of the utilitarian theory is always 0. In order to accommodate this lexicographic ordering, the idea would be to endorse a two-step metanormative theory: maximize the expected choice-worthiness of the first element of each theory's choice-worthiness vector; then, in case of a tie, maximize the expected choice-worthiness of the second element of each theory's choice-worthiness

vector. If so, then, even using the lexicographic representation, it's appropriate for Doug to refrain from lying, no matter how small Doug's credence is in absolutism.

But it's by no means obvious that we should make the intertheoretic comparison between utilitarianism and absolutism in this way. We know that utilitarianism thinks that the difference between saving ten lives and doing nothing is 10 units, and of doing nothing and of lying is 0.1 units. And we know that not-lying lexicographically dominates lying, according to absolutism. Choosing to represent the two theories such that they agree on the choice-worthiness of saving lives is only one way to make the intertheoretic comparison. Instead, we could make the intertheoretic comparison such that they agree on the un-choice-worthiness of lying.[35] If so, then the correct way to represent the two theories is as follows:

	Utilitarianism	Absolutism
Lie	(9.9, 0)	(-0.1, 10)
Don't lie	(0, 0)	(0, 0)

That is, we normalize the two theories at the higher lexicographical order rather than the lower lexicographical order. And if *this* is the correct way to make the intertheoretic comparison, then Absolutism has almost no say at all over the decision and, unless Doug was very confident in Absolutism indeed, then it would be appropriate for Doug to lie.

[35] Similarly, in the insects and humans case, we assumed that the choice-worthiness of saving a human was equal according to both the 'insects have vanishingly small moral value' view and the 'insects have one millionth of the moral value of humans'. But we could have normalized such that the choice-worthiness of saving an insect is the same according to both views. If so, then the 'insects have one millionth the moral value of humans" view would not swamp the expected choice-worthiness calculations.

Admittedly, the first way of making the intertheoretic comparison seems more plausible, intuitively, to me. But I'm not certain that that's true. So a decision-maker like Doug, perhaps, should split his credence between the two different ways of making the intertheoretic comparison, giving higher credence to the one that seems more intuitively plausible.[36] In which case, Doug's credences over normalized moral theories might look as follows:

	Utilitarianism$_1$ – 98.5%	Utilitarianism$_2$ – 0.5%	Absolutism – 1%
Lie	(0, 9.9)	(9.9, 0)	(-0.1, 10)
Don't lie	(0, 0)	(0, 0)	(0, 0)

If these were Doug's credences, it would be appropriate for Doug to lie: which options are appropriate is almost entirely determined by Utilitarianism$_2$ and Absolutism and, though Absolutism has twice the credence, the situation is much higher stakes for Utilitarianism$_2$. So the appropriate option is to lie. Taking into account uncertainty about how to normalise across theories therefore seems to get reasonably intuitive conclusions concerning what it is appropriate for one to do in real-life cases even when one has credence in what is seems initially to be a 'fanatical' moral theory.

VIII. Conclusion

In this chapter I have argued that, in conditions of cardinal and intertheoretically comparable choice-worthiness, moral and empirical uncertainty should be treated in

[36] In chapter 4 I will give grounds for thinking that we should think of this as Doug splitting his credence between two different theories, one an 'amplified' version of the other, rather than splitting his credence between two different ways of normalizing the same theory. But nothing will turn on this for now.

the same way. Because I take expected utility theory to provide the default formal framework for taking empirical uncertainty into account, that means I think that maximise expected choice-worthiness is the default metanormative theory.

In the next chapter, I try to work out what it's appropriate to do in situations where the condition of cardinality is not met.

Chapter 2: Ordinal Theories and the Social Choice Analogy

Introduction

In the previous chapter, I argued that when the decision-maker has non-zero credence only in cardinal and intertheoretically comparable theories, it's appropriate to maximize expected choice-worthiness.

But when we look to apply MEC in generally, a couple of problems immediately rear their heads. First, what should you do if one of the theories you have credence in doesn't give sense to the idea of magnitudes of choice-worthiness? Some theories will tell you that murder is a more serious wrong than lying is, but will not give any way of determining of *how much* more serious a wrong murder is than lying. But if it doesn't make sense to talk about magnitudes of choice-worthiness, on a particular theory, then we won't be able to take an expectation over that theory. I'll call this *the problem of merely ordinal theories*.

A second problem is that, even when all theories under consideration give sense to the idea of magnitudes of choice-worthiness, we need to be able to compare these magnitudes of choice-worthiness across different theories. But it seems that we can't always do this. A rights-based theory claims that it would be wrong to kill one person in order to save fifty; utilitarianism claims that it would be wrong not to do so. But how can we compare the seriousness of the wrongs, according to these different theories? For

which theory is there more at stake? In line with the literature,[37] I'll call this *the problem of intertheoretic comparisons*.

Some philosophers have suggested that these problems are fatal to the project of developing a normative account of decision-making under moral uncertainty.[38] The primary purpose of this chapter and the next is to show that this is not the case.

I discuss these problems in more depth in section I. In section II, I introduce the analogy between decision-making under moral uncertainty and social choice, and explain how what I call the Hybrid Account can allow us to overcome these problems. The rest of the paper is spent fleshing out how this idea can help us to develop a metanormative theory that is applicable to merely ordinal theories. In section III, I show how the social choice analogy gives fertile ground for coming up with new metanormative theories. I reject the accounts in the literature and, in section IV, I defend the idea that, when maximizing choice-worthiness is not possible, one should use the Borda Rule instead. In section V, I respond to objections stemming from option-individuation and Arrow's impossibility theorem.

II. Intertheoretic Comparisons and Ordinal Theories

If you want to take an expectation over moral theories, two conditions need to hold. First, each moral theory in which you have credence needs to provide a concept of

[37] E.g. (Lockhart 2000; Ross 2006; Sepielli 2009).
[38] E.g. (Gracely 1996; Hudson 1989; Ross 2006) In conversation, John Broome has suggested that the problem is 'devastating' for accounts of decision-making under moral uncertainty; Derek Parfit has described the problem as 'fatal'.

choice-worthiness that is at least cardinally measurable. That is, you need to be able to make sense, on every theory in which you have credence, of the idea that the difference in choice-worthiness between *A* and *B* is equal to the difference in choice-worthiness between *C* and *D*. Intuitively, you need to be able to make sense of the idea of *how much* less choice-worthy one option is than another — that, though lying and murdering are both wrong, murdering is *far* more wrong than lying is. Second, you need to be able to make comparisons of magnitudes of choice-worthiness across different moral theories. That is, you need to be able to tell whether the difference in choice-worthiness between *A* and *B*, on T_i, is greater than, smaller than, or equal to, the difference in choice-worthiness between *C* and *D*, on T_j. Moreover, you need to be able to tell, at least roughly, *how much* greater the choice-worthiness difference between *A* and *B* on T_i is than the choice-worthiness difference between *C* and *D* on T_j.

The problem for maximizing expected choice-worthiness accounts is that sometimes these conditions don't hold. First, consider the problem of merely ordinal theories. Many theories do provide cardinally measurable choice-worthiness: in general, if a theory orders empirically uncertain prospects in terms of their choice-worthiness, such that the choice-worthiness relation satisfies the axioms of expected utility theory, then the theory provides cardinally measurable choice-worthiness.[39] Many theories satisfy this condition. Consider, for example, decision-theoretic utilitarianism, according to which one should maximise expected wellbeing (and which therefore satisfies the axioms of expected utility theory). If, according to decision-theoretic utilitarianism, a guarantee of saving Person A is equal to a 50% chance of saving no-one and a 50%

[39] As shown in (von Neumann et al. 2007). The application of this idea to moral theories is discussed at length in (Broome 1995).

chance of saving both Persons B and C, then we would know that, according to decision-theoretic utilitarianism, the difference in choice-worthiness between saving person B and C and saving person A is the same as the difference in choice-worthiness between saving person A and saving no-one. We give meaning to the idea of 'how much' more choice-worthy one option is than another by appealing to what the theory says in cases of uncertainty.

However, this method cannot be applied to all theories, for two reasons. First, if the theory does not order empirically uncertain prospects, then the axioms of expected utility theory are inapplicable. This problem arises even for some consequentialist theories: if the theory orders options by the value of the consequences the option *actually* produces, rather than the value of the consequences it is *expected* to produce, then the theory has not given enough structure such that we can use probabilities to measure choice-worthiness on an interval scale. For virtue-ethical theories, or theories that focus on the intention of the agent, this problem looms even larger. Second, the axioms of expected utility theory sometimes clash with common-sense intuition, such as in the Allais paradox.[40] If a theory is designed to cohere closely with common-sense intuition, as many non-consequentialist theories are, then it may violate these axioms. And if the theory does violate these axioms, then, again, we cannot use probabilities in order to make sense of cardinal choice-worthiness. Plausibly, Kant's ethical theory is an example of a merely ordinally measurable theory. According to Kant, murder is less choice-worthy than lying, which is less choice-worthy than failing to aid someone in need. But I don't think it makes sense to say, even roughly, that on Kant's view the difference in

[40] (Allais 1953).

choice-worthiness between murder and lying is greater than or less than the difference in choice-worthiness between lying and failing to aid someone in need. So someone who has non-zero credence in Kant's ethical theory simply can't use expected choice-worthiness maximization over all theories in which she had credence.

The second problem for the maximizing expected choice-worthiness account is the problem of intertheoretic comparisons. Even when theories do provide cardinally measurable choice-worthiness, there is no guarantee that we will be able to compare magnitudes of choice-worthiness between one theory and another. Previously I gave the example of comparing the difference in choice-worthiness between killing one person to save fifty and refraining from doing so according to a rights-based moral theory and according to utilitarianism. In this case, there's no intuitive answer to the question whether the situation is higher-stakes for the rights-based theory than it is for utilitarianism or vice-versa. And in the absence of intuitions about the case, it's difficult to see how there could be any way of determining an answer.

Even worse, the problem of intertheoretic comparisons arises even for theories that seem very similar. Consider two consequentialist theories: utilitarianism and prioritarianism. Prioritarianism gives more weight to gains in wellbeing to the worse off than it does to gains in wellbeing to the better off. But does it give more weight to gains in wellbeing to the worse off than utilitarianism does? That is, is prioritarianism like utilitarianism but with additional concern for the worse-off; or is prioritarianism like utilitarianism but with *less* concern for the better off? We could represent the prioritarian's idea of favouring the worse-off over the better-off equally well either way.

And there seems to be no information that could let us determine which of these two ideas is the 'correct' way to represent prioritarianism *vis-à-vis* utilitarianism.[41]

However, the question of what to do when one or more of these conditions do not hold has not been discussed in the literature. At best, it has been assumed that, in the absence of intertheoretic comparisons, the only alternative to maximizing expected choice-worthiness is the account according to which one should simply act in accordance with the theory that one thinks is most probable.[42] For that reason, it has been assumed that the lack of intertheoretic comparisons would have drastic consequences. For example, because intertheoretic incomparability entails that *maximise expected choice-worthiness* cannot be applied, Jacob Ross says:

> the denial of the possibility of intertheoretic value comparisons would imply that among most of our options there is no basis for rational choice. In other words, it would imply the near impotence of practical reason.[43]

In a similar vein, other commentators have regarded the problem of intertheoretic comparisons as fatal to the very idea of developing a normative account of decision-making under moral uncertainty. In the first modern article to discuss decision-making under moral uncertainty, James Hudson says:

> Hedging will be quite impossible for the ethically uncertain agent... Under the circumstances, the two units [of value, according to different theories] must be incomparable by the agent, and so there can be no way for her [moral]

[41] I give further arguments for the view that choice-worthiness differences are sometimes incomparable across theories in the next chapter.
[42] E.g. (Ross 2006, 762 fn.11).
[43] (Ross 2006, 763) Note that Ross uses this purported impotence as a *reductio* of the idea that different theories' choice-worthiness rankings can be incomparable. However, if my argument in the preceding paragraphs is sound, then Ross's position is not tenable.

> uncertainty to be taken into account in a reasonable decision procedure. Clearly this second-order hedging is impossible.[44]

Likewise, Edward Gracely argues, on the basis of intertheoretic incomparability, that:

> the proper approach to uncertainty about the rightness of ethical theories is to determine the one most likely to be right, and to act in accord with its dictates. Trying to weigh the importance attached by rival theories to a particular act is ultimately meaningless and fruitless.[45]

Similarly, Sepielli suggests that *maximise expected choice-worthiness* is 'the' correct principle for decision-making under moral uncertainty.[46] None of the above philosophers consider the idea that different criteria could apply depending on the informational situation of the agent. It is this assumption that leads to the thought that the problem of intertheoretic comparisons of value is fatal for accounts of decision-making under moral uncertainty. Against Ross and others, I'll argue that decision-making in conditions of moral uncertainty and intertheoretic incomparability is not at all hopeless. In this chapter, I focus on decision-making in conditions of merely ordinal theories. In the next chapter, I focus on decision-making in conditions of cardinal but incomparable theories. In both cases, I will exploit an analogy between decision-making under moral uncertainty and social choice. So let's turn to that now.

[44] (Hudson 1989, 224)
[45] (Gracely 1996, 331–2).
[46] (Sepielli 2009, 12).

III. Moral Uncertainty and the Social Choice Analogy

Social choice theory, in the framework developed by Amartya Sen,[47] studies how to aggregate individuals' utility functions (where each utility function is a numerical representation of that individuals' preferences) into a single 'social' utility function, which represents 'social' preferences. A *social welfare functional* is a function from sets of utility functions to a 'social' utility function. Familiar examples of social welfare functionals include: utilitarianism, according to which *x* has higher social utility than *y* iff the sum total of utility over all individuals is greater for *x* than for *y*; and maximin, according to which *x* has higher social utility than *y* iff *x* has more utility than *y* for the worst-off member of society.

Similarly, the theory of decision-making under moral uncertainty studies how to aggregate different theories' choice-worthiness functions into a single appropriateness function. The formal analogy between these two disciplines should be clear.[48] Rather than individuals we have theories; rather than preferences we have choice-worthiness orderings; and rather than a social welfare functional, we have a metanormative theory. And, just as social choice theorists try to work out what the correct social welfare functional is, so we are trying to work out what the correct metanormative theory is. Moreover, just as social choice theorists tend to be attracted to weighted utilitarianism ('weighted' because the weights assigned to each individual's welfare need not be equal)

[47] (A. K. Sen 1970).

[48] Note that this analogy is importantly different from other analogies between decision theory and social choice theory that have recently been drawn in the literature. Rachael Briggs's (2010) analogy is quite different from mine: in her analogy, a decision theory is like a voting theory but where the voters are the decision-maker's future selves. Samir Okasha's (2011) analogy is formally similar to mine, but his analogy is between the problem of social choice and the problem of aggregating different values within a pluralist epistemological theory, rather than with normative uncertainty.

when information permits,[49] so decision-theorists are attracted to its analogue under normative uncertainty, *maximize expected choice-worthiness,* when information permits.

The formal structure of the two problems is very similar. But the two problems are similar on a more intuitive level as well. The problem of social choice is to find the best compromise in a situation where there are many people with competing preferences. The problem of normative uncertainty is to find the best compromise in a situation where there are many possible normative theories with competing recommendations about what to do.

What's particularly enticing about this analogy is that the literature on social choice theory is well developed, and results from social choice theory might be transferable to moral uncertainty, shedding light on that issue. In particular, since the publication of Amartya Sen's *Collective Choice and Social Welfare*,[50] social choice theory has studied how different social welfare functionals may be axiomatised under different *informational assumptions*. One can vary informational assumptions in one of two ways. First, one can vary the *measurability assumptions*, and, for example, assume that utility is merely ordinally measurable, or assume that it is cardinally measurable. Second, one can vary the *comparability assumptions*: one can assume that we can compare differences in utility between options across different individuals; or one can assume that such comparisons are meaningless. The problem of determining how such comparisons are possible is known as the problem of interpersonal comparisons of utility. As should be clear from the discussion in the previous section, exactly the same distinctions can be made for

[49] See, for example, (Blackorby, Donaldson, and Weymark 1984) for the reasons why, given cardinal and interpersonally comparable utility, weighted utilitarianism is regarded as the most desirable social choice function.
[50] (A. K. Sen 1970).

moral theories: choice-worthiness can be ordinally or cardinally measurable; and it can be intertheoretically comparable or incomparable.

Very roughly, what is called voting theory is social choice theory in the context of preferences that are non-comparable and merely ordinally measurable. Similarly, the problem with which I'm concerned in this article is how to aggregate individual theories' choice-worthiness functions into a single appropriateness function in conditions where choice-worthiness is merely ordinally measurable.[51] So we should explore the idea that voting theory will give us the resources to work out how to take normative uncertainty into account when the decision-maker has non-zero credence only in merely ordinal theories.

However, before we begin we should note two important disanalogies between voting theory and decision-making under normative uncertainty. First, theories, unlike individuals, don't all count for the same: theories are objects of credences. The answer to this disanalogy is obvious. We treat each theory like an individual, but we weight each theory's choice-worthiness function in proportion with the credence the decision-maker has in that the theory. So the closer analogy is with weighted voting.

The second and more important disanalogy is that, unlike in social choice, a decision-maker under moral uncertainty will face varying information from different theories at one and the same time. For a typical decision-maker under normative uncertainty, some of the theories in which she has credence will be cardinally measurable and intertheoretically comparable; others will be cardinally measurable but

[51] And, as noted previously, I assume that comparisons of *levels* of choice-worthiness are not possible between theories.

intertheoretically incomparable; others again will be merely ordinally measurable. In contrast, when social choice theorists study different informational set-ups, they generally assume that the same informational assumptions apply to all individuals.

I discuss this issue at the end of chapter 3, providing a general metanormative theory where the precise method of aggregating the decision-maker's uncertainty is sensitive to the information provided by the theories in which she has credence, but which can be applied even in cases of varying informational conditions. In this chapter, however, I assume that all theories in which the decision-maker has credence are merely ordinal. With these caveats, the obvious next question is: which voting system should we use?

IV. Some Voting Systems

In the previous chapter, we looked at *My Favourite Theory* and *My Favourite Option*. One key argument against them was that they are insensitive to magnitudes of choice-worthiness. But if we are considering how to take normative uncertainty into account given non-zero credence in only merely ordinal theories, then this objection does not apply. So one might think MFT or MFO gets it right in conditions of merely ordinal theories. However, even in this situation, I think we have good reason to reject these accounts. Consider the following case:[52]

[52] In the cases that follow, and in general when I'm discussing merely ordinal theories, I will refer to a theory's choice-worthiness ordering directly, rather than its choice-worthiness function (e.g. I write $A \succ_i B \succ_i C$ rather than $CW_i(A)=3$, $CW_i(B)=2$, $CW_i(C)=1$). I do this in order to make it clear which theories are to be understood as ordinal, and which are to be understood as cardinal. However, strictly speaking the metanormative theories I discuss take choice-worthiness functions as inputs, rather than choice-worthiness orderings. Also, when it is clear which theory the choice-worthiness ordering belongs to, I leave out the subscript on the symbol '$\succ$'.

Judge

Julia is a judge who is about to pass a verdict on whether Smith is guilty for murder. She is very confident that Smith is innocent. There is a crowd outside, who are desperate to see Smith convicted. Julia has three options:

A: Pass a verdict of 'guilty'.

B: Call for a retrial.

C: Pass a verdict of 'innocent'.

Julia knows that the crowd will riot if Smith is found innocent, causing mayhem on the streets and the deaths of several people. If she calls for a retrial, she knows that he will be found innocent at a later date, and that it is much less likely that the crowd will riot at that later date. If she declares Smith guilty, the crowd will be appeased and go home peacefully. She has credence in three moral theories:

35% credence in a variant of utilitarianism, according to which A≻B≻C.

34% credence in a variant of common sense, according to which B≻C≻A.

31% credence in a deontological theory, according to which C≻B≻A.

MFT and MFO both regard A as most appropriate, because A both is most choice-worthy according to the theory in which the decision-maker has highest credence, and has the greatest probably of being right. But note that Julia thinks that there's only slightly less chance of B being right than A; and she's 100% certain that B is at least

second best. It seems highly plausible that this certainty in B being at least the second best option should outweigh the slightly lower probability of B being maximally choice-worthy. So it seems, intuitively, that B is the most appropriate option: it is well supported in general by the theories in which the decision-maker has credence. But neither MFT nor MFO can take account of that fact. Indeed, MFT and MFO are completely insensitive to how theories rank options that are not maximally choice-worthy. But to be insensitive in this way, it seems, is simply to ignore decision-relevant information. So we should reject these theories.

If we turn to the literature on voting theory, can we do better? Within voting theory, the gold standard voting systems are *Condorcet Extensions*.[53] The idea behind such voting systems is that we should think how candidates would perform in a round-robin head-to-head tournament. A voting system is a Condorcet extension if it satisfies the following condition: that, if, for every other option *B*, the majority of voters prefer *A* to *B*, then *A* is elected.

We can translate this idea into our terminology as follows. Let's say that *A beats B* (or *B is defeated by A*) iff it is true that, in a pairwise comparison between *A* and *B*, the decision-maker thinks it more likely that *A* is more choice-worthy than *B* than that *B* is more choice-worthy than *A*. *A* is the *Condorcet winner* iff *A* beats every other option within the option-set. A metanormative theory is a *Condorcet Extension* if it elects a Condorcet

[53] A brief comment on some voting systems I don't consider: I don't consider range voting because I'm considering the situation where theories give us only ordinal choice-worthiness, whereas range voting requires interval-scale choice-worthiness. I don't consider instant-runoff (or "alternative vote"), because it violates monotonicity: that is, one can cause A to win over B by choosing to vote for B over A rather than vice-versa. This is seen to be a devastating flaw within voting theory, and I agree: none of the voting systems I consider violate this property.

winner whenever one exists. Condorcet Extensions get the right answer in *Judge*, because B beats both A and C.

However, often Condorcet winners do not exist. Consider the following case:

Hiring Decision

Jason is a manager at a large sales company. He has to make a new hire, and he has three candidates to choose from. They each have very different attributes, and he's not sure what attributes are morally relevant to his decision. In terms of qualifications for the role, applicant B is best, then applicant C, then applicant A. However, he's not certain that that's the only relevant consideration. Applicant A is a single mother, with no other options for work. Applicant B is a recent university graduate with a strong CV from a privileged background. And applicant C is a young black male from a poor background, but with other work options. Jason has credence in three competing views.

30% credence in a form of virtue theory. On this view, hiring the single mother would be the compassionate thing to do, and hiring simply on the basis of positive discrimination would be disrespectful. So, according to this view, A>B>C.

30% credence in a form of non-consequentialism. On this view, Jason should just choose in accordance with qualification for the role. According to this view, B>C>A.

> 40% credence in a form of consequentialism. On this view, Jason should just choose so as to maximise societal benefit. According to this view, C≻A≻B.

In this case, no Condorcet winner exists: B beats C, C beats A, but A beats B. But, intuitively, C is more appropriate than A or B: A≻B≻C, B≻C≻A, and C≻A≻B are just 'rotated' versions of each other, with each option appearing in each position in the ranking exactly once. Given this, then the ranking with the highest credence should win out, and C should be the most appropriate option.

So Condorcet extensions need some way to determine a winner even when no Condorcet winner exists. Let us say that the *magnitude* of a defeat is the difference between the credence the decision-maker has that *A* is more choice-worthy than *B* and the credence the decision-maker has that *B* is more choice-worthy than *A*. A simple but popular Condorcet extension is the Simpson-Kramer method:

> *Simpson-Kramer Method*: *A* is more appropriate than *B* iff *A* has a smaller biggest pairwise defeat than *B*; *A* is equally as appropriate as *B* iff *A* and *B*'s biggest defeats are equal in magnitude.

In *Hiring Decision*, the magnitude of the biggest defeat for A and B is 70%, whereas the magnitude of the biggest defeat for C is only 60%. So, according to the Simpson-Kramer method, C is the most appropriate option, which seems intuitively correct in this case.

In what follows, I'll use the Simpson-Kramer Method as a prototypical Condorcet Extension.[54] Though Condorcet Extensions are the gold standard within voting theory, they are not right for our purposes. The reason is that, whereas voting systems rarely have to handle an electorate of variable size, metanormative theories do: varying the size of the electorate is analogous to changing one's credences in different normative theories. It's obvious that our credences in different normative theories should often change. But Condorcet Extensions handle that fact very poorly. A minimal condition of adequacy for handling variable electorates is:[55]

> *Twin Condition*: If an additional voter who has exactly the same preferences as a voter who is already part of the electorate joins the electorate and votes, that does not make the outcome of the vote worse by the lights of the additional voter.

The parallel condition in the case of decision-making under normative uncertainty is:

> *Updating Consistency*: Increasing one's credence in some theory one theory does not make the appropriateness ordering worse by the lights of that theory. More precisely: For all T_i, A, B, if $A \succ_i B$ and $A \succ_A B$, then if the decision-maker increases her credence in T_i, decreasing her credence in all other theories proportionally, it is still true that $A \succ_A B$.

[54] There are other Condorcet extensions that in my view are better than the Simpson-Kramer method, such as the Schulze method (Schulze 2010) and Tideman's Ranked Pairs (Tideman 1987), because they satisfy some other desirable properties that the Simpson-Kramer method fails to satisfy. However, these are considerably more complex than the Simpson-Kramer Method, and fail to be satisfactory for exactly the same reasons why the Simpson-Kramer method fails to be satisfactory. So in what follows I will use the Simpson-Kramer method as my example of a Condorcet extension.

[55] First given in (Moulin 1988).

Updating Consistency seems to me to be a necessary condition for any metanormative theory. When all theories in which the decision-maker has non-zero credence are merely ordinally measurable, appropriateness should be determined by two things only: first, how highly ranked the option is according to the theories in which the decision-maker has non-zero credence; and, second, how much credence the decision-maker has in those theories. It would be perverse, therefore, if, according to some metanormative theory, increasing one's credence in a particular theory on which A is more choice-worthy than B makes A less appropriate than B.

However, all Condorcet Extensions violate that condition. To see this, consider the following case:

> *Tactical Decisions*
>
> Jane is a military commander. She needs to take aid to a distant town, through enemy territory. She has four options available to her:
>
> *A*: Bomb and destroy an enemy hospital in order to distract the enemy troops in the area. This kills 10 enemy civilians. All 100 of her soldiers and all 100 enemy soldiers survive.
>
> *B*: Bomb and destroy an enemy ammunitions factory, restricting the scale of the inevitable skirmish. This kills 10 enemy engineers, who help enemy soldiers though they are not soldiers themselves. As a result, 90 of her soldiers and 90 enemy soldiers survive.

C: Status quo: don't make any pre-emptive attacks and go through the enemy territory only moderately well armed. 75 of her soldiers and 75 enemy soldiers survive.

D: Equip her soldiers with much more extensive weaponry and explosives. 95 of her soldiers and none of the enemy soldiers survive.

Jane has credence in five different moral views:

She has 5/16 credence in T_1 (utilitarianism), according to which one should simply minimise the number of deaths. According to T_1, A≻B≻C≻D.

She has 3/16 credence in T_2 (partialist consequentialism), according to which one should minimise the number of deaths of home soldiers and enemy civilians and engineers, but that deaths of enemy soldiers don't matter. According to T_2, D≻A≻B≻C.

She has 3/16 credence in T_3 (mild non-consequentialism), according to which one should minimize the number of deaths of home soldiers and enemy civilians and engineers, that deaths of enemy soldiers don't matter, and that it's mildly worse to kill someone as a means to an end than it is to let them die in battle. According to T_3, D≻A≻C≻B.

She has 4/16 credence in T_4 (moderate non-consequentialism), according to which one should minimise the number of deaths of all parties, but that there is a side-constraint against killing a civilian (but not an engineer or soldier) as a means to an end. According to T_4, B≻C≻D≻A.

She has 1/16 credence in T_5 (thoroughgoing non-consequentialism), according to which one should minimize the number of deaths, but that there is a side-constraint against killing enemy civilians or engineers as a means to and end, and that killing enemy civilians as a means to an end is much worse than killing enemy engineers. According to T_5, C≻D≻B≻A.

Given her credences, according to the Simpson-Kramer method D is the most appropriate option.[56] The above case is highly complicated, and I have no intuitions about what the most appropriate option is for Jane, so I don't question that answer. However, what's certain is that gaining new evidence in favour of one moral theory, and increasing one's credence in a moral theory, should not have the consequence of making an option which is *worse* by the lights of the theory in which one has increased one's credence *more appropriate*. But that's exactly what happens on the Simpson-Kramer method. Let us suppose that Jane hears new arguments, and increased her credence in T_5 so that now she has 5/19 credence in T_5. The ratios of her credences in all other theories stays the same: she has 4/19 in T_1, 3/19 in T_2, 3/19 in T_3 and 4/19 in T_4. After updating in favour of T_5, B is the most appropriate option, according to the Simpson-Kramer method.[57] But T_5 regards D as more choice-worthy than B. So the fact that Jane has updated in favour of T_5 has made the most appropriate option worse

[56] Option A is defeated by D by 11/16; B is defeated by A by 11/16; C is defeated by A by 11/16 and by B by 12/16; D is defeated by B by 9/16 and C by 10/16. D has the smallest biggest defeat. So D is the most appropriate option according to the Simpson-Kramer method.

[57] Option A is defeated by D by 14/19; B is defeated by A by 11/19 and D by 10/19; C is defeated by A by 11/16 and by C by 12/16; D is defeated by B by 13/16. B has the smallest biggest defeat. So B is the most appropriate option according to the Simpson-Kramer method. The Schulze method and Ranked Pairs (mentioned in fn.23) both give exactly the same answers in both versions of *Tactical Decisions*, so this case is a counterexample to them too.

by T_5's lights. That just shouldn't happen. So we should reject the Simpson-Kramer method.

In fact, it has been shown that *any* Condorcet extension will violate the *Twin Condition* described above;[58] and so any metanormative theory analogue will violate *Updating Consistency*. So, rather than just a reason to reject the Simpson-Kramer method, violation of *Updating Consistency* gives us a reason to reject all Condorcet extensions as metanormative theories.

Before moving on to a voting system that does better in the context of decision-making under moral uncertainty, I'll highlight one additional reason that is often advanced in favour of Condorcet extensions. This is that Condorcet extensions are particularly immune to strategic voting: that is, if a Condorcet extension voting system is used, there are not many situations in which a voter can lie about her preferences in order to bring about a more desirable outcome than if she had been honest about her preferences.

It should be clear that this consideration should bear no weight in the context of decision-making under moral uncertainty. We have no need to worry about theories "lying" about their choice-worthiness function (whatever that would mean). The decision-maker knows what moral theories she has credence in, and she knows their choice-worthiness functions. So, unlike in the case of voting, there is no gap between an individual's stated preferences and an individual's true preferences.

[58] The proof of this is too complex to provide here, but can be found in (Moulin 1988).

V. The Borda Rule

We have surveyed some metanormative theories that fail. Let's now look at a voting system that does better: the Borda Rule. To see both the Borda Rule's similarity to, and difference from, Condorcet extensions, again we should imagine that all options compete against each other in a round-robin head-to-head tournament. Like the Simpson-Kramer method, the magnitudes of the victories and defeats in these pairwise comparisons matter. However, rather than focusing on the size of the biggest pairwise defeat, as the Simpson-Kramer method does, the Borda Rule regards the success of an option as equal to the sum of the magnitudes of its pairwise victories against all other options. The most appropriate option is the option whose sum total of magnitudes of victories is greatest. To see the difference, imagine a round-robin tennis tournament, with players A-Z. A beats all other players, but in every case wins during a tiebreaker in the final set. B loses by only two points to A, but beats all other players in straight sets. Condorcet extensions care first and foremost about whether a player beats everyone else, and would regard A as the winner of the tournament. The Borda Rule cares about how many points a player wins in total, and would regard B as the winner of the tournament. In this case, it's not obvious to me which view is correct: the arguments for choosing A or B both have something going for them. But the fact that it's not obvious shows that we shouldn't reject outright any metanormative theory that isn't a Condorcet extension.

Now that we have an intuitive understanding of the Borda Rule, let's define it more precisely:

> An option *A*'s *Borda Score*, for any theory *Ti*, is equal to the number of options within the option-set that are less choice-worthy than *A* according to theory *Ti*'s choice-worthiness function, minus the number of options within the option-set that are more choice-worthy than *A* according to *Ti*'s choice-worthiness function.[59]
>
> An option *A*'s *Credence-Weighted Borda Score* is the sum, for all theories *Ti*, of the Borda Score of *A* according to theory *Ti* multiplied by the credence that the decision-maker has in theory *Ti*.

These definitions allow us to state the Borda Rule:

> *Borda Rule*. An option *A* is more appropriate than an option *B* iff *A* has a higher Credence-Weighted Borda Score than *B*; *A* is equally as appropriate as *B* iff *A* and *B* have an equal Credence-Weighted Borda Score.

In this way, the Borda Rule generates not just a set of maximally appropriate actions, but also an appropriateness function.

We can argue for the Borda Rule in two ways. First, we can appeal to cases. Consider again the *Judge* case. We criticised MFT and MFO for not being sensitive to the entirety of the decision-maker's credence distribution, and for not being sensitive to the entire

[59] The reader familiar with the Borda Rule might previously have seen an option's Borda Score defined as equal simply to the number of options below it. The addition of "minus the number of options that rank higher" clause is to account for tied options. The motivation for this way of dealing with ties is that we want the sum total of Borda Scores over all options to be the same for each theory, whether or not that theory claims there are tied options; if we did not do this we would be giving some moral theories greater voting power on arbitrary grounds. My definition gets this right. The reader may also have seen a Borda Score defined such that an option ranked *i*th receives *n-i* points plus 0.5 for every option with which it is tied, where *n* is the total number of options in the option-set. This definition is equivalent to mine; however, mine will prove easier to use when it comes to extending the account in the next section.

range of each theory's choice-worthiness ordering. The Borda Rule does not make the same error. In *Judge*, the Borda Rule ranks B as most appropriate, then C, then A.[60] This seemed to me to be the intuitively correct result: favouring an option that is generally well-supported rather than an option that is most choice-worthy according to one theory but least choice-worthy according to all others. In *Hiring Decision*, according to the Borda Rule C is the most appropriate candidate, then A then B.[61] Again, this seems to me to be obviously the correct answer.

Finally, consider the *Tactical Decisions* case. In this case, according to the Borda Rule, before updating the most appropriate option for Jane is A, followed by B, then D, then C.[62] As I said before, I don't have any intuitions in this case about which option is most appropriate. But I do know that Jane increasing her credence in T_5 (which ranks C≻D≻B≻A) shouldn't make the most appropriate option worse by T_5's lights. Indeed, given that it seems unclear which option is most appropriate, I would expect that a substantial increase in her credence in T_5 to improve the appropriateness ranking by T_5's lights. And that's what we find. After updating in favour of T_5, according to the Borda Rule the appropriateness ranking is D, followed by B, then C, then A.[63]

However, appeal to cases is limited in its value because we can't know whether the cases we have come up with are representative, or whether there exist other cases that are

[60] Because it doesn't affect the ranking when there are no ties, when giving working I will use a simpler definition of a Borda score: that an option's Borda Score, for some theory *i*, is equal to the number of options below it on theory *i's* choice-worthiness ranking. Given this, definition, *a* receives a score of 35*2 + 0 + 0 = 70; option *b* receives a score of 35 + 34*2 + 31 = 134; option *c* receives a score of 0 + 34 + 31*2 = 96.

[61] Option *a* receives a score of 30*2 + 0 + 10*1 = 100. Option b receives a score of 30*1 + 30*2 + 0 = 90. Option c receives a score of 0 + 30*1 + 40*2 = 110. So, on the Borda Rule, *c>a>b.*

[62] Option *a* receives a score of 5*3 + 3*2 + 3*2 +0 + 0 = 27. *b*'s score is 5*2 + 3*1 + 0 + 4*3 + 1*1 = 26. *c*'s score is 5*1 + 0 + 3*1 + 4*2 + 1*3 = 19. *d*'s score is 0 + 3*3 + 3*3 + 4*1 + 1*2 = 24.

[63] Option A receives a score of 5*3 + 3*2 + 3*2 +0 + 0 = 27. B's score is 5*2 + 3*1 + 0 + 4*3 + 4*1 = 29. C's score is 5*1 + 0 + 3*1 + 4*2 + 4*3 = 28. *d*'s score is 0 + 3*3 + 3*3 + 4*1 + 4*2 = 30.

highly damaging for our favoured proposal that we simply haven't thought of. A better method is to appeal to general desirable properties. One such property is *Updating Consistency*. In the context of voting theory, it has been shown that only *scoring rules* satisfy the equivalent property, where a scoring rule is a rule that gives a score to an option based on its position in an individuals' preference ranking, and claims you should maximize the sum of that score across individuals.[64] The Borda Rule is an example of a scoring rule, as is MFO, whereas MFT and the Simpson-Kramer Method are not. But we rejected MFO on the grounds that it wasn't sensitive to the entirety of theories' choice-worthiness rankings. So we could add in another condition that the score of each option in ith position has to be strictly greater than the score given to an option in $i+1$th position. This wouldn't quite motivate the Borda Rule, but it would come close.

In order to fully axiomatise the Borda Rule, we need another condition, as follows:

> *Cancellation*: If, for all pairs of options (A,B), S thinks it equally likely that $A \succ B$ as that $B \succ A$, then all options are equally appropriate.[65]

It has been shown that the only scoring function that satisfies the voting system analogue of *Cancellation* is the Borda Rule.[66] One might question *Cancellation* on the following grounds: one might think, for example, in a case where one has 50% credence in a theory according to which $A \succ B \succ C$ and 50% credence in a theory according to which $C \succ B \succ A$, that B is the most appropriate option (even though, according to *Cancellation*, all three options are equally appropriate). The grounds for this might be the

[64] See (Moulin 1988).
[65] The voting system analogue is: if for all pairs of alternative (x,y), the number of voters preferring x to y equals the number of voters preferring y to x, then a tie between all options should be declared. See (Young 1974).
[66] (Young 1974).

ordinal equivalent of risk-aversion, whereas the Borda Rule incorporates the equivalent of risk-neutrality. However, in chapter 1 I endorsed risk-neutral MEC as a default view. But if you should be risk-neutral when you can maximize expected choice-worthiness, then surely you should be risk neutral in the ordinal case as well. So for that reason I suggest that the Borda rule should be the default metanormative theory to aggregate merely ordinal theories.

With my argument in favour of the Borda Rule complete, let's consider objections that can be made to the Borda Rule.

VI. Objections and Extensions

Option-individuation

One objection to the Borda rule is that it is extremely sensitive to how one individuates options.[67] Consider the following case:

> *Trolley Problems*
>
> Sophie is watching as an out-of-control train hurtles towards five people working on the train track. If she flicks a switch, she will redirect the train, killing one person working on a different track. Alternatively, she could push a large man onto the track, killing him but stopping the train. Or she could do nothing. So she has three options available to her.

[67] This problem is analogous to the problem of 'clone-dependence' in voting theory, which itself is a generalization of the idea of vote-splitting. For discussion of clone-dependence, see (Tideman 1987). I thank Graham Oddie and an anonymous referee for pressing this criticism of the Borda Rule. The example is Oddie's.

A: Do nothing.

B: Flick the switch.

C: Push the large man.

She has credence in three moral theories.

40% in utilitarianism, according to which: B≻C≻A

30% in simple Kantianism, according to which: A≻B~C

30% in sophisticated Kantianism, according to which: A≻B≻C

In this case, according to the Borda Rule, B is the most appropriate option, followed by A and then C.[68] But now let us suppose that there are actually two Switching options:

A: Do nothing.

B': Flick the switch to the left.

B'': Flick the switch to the right.

C: Push the large man over the railing to stop the track

Sophie has the same credences in moral theories as before. Their recommendations are as follows:

Utilitarianism: B'~B''≻C≻A

[68] Now that some theories posit tied options, I return to using my 'official' definition of a Borda score in my working. Option A receives a score of 0 + 30*2 + 30*2 – (40*2 + 0 + 0) = 40. B's score is 40*2 + 0 + 30*1 – (0 + 30*1 + 30*1) = 50. C's score is 40*1 + 0 + 0 – (40*1 + 30*1 + 30*2) = –90.

Simple Kantianism: A≻B'∼B''∼C

Sophisticated Kantianism*:* A≻B'∼B''≻C

Given these choice-worthiness rankings, according to the Borda Rule A is the most appropriate option, then B' and B'' equally, then C.[69] So, according to the Borda Rule, it makes a crucial difference to Sophie whether she has just one way of flicking the switch or whether she has two: and if she has two ways of flicking the switch, it's of crucial importance to her to know whether that only counts as one option or not. But that seems bizarre.

Indeed, if this objection were successful, then it would be devastating, in my view. But there is a principled and satisfying response: what this objection shows is that we need to have a *measure* over possibility space, and that we were neglectful when we didn't initially include a measure in our definition of the Borda Rule.[70] A measure is a function from regions of some space to non-negative real numbers, with the following property: that, if A is a region that divides exactly into B and C, where B and C are non-overlapping, the measure of A is equal to the measure of B plus the measure of C. If a region D is empty, then it receives measure 0. When a measure is a *probability measure*, the measure of the domain of the function (that is, the whole space), is equal to 1. Though I talked about 'regions' in the preceding sentences, measures are often defined over sets, where the crucial property is that if set A is the union of the disjoint sets B and C, then the measure of A equals the sum of the measures of B and C.

[69] Option *a* receives a score of 0 + 30*3 + 30*3 – (40*3 + 0 + 0) = 60. *b'* and *b''* each receive a score of 40*2 + 0 + 30*1 – (0 + 30*1 + 30*1) = 50. *c*'s score is 40*1 + 0 + 0 – (40*2 + 30*1 + 30*3) = –160.
[70] I thank Owen Cotton-Barratt for this suggestion.

Intuitively, the measure of a region is supposed to represent the *size* of that region. Geographical area defined over regions of land is an example of a measure: if region A consists of region B and region C, and region B and C are non-overlapping, then the area of region A is equal to the area of region B plus the area of region C. It's clear that we have an intuitive notion of the 'size' of possibility space: that there's a clear sense in which the proposition *that Tim picks up a cup* carves out a larger portion of possibility space than the proposition *that Tim picks up a whisky-filled cup at 2am on the 25th December 2013*. We might instinctively want to explain this by saying that the latter proposition excludes a greater number of possible worlds than the former proposition. But when we are dealing with an infinite number of possible worlds that explanation is not available to us, and we have to bring in the notion of a measure.

In order to adequately define the Borda Rule, we need to have a measure defined over possibility space (that is, the set of all possible options) that represents the size of each option. Plausibly, the measure over the set of all possible options should be 1, so the measure is a probability measure. But I leave it open what exact choice of measure to use, for several reasons. First, it would simply take too long for this thesis — providing a convincing arguing for one measure over possibility space being the correct one would take volumes of argumentation. Second, having a theoretical grounding for the choice of measure is not that important for practical purposes. We have an intuitive grasp of the relative 'size' of an option, and the choice of measure is simply a way to make precise that intuitive grasp. So, even if we don't know the theoretical underpinning, we can still use that intuitive understanding in the context of a voting system.

However, to illustrate, some contenders for our choice of measure are as follows. One could use the Lebesgue measure over phase space: that is, roughly speaking, a uniform measure over all possible locations and velocities of all particles.[71] One could use the Kolmogorov complexity of a world: that is, the length of the shortest possible description of the world in some fixed language.[72] One could use the principle of maximum entropy, which is a generalisation of the principle of indifference.[73] Or, finally, the choice of measure might be subjective: equal to one's fundamental prior probability distribution over possible worlds (that is, the probabilities one assigns to different sets of worlds in the absence of *any* evidence at all). But, as I have said, in using the Borda Rule I would prefer to go with our intuitive grasp of the 'size' of an option, rather than privilege any particular theoretical account.

With the concept of a measure on board, we can reformulate the definition of an option's Borda score as follows: that an option's *Borda Score* is equal to the sum on the measure of the options below it minus the sum of the measure of the options above it. Once we've defined a Borda Score in this way, then we can use all the other definitions as stated. Nothing will change in terms of its recommendations in the cases we've previously discussed. But it resolves the option-individuation problem.

To see how this resolves the option-individuation problem, consider again the case given above. Let us suppose that the measure of each option, A, B and C, is 1/3.[74] If so, then, as before, according to the Borda Rule, B is the most appropriate option, followed

[71] See, for example, (Handfield 2012, chap. 5–6).
[72] See, for example, (Solomonoff 1964).
[73] See, for example, (Jaynes 1957).
[74] Note that there would be no difference to my argument if the measure were split unequally among options A, B and C.

by A and then C.[75] Now, however, when we split the option B into options B' and B'', we have to also split the measure: let us suppose that the measure splits equally, so that B' and B'' each have measure 1/6.[76] If so, then according to the Borda Rule, B is still the most appropriate option, followed by A and then C.[77] In general, the addition of a measure means that we can make sense of a 'size' of an option, and will therefore avoid the option-individuation problem.

There are two additional benefits of incorporating a measure into the Borda Rule. First, it makes the Borda Rule easier to use. Previously, we might have wondered how on earth we could determine how many options are available to us at any one time – and, for that reason, how on earth we were meant to apply the Borda Rule in any real-life situation. With the concept of measure, things become much easier: we can simply decide upon a formulation of the options such that they form a partition of the possibility space, giving each option a measure that intuitively represents the 'size' of that option. The exact way in which we decide to individuate options is no longer crucial.

Second, it means that the Borda Rule can handle situations in which the decision-maker faces an infinite number of options.[78] Before we had defined a measure over possibility space and incorporated that into an option's Borda score, one could have

[75] Option A receives a score of 0 + 30*(2/3) + 30*(2/3) – (40*(2/3) + 0 + 0) = 13 1/3. B's score is 40*(2/3) + 0 + 30*(1/3) – (0 + 30*(1/3) + 30*(1/3)) = 16 2/3. C's score is 40*(1/3) + 0 + 0 – (40*(1/3) + 30*(1/3) + 30*(2/3)) = –30.

[76] Note that there would be no difference to my argument if the measure did not divide evenly between B' and B''.

[77] Option A receives a score of 0 + 30*(2/3) + 30*(2/3) – (40*(2/3) + 0 + 0) = 13 1/3. B' and B'' each receive a score of 40*(2/3) + 0 + 30*(1/3) – (0 + 30*(1/3) + 30*(1/3)) = 16 2/3. C's score is 40*(1/3) + 0 + 0 – (40*(1/3) + 30*(1/3) + 30*(2/3)) = –30. That is, the scores are just the same as they were prior to the more fine-grained individuation of option *b*.

[78] I thank an anonymous referee for pressing this objection.

objected that the Borda Rule can't handle infinite option sets. For if the number of options below or above one option *A* were infinite, then there would have been no answer to the question of what that option's Borda score is.

Having a measure over possibility space resolves this problem, because one can have an infinite number of options with a measure that sums to some finite number. For example, suppose that below option *x* there are an infinite number of options, with measure ¼, 1/8, 1/16, 1/32… In this case, even though there are an infinite number of options there is a fact about the sum of the measure of options below *A*: namely, ½. Indeed, because the measure of the set of all possible options is 1, then the measure of options above or below any particular action will always be finite. So the Borda score of an option will always be well defined, even when there are an infinite number of options available to the decision-maker.

Arrow's Impossibility Theorem

Given the voting analogy, an obvious issue is how to respond to Arrow's impossibility result. In this final section I discuss the implications of Arrow's result for my account.

Arrow's impossibility theorem can be formulated in many ways. The strongest, in my view, is as follows: in conditions of ordinal preferences and interpersonal incomparability, there is no social welfare functional that satisfies *Pareto, Non-Dictatorship, Unlimited Domain* and *Contraction Consistency*.[79] The condition that the Borda Rule violates

[79] The reader might be more familiar with Arrow's final condition being Independence of Irrelevant Alternatives (IIA), which states that how candidate *x* fares against candidate *y* should be independent of

is Contraction Consistency. Within the context of decision-making under moral uncertainty, we may define this condition as follows:

> *Contraction Consistency*: Let $\mathcal{C}$ be the option-set, $\mathcal{M}$ be the set of maximally appropriate options given the option-set $\mathcal{C}$, and $\mathcal{S}$ be a subset of $\mathcal{C}$ that contains all the members of $\mathcal{M}$. The set $\mathcal{M}^*$ of the maximally appropriate options given the option-set $\mathcal{S}$ has all and only the same members as $\mathcal{M}$.

To see that the Borda Rule violates Contraction Consistency, consider *Hiring Decision* again. In that case, C was the uniquely most appropriate option. Now, however, suppose that it was no longer possible to hire candidate A. In which case the credence distribution looks as follows:

> 30% credence in virtue theory, according to which B≻C.
>
> 30% credence in non-consequentialism, according to which B≻C.
>
> 40% credence in consequentialism, according to which C≻B.

In this new decision-situation, B is now the uniquely most appropriate option. Similarly, if the option-set were (A, B), A would be most appropriate, and if the option-set were (A. C), C would be most appropriate. The appropriateness of options is highly sensitive to which other options are within the option-set. According to the objection, Contraction

---

the voters' views on *X* vis-à-vis any third candidate, *z*. However, though IIA is what Arrow uses in his proof, when he justifies the condition he gives an argument in favour of Contraction Consistency, apparently confusing the two conditions: see (Bordes and Tideman 1991) for discussion of this. Contraction consistency is a more plausible condition than IIA, in my view (IIA, for example, rules out the Borda Rule almost by fiat), and can be used in place of IIA to get the impossibility result, and so I use Contraction Consistency instead of IIA. A proof of Arrow's theorem using Contraction Consistency can be found in (Tideman 2006, 123–42).

Consistency is a requirement on any metanormative theory. So my account can't be correct.

Before responding to this objection, I'll note what it can't do. The objection can't use Contraction Consistency violation as an argument against the Borda Rule and in favour of some other voting system. Though Contraction Consistency is plausible, the other conditions listed are essential. All voting systems endorsed in the voting theory literature satisfy the other conditions but violate Contraction Consistency. Instead, therefore, the objection must be an objection to the Hybrid Account itself: to the very idea of using some voting system as a way to aggregate the part of one's credence devoted to merely ordinal theories.

Even more importantly, the violation of Contraction Consistency is not as bad as one might think, for two reasons.

First, a reason why Contraction Consistency is thought desirable in the voting context is that violating it leads to susceptibility to tactical voting. Again, consider *Hiring Decision*. If virtue theory could pretend that its preference ordering was B≻A≻C rather than A≻B≻C, then it could guarantee that its second-favoured option 'won', rather than its least-favoured option. And, indeed, the Borda Rule is often dismissed for being extremely susceptible to tactical voting. However, as I have noted, while tactical voting is a real problem when it comes to aggregating the stated preferences of people, it is no problem at all in the context of decision-making under moral uncertainty. Theories aren't agents, and so there's no way that they can conceal their choice-worthiness

ordering. If a decision-maker were to pretend that one theory's choice-worthiness ordering were different than it is, she would only be deceiving herself.

Second, we have reasons for positively expecting violations of Contraction Consistency in this context. To see this, I'll introduce an analogy with aggregating the competing claims of individuals. A fairly common view within first-order ethics is that small benefits to many can't outweigh a sufficiently large benefit to one person, but some other trade-offs are acceptable.[80] So, perhaps, one should choose to cure a huge number of headaches over a merely large number of people with broken arms, in the choice between those two options; and one should choose to cure a large number of broken arms over curing one person of AIDS. Nonetheless, according to the view I'm considering, in the case of a choice between curing one person of AIDS or a huge number of headaches, one should cure the one person of AIDS. This view, as I understand it, violates Contraction Consistency. When all three options are available, on this view, one should cure the broken arms, because it is better to cure the many broken arms than to cure the one person of AIDs, and the claim of those suffering from headaches has no legitimate force when the option of curing someone of AIDS is available. However, when the option of curing the person of AIDS is removed, then one should cure the headaches, because the claims of those suffering from headaches do have legitimate force against the claims of those suffering from broken arms. One potential explanation for this phenomenon is that one *disrespects* the person with AIDS by curing the headaches, but one would not disrespect them by curing broken arms.

[80] E.g. (Brink 1993; Scanlon 1998, 238–41).

The reason why contraction consistency is violated is that whether or not one shows disrespect by choosing a particular option depends on what other options are available.

Now, one might not find the above view plausible. But, whether or not that view is plausible in that context, we can use a very similar analysis to understand our voting situation. Consider again our problematic case:

> 30% credence in virtue theory, according to which A≻B≻C.
>
> 30% credence in consequentialism, according to which B≻C≻A.
>
> 40% credence in non-consequentialism, according to which C≻A≻B.

In this case, there is no reasonable complaint against choosing C. Virtue theory and consequentialism could 'complain' that B is preferred by a majority to C, so B should be chosen over C. But non-consequentialism could legitimately respond by noting that the exact same argument would apply to choosing A over B. There is a deadlock of majority opinion, so the option that has the largest majority opinion behind it, namely option C, should win. In the two-option case, however, when A is dropped from the option-set, this response from non-consequentialism is not possible. The majority prefers B to C, but there is no corresponding reason against choosing C. So virtue theory and consequentialism have a stronger claim to B being the winning option, and non-consequentialism has no good response.

If we understand decision-making under moral uncertainty as adjudicating among competing claims of different moral theories, violations of Contraction Consistency are to be expected. The claims that different theories have depend crucially on the other

options that are available in this option-set, and when the desirability of some option depends on which other options are available, violation of Contraction Consistency is to be expected.[81] So we should not take violation of Contraction Consistency to be a reason to reject the Borda Rule as a way of taking normative uncertainty across merely ordinal theories.

VII. Conclusion

The problem of intertheoretic comparisons is generally considered to be *the* problem facing normative accounts of decision-making under moral uncertainty. It is often assumed that, if theories are intertheoretically incomparable, then all accounts of decision-making under moral uncertainty are doomed — we should just go back to ignoring moral uncertainty, or to assuming our favourite moral theory to be true when deciding what to do.

This chapter has shown the above assumption to be false. Our metanormative theory should be sensitive to the information that theories give the decision-maker. And even in the situation in which choice-worthiness is merely ordinally measurable across all theories in which the decision-maker has non-zero credence, there is a plausible way to take decision-theoretic uncertainty into account, namely the Borda Rule.

However, even in conditions of intertheoretic incomparability we often have more information than merely ordinal information. Theories can give cardinal choice-

[81] This is argued convincingly, and at length, by (A. Sen 1993).

worthiness, yet be incomparable with each other. How to take normative uncertainty into account in that informational condition is what I turn to in the next chapter.

Chapter 3: Variance Voting

Introduction

In the last chapter, we discussed how to take into account normative uncertainty over merely ordinal and non-comparable theories. But, very often, theories will provide cardinally measurable choice-worthiness functions. This chapter discusses how to take into account normative uncertainty over cardinally measurable but non-comparable theories. Once again, I make use of the analogy between decision-making in this context and with voting.

In section I, I give examples of cardinal theories where it's plausible to think that these theories are incomparable with each other. From section II onwards I discuss what to do in such cases. In section II, I consider but reject the idea that one should use the Borda Rule in such situations. I then consider Ted Lockhart's idea that, in conditions of intertheoretic incomparability, one should treat each theory's maximum and minimum degree of choice-worthiness within a decision-situation as equal, and then aggregate using MEC. This is the analogue of Range voting.

I consider Sepielli's objection that the principle is arbitrary, but argue that the idea of giving every theory 'equal say' has the potential to make the account non-arbitrary. However, in section III, I argue that Lockhart's suggestion fails by this principle, and that what I call Variance Voting is uniquely privileged as the account that gives incomparable theories equal say. I give intuitive examples in favour of this, and then

show, in section IV that, on either of two ways of making the principle of 'equal say' precise, it is only Variance Voting that gives each theory equal say.

In section V, I discuss of how to aggregate moral uncertainty in conditions where one has positive credence in some merely ordinal theories, some cardinal but non-comparable theories, and some theories that are cardinal and comparable with each other. In section VI, I discuss whether this account should be defined only within a particular decision-situation, or if it should be defined over all possible decision-situations.

I. Intertheoretic Incomparability

A problem that has dogged accounts of decision-making under moral uncertainty is how to make intertheoretic comparisons of choice-worthiness differences.[82] The problem is as follows. All a normative theory needs to provide, one might suppose,[83] is a statement of all the truths of the relation "*A* is at least as choice-worthy as *B*," where *A* and *B* represent possible options. If the choice-worthiness relation of the moral theory orders all options and satisfies the von Neumann-Morgenstern axioms,[84] then we can interpret it as giving a concept of choice-worthiness that is measurable on an interval scale. This means that we can represent this choice-worthiness relation using a choice-

[82] The problem is normally called the 'problem of intertheoretic comparisons of value'. But this is somewhat misleading. What I and the others who have explored decision-making under normative uncertainty are primarily interested in is comparing *choice-worthiness* across moral theories, rather than comparing *value* across theories.

[83] I deny this supposition in the following chapter; but assuming it provides a particularly clear way of understanding of where the problem comes from.

[84] Namely, Transitivity, Completeness, Continuity and Independence. See (Broome 1995) for discussion of these axioms in relation to moral theory.

worthiness function such that it's meaningful to say that the difference in choice-worthiness between two options *A* and *B*, according to the theory, is greater than, less than, or equal to, the difference in choice-worthiness between two other options *C* and *D*. But, importantly, the choice-worthiness function is only unique up to a positive affine transformation: if one multiplies that numerical representation by a positive constant, or adds any constant, then one still represents the same choice-worthiness ordering. The choice of unit is arbitrary, and so, from the normative theories alone, even though we can meaningfully talk about magnitudes of choice-worthiness *within* a normative theory, we just don't have enough information to enable us to compare magnitudes of choice-worthiness *across* normative theories.[85] But if so, then we cannot apply MEC.

There are really two problems that fall under the label of 'the problem of intertheoretic choice-worthiness comparisons'. The first problem is: "When, if ever, are intertheoretic choice-worthiness comparisons possible, and in virtue of what are true intertheoretic comparisons true?" I address this question in the next chapter. The second problem is: "Given that choice-worthiness sometimes is incomparable across first-order normative theories, what is it appropriate to do in conditions of normative uncertainty?" I focus on this latter question in in this chapter, in the situation where the non-comparable theories are cardinal.

To show that it's plausible that theories sometimes are cardinal but incomparable, I'll give three examples. First, consider prioritarianism and utilitarianism. Both of these views make the same recommendations in situations that involve saving identical lives

[85] A similar problem arises in the study of social welfare in economics: it is desirable to be able to compare the strength of preferences of different people, but even if you represent preferences by cardinally measurable utility functions you need more information to make them comparable.

under uncertainty. On both views, a 50% chance of saving two lives with the same lifetime wellbeing and a guarantee of saving one of those lives are equally choice-worthy. So, according to both of these theories, saving two identical lives is twice as good as saving one. So one might think that one can use this 'agreement' between the two theories on the difference in choice-worthiness between saving one life and saving two as a common measure.[86]

To see that this doesn't work, consider Annie and Betty. For each of these people, if you administer a certain drug they'll each live for 9 more years. Both utilitarianism and prioritarianism agree that the difference in choice-worthiness between doing nothing and saving both Annie and Betty is exactly twice as great as the difference in choice-worthiness between doing nothing and saving Annie alone. For concreteness, we'll assume that the prioritarian's concave function is the square root function. And we'll begin by assuming that Annie and Betty have lived for 16 years so far. If so, then the prioritarian claims that the choice-worthiness difference between saving both Betty and Annie's lives and saving Annie's life alone is $\sqrt{25} - \sqrt{16}$, which equals 1. The utilitarian claims that this difference is 25 – 16, which equals 9. So if we are normalizing the two theories at the difference between saving one life and saving two, then 1 unit of choice-worthiness, on prioritarianism, equals 9 units of choice-worthiness, on utilitarianism.

But now suppose that both Annie and Betty had lived much longer. Suppose they had lived for 64 years each. In which case, then the difference in choice-worthiness, on prioritarianism, between saving both Betty and Annie's lives, and saving Annie's life alone is $\sqrt{73} - \sqrt{64}$, which equals ~0.5. The utilitarian, in contrast, claims that this

[86] Both (Ross 2006, 764) and (Sepielli 2009) make suggestions that could be interpreted on these lines.

difference is 74 – 64, which equals 9. So, if we are normalizing the two theories at the difference between saving one life and saving two in this case, then 1 unit of choice-worthiness, on prioritarianism, equals approximately 18 units of choice-worthiness, on utilitarianism. But this is inconsistent with our previous conclusion. Applying the "normalize at the difference between saving one life and saving two" rule gives different answers depending on which two lives we're talking about.

So we cannot consistently normalize utilitarianism and prioritarianism at the difference ratio between saving one life and saving two lives, and saving two lives and saving no lives. Utilitarians and prioritarians agree on the ratio of choice-worthiness differences in every instance of saving identical lives. But whereas the prioritarian claims that that difference in value between saving one life and saving two lives gets smaller the longer those lives have already been lived for, the utilitarian claims that the value difference is always the same, no matter how long those lives have already been lived for. But if we cannot normalize utilitarianism and prioritarianism in this way, then it seems very difficult to see how there could be any principled way of claiming that there is a unit of value that is shared between utilitarianism and prioritarianism. So one might reasonably think that they cannot be placed on a common scale.

Second, consider average utilitarianism (AU) and total utilitarianism (TU).[87] Suppose that the decision-maker is equally divided between these two theories of population ethics. And suppose that there are three possible outcomes to choose between:

A: N people at 100 wellbeing

[87] I thank Toby Ord for this example.

B: 10*N people at 99 wellbeing

C: 1000*N people at 1 wellbeing

A has an average wellbeing of 100, and a total wellbeing of 100N. B has an average wellbeing of 99, and a total wellbeing of 990N. C has an average wellbeing of 1, and a total wellbeing of 1000N. According to average utilitarianism, B is 99% as valuable as the best option (option A). According to total utilitarianism, B is 99% as valuable as the best option (option C). And this is true, for both theories, no matter what number N is. So you might think, intuitively, that if you are unsure between average and total utilitarianism, then it's appropriate to choose option B, no matter what number N is. B seems to represent the best hedge between the two views.

However, it is impossible to consistently do this for all values of N. If you think that B should be chosen for some value of N, then for some other value you must say that A should be chosen, and for some other value you must say that C should be chosen.

Suppose, for example, that N = 1000. If so, then the value of A, according to AU, is 100. The value of C, according to TU, is 1 million. If B is the appropriate choice, then the following must be true (where M is the conversion factor from AU to TU):

Expected choice-worthiness (B) > Expected choice-worthiness (C)

That is:

0.5*99M+ 0.5*990,000 > 0.5*1M + 0.5*1,000,000

Solving for M, we have: M > 1000/9

But now suppose that N = 1. If so, then the value of A, according to AU, is again 100. The value of C, according to TU, is 1. If B is still the appropriate choice, then the following must be true (where, again, M is the conversion factor from AU to TU):

> Expected choice-worthiness (B) > Expected choice-worthiness (A)
>
> That is:
>
> 0.5*99M+ 0.5*990 > 0.5*100M + 0.5*1000
>
> Solving for M, we have: 10 > M

But these two statements are inconsistent. If you want to hedge between AU and TU when the number of people at stake is small, then, when the number of people at stake is large, TU swamps AU. If you want to hedge between AU and TU when the number of people at stake is very large, then, when the number of people at stake is small, AU swamps TU. So, again, our intuitions about how two theories falter. And in the absence of those intuitions, it's unclear how any way of comparing AU and TU could be correct.

For a third example, consider Jason. He is certain in consequentialism, but splits his credence between three axiological views:

> T_1: Only pleasure and pain are of intrinsic value. Objective goods are not of intrinsic value.
>
> T_2: Only objective goods, such as friendship, achievement, and appreciation of beauty, are of intrinsic value. Pleasure and pain are not of intrinsic value.
>
> T_3: Pleasure and pain and objective goods are of intrinsic value.

Jason is also certain that, if T_3 is true, then objective goods are absolutely incommensurable with pleasure and pain: there is simply no fact of the matter about how an increase in objective goods compares with an increase in pleasure, or a decrease in pain. Given Jason's credence distribution, it would be bizarre if all three theories were comparable with one another.

To see this, suppose that the value of pleasure is comparable between T_1 and T_3, and that the value of objective goods is comparable between T_2 and T_3. And now assume that T_1 and T_2 are comparable. If so, then we have a way of comparing the value of pleasure (on T_1) with the value of objective goods (on T_2). So it then seems incoherent for Jason to be certain that pleasure and objective goods are absolutely incommensurable on T_3.

Alternatively, suppose the value of pleasure on T_3 is comparable with both the value of pleasure on T_1 and the value of objective goods on T_2. If so, then we have a way of comparing the value of pleasure on T_1 with the value of objective goods on T_2. But if so, then, again, it seems incoherent for Jason to be certain that pleasure and objective goods are absolutely incommensurable on T_3.

Finally, suppose that the value of objective goods on T_3 is comparable with both the value of pleasure on T_1 and the value of objective goods on T_2. If so, then exactly the same argument as before would apply: these comparisons would provide a way of comparing T_1 and T_2, making it seem incoherent for Jason to be certain that pleasure and objective goods are incomparable on T_3.

So these are three cases where choice-worthiness differences seem to be incomparable between different theories. But if we have no way of making the intertheoretic comparison, then we cannot take an expectation over moral theories. So it's unclear what a decision-maker under moral uncertainty should do if she faces theories that are cardinally measurable but intertheoretically incomparable. So we need an account of what it's appropriate to do in conditions where we cannot put two different normative theories on a common scale. Let us now look at some contenders.

II. Two unsatisfactory proposals

One might initially think that my work in the previous chapter gives a solution. When theories are intertheoretically incomparable, one should aggregate those theories' choice-worthiness orderings using the Borda Rule.

The problem with this proposal should be obvious. Consider the following decision-situation:

	T_1 – 50%	T_2 – 50%
A	10	0
B	9	90
C	0	100

In this case, the difference between B and C, on T_1, is far greater than the difference between A and B. Similarly, the difference between A and B, on T_2, is far greater than the difference between B and C. Yet the difference between the Borda Scores of A and

B is the same as the difference in the Borda Scores between B and C, on both theories. The Borda Rule therefore seems to misrepresent the theories themselves, throwing away cardinal information when we have it. The voting analogy might prove useful, but ignoring cardinal information when we have it is not the way to proceed.

Lockhart has suggested a different account: what he calls the 'Principle of Equity among Moral Theories'. He defines it as follows:[88]

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

It's ambiguous whether Lockhart thinks that the PEMT is giving an account of how two theories actually compare, or whether he is giving an account of what to do, given that all theories are incomparable. In the above quote it sounds like the latter, because he says, "should be considered" rather than "is". But in other parts of the text, he gives the impression that he believes this is how moral theories actually compare. In this chapter I'll just consider the idea that the PEMT can be used as an account of what to do given that all theories are incomparable. (In the next chapter I will consider whether accounts similar to Lockhart's are plausible as accounts of how choice-worthiness actually compares intertheoretically, and argue that they are not.)

On this understanding, Lockhart's account is analogous to Range Voting. On Range Voting, every voter can give each candidate a score of between 1 and 10. The elected candidate is the candidate whose sum total of scores across all voters is highest. Range

[88] (Lockhart 2000, 84)

Voting, as I understand it, is not typically suggested as a way of actually comparing strengths of preference across different voters. Rather, it's an account of how to aggregate voters' preferences in conditions where there is no fact of the matter about how voters' preferences compare, where there is no way of determining how voters' preferences compare, or where, due to reasons of fairness, it is not desirable to take into account voters' actual preference-strengths.

To illustrate Lockhart's account, let's look again at the previous table. If we were to take the numbers in the table at face value, then we would suppose that the difference between B and C, on T_2, is ten times as great as the difference between A and B, on T_1. But to do so would be to forget that the theory's choice-worthiness functions are unique only up to a positive affine transformation. According to Lockhart's proposal we should treat the best and worst options as equally choice-worthy. So we should treat the choice-worthiness of $CW_1(A)$ as the same as the choice-worthiness of $CW_2(C)$ and we should treat the choice-worthiness of $CW_1(C)$ as the same as the choice-worthiness of $CW_2(A)$. Treated in this way, the theories look as follows:

	T_1 – 50%	T_2 – 0.5%
A	10	0
B	9	9
C	0	10

What seems promising about Lockhart's account is that it provides a way of taking into account normative uncertainty across cardinal theories that are incomparable. However, Lockhart's account has come under serious fire in a recent article by Andrew

Sepielli.[89] Most of the problems with his account arise from the fact that his account treats maximum and minimum degrees of choice-worthiness as the same within a decision-situation rather than across all possible decision-situations. This issue arises for all the accounts I consider in this chapter. I'll discuss this issue further in section VI, ultimately deciding that we should endorse the 'across all possible decision-situations' formulation. But before then, I want to discuss a different problem, which is that the PEMT is *arbitrary*.[90]

There is a wide array of alternatives to Lockhart's view. Why, one might ask, should one treat the maximum and minimum choice-worthiness as the same, rather than the difference between the most choice-worthy option and the mean option, or between the least choice-worthy option and the mean option? Or why not treat the mean difference in choice-worthiness between options as the same for all theories?

Lockhart anticipates this objection, stating:[91]

> It may appear that I have, in an ad hoc manner, concocted the PEMT for the sole purpose of defending the otherwise indefensible claim that moral hedging is possible.

However, he responds as follows:

> The PEMT might be thought of as a principle of fair competition among moral theories, analogous to democratic principles that support the equal counting of

[89] (Sepielli 2012).
[90] Sepielli says (2012, 587): "perhaps the most telling problem with the PEMT is that it is arbitrary."

[91] (2000, 86).

> the votes of all qualified voters in an election regardless of any actual differences in preference intensity among the voters…. PEMT appears not to play favorites among moral theories or to give some type(s) of moral theories unfair advantages over others.

That is, he appeals to what I'll call the *principle of equal say*: the idea, stated imprecisely for now, that we want to give equally likely moral theories equal weight when considering what it's appropriate to do.

As Sepielli points out, this idea doesn't seem at all plausible if we're trying to use the PEMT as a way of actually making intertheoretic comparisons. Considerations of fairness are relevant to issues about how to treat *people*: one can be unfair to a person. But one cannot be unfair to a theory. Perhaps by saying that one was being 'unfair' to Kantianism, one could mean that one's degree of belief was too low in it. But one can't be unfair to it insofar as it 'loses out' in the calculation of what it's appropriate to do. If a theory thinks that a situation is low stakes, we should represent it as such.

But I think the idea of equal say has more plausibility if we are talking about how to come to a decision in the face of genuine intertheoretic incomparability. In developing an account of decision-making under moral uncertainty, we want to remain neutral on what the correct moral theory is: we don't want to bias the outcome of the decision-making in favour of some theories over others. And if we're in a condition where there really is no fact of the matter about how two theories compare, then we can't make sense of the idea that things might be higher-stakes in general for one theory rather than

the other. So we need a way of taking uncertainty over those theories into account that isn't biased towards one theory rather than another.

To see a specific case of how this could go awry, consider average and total utilitarianism, and assume that they are indeed incomparable. And suppose that, in order to take an expectation over those theories, we choose to treat them as agreeing on the choice-worthiness ordering of options concerning worlds with only one person in them. If we do this, then, for almost all decisions about population ethics, the appropriate action will be in line with what total utilitarianism regards as most choice-worthy because, for almost all decisions, the stakes are huge for total utilitarianism, but not very large for average utilitarianism. So it seems that, if we treat the theories in this way, we are being partisan to total utilitarianism. In contrast, if we chose to treat the two theories as agreeing on the choice-worthiness differences between options with worlds involving 10^100 people then, for almost all real-world decisions, what it's appropriate to do will be the same as what average utilitarianism regards as most choice-worthy. This is because we're representing average utilitarianism as claiming that, for almost all decisions, the stakes are much higher than for total utilitarianism. In which case, it seems that we are being partisan to average utilitarianism, whereas what we want is to have a way of normalising such that each theory gets equal influence.

Lockhart states that the PEMT is the best way to give every theory equal say. But he doesn't argue for that conclusion, as Sepielli notes:[92]

> But even granting that some "equalization" of moral theories is appropriate, Lockhart's proposal seems arbitrary. Why equalize the maximum and minimum

[92] (2012, 587–8).

> value, rather than, say, the mean value and the maximum value?... It seems as though we could find other ways to treat theories equally, while still acknowledging that the moral significance of a situation can be different for different theories. Thus, even if we accept Lockhart's voting analogy, there is no particularly good reason for us to use PEMT rather than any of the other available methods.

In a very similar vein, Amartya Sen has argued against an analogue of the PEMT within social choice theory, the 'zero-one' rule:[93]

> It may be argued that some systems, e. g., assigning in each person's scale the value 0 to the worst alternative and the value 1 to his best alternative are interpersonally "fair" but such an argument is dubious. First, there are other systems with comparable symmetry, e.g., the system we discussed earlier of assigning 0 to the worst alternative and the value 1 to the sum of utilities from all alternatives.

I think both Sen and Sepielli are right that principled reasons for endorsing the PEMT over its rivals have not been given. But, further to that, I think that it's demonstrably *false* that the PEMT is the best way of giving each theory equal say. Instead, I think that what I'll call Variance Voting is the best account of taking normative uncertainty across cardinal and incomparable theories, because it is the best way of giving each theory equal say. I'll now turn to the reasons why I think this.

III. Variance Voting

I'll call Lockhart's view and its rivals *cardinal voting systems*. To begin to get a sense of how different cardinal voting systems can differ in how they apportion 'say' between theories, let's consider some examples. Let's consider four different cardinal voting systems using

[93] (1970, 98). In contrast to Range Voting, the zero-one rule is normally understood as an account of how preferences actually compare across people.

the 'across all decision-situations' formulation of each: (i) Lockhart's PEMT, which treats the range of the choice-worthiness function as the same across all cardinal and incomparable theories; (ii) what I'll call Max-Mean, which treats the difference between the mean choice-worthiness and the maximum choice-worthiness as the same across all cardinal and incomparable theories; (iii) what I'll call Mean-Min, which treats the difference between the mean choice-worthiness and the minimum choice-worthiness of all cardinal and incomparable theories as the same (this is the account that Sen suggests in the above quote); and (iv) Variance Voting, which treats the average of the squared differences in choice-worthiness from the mean choice-worthiness as the same across all theories. Intuitively, the variance is a measure of how spread out choice-worthiness is over different options; normalising at variance is the same as normalising at the difference between the mean choice-worthiness and one standard deviation from the mean choice-worthiness.[94]

As well as considering four different cardinal voting systems, let's consider four types of first-order normative theory. We'll call the first type *Bipolar* theories. According to Bipolar theories, the differences in choice-worthiness between different highly choice-worthy options, and between different highly un-choice-worthy options, are zero or tiny compared to the difference in choice-worthiness between highly choice-worthy options and highly un-choice-worthy options. For example, a view according to which violating

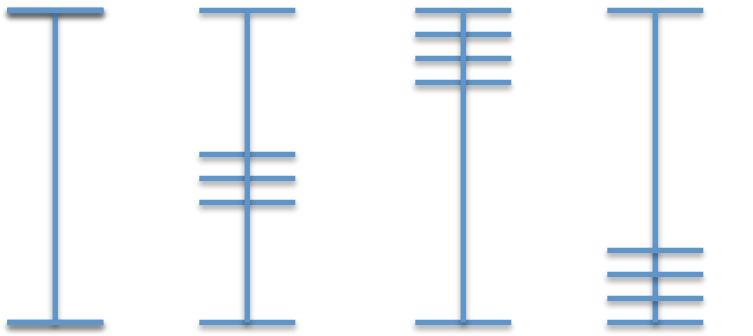
[94] In order to make sense of the variance of a choice-worthiness function, we need a notion of *measure* over possibility space. This is discussed in chapter 2, section VI, in relation to the Borda Rule. I assume that we should use the same choice of measure when using Variance Normalisation as we did when using the Borda Rule. Having a measure over the option-set allows variance normalization to apply to many unbounded moral theories: I take this to be yet another advantage of Variance Normalisation over the PEMT.

rights is impermissible, everything else is permissible, and where there is very little difference in wrongness between different wrong action, would be a Bipolar theory.

We'll call the second type of theory *Outlier* theories. According to this view, most options are roughly similar in choice-worthiness, but there are some options that are extremely choice-worthy, and some options that are extremely un-choice-worthy. A bounded total utilitarian view with a very high and very low bounds might be like this: the differences in value between most options are about the same, but there are some possible worlds which, though unlikely, are very good indeed, and some other worlds which, though unlikely, are very bad indeed.

We'll call the third type of theory *Top-Heavy*. According to this type of theory, there are a small number of outliers in choice-worthiness, but they are only in one direction: there are just a small number of extremely un-choice-worthy possible options. Any consequentialist theory that has a low upper bound on value, but a very low lower bound on value, such that most options are close to the upper bound and far away from the lower bound, would count as a Top-Heavy moral theory. The fourth type of theory is *Bottom-Heavy*. These are the inverse of Top-Heavy theories.

We can represent these theories visually, where horizontal lines represent different options, which are connected by a vertical line, representing the choice-worthiness function. The greater the distance between the two horizontal lines, the greater the difference in choice-worthiness between those two options. If we used PEMT, the four theories would look as follows:



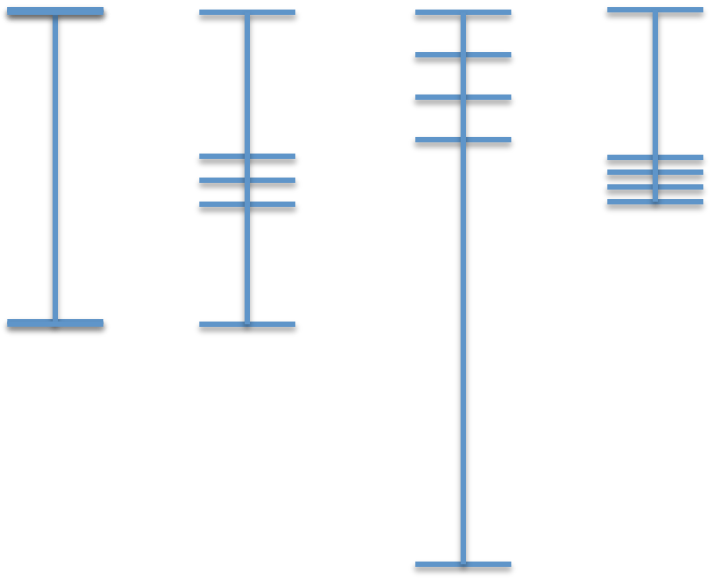
For Top-Heavy and Bottom-Heavy, the PEMT yields the right result. Top-Heavy and Bottom-Heavy are simply inversions of each other, so it seems very plausible that one should treat the magnitudes of choice-worthiness differences as the same according to both theories, just of opposite sign.

For Bipolar and Outlier, however, the PEMT does not yield the right result. Because it *only* cares about the maximal and minimal values of choice-worthiness, it is insensitive to how choice-worthiness is distributed among options that are not maximally or minimally choice-worthy. This means that Bipolar theories have much more power, relative to Outlier theories, than they should.

This might not be immediately obvious, so let us consider a concrete case. Suppose that Sophie is uncertain between an absolutist moral theory, and a form of utilitarianism that has an upper limit of value of saving ten billion lives, and a lower limit of forcing ten billion people to live lives of agony, and suppose that those views are incomparable with each other. She has 1% credence in the absolutist theory, and 99% credence in

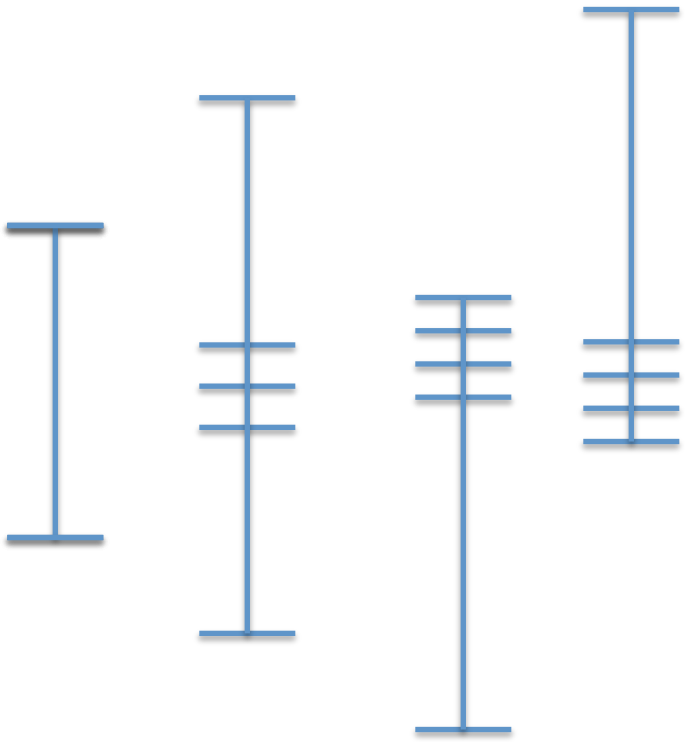
bounded utilitarianism. If the PEMT normalisation is correct, then in almost every decision-situation she faces she ought to side with the absolutist theory. Let's suppose she is confronted with a murderer at her door, and she could lie in order to save her family: an action required by utilitarianism, but absolutely wrong according to the absolutist view. Given the PEMT, it's as bad to lie, according to the absolutist view, as it is to force ten billion people to live lives of agony, according to utilitarianism. So her 1% credence in the absolutist view means that she shouldn't lie to the murderer at the door. In fact, she shouldn't lie even if her credence was as low as 0.000001%. That seems incredible. The PEMT is supposed to be motivated by the idea of giving each moral theory 'equal say', but it flagrantly fails to do this in cases where some theories put almost all options into just two categories.

For a second illustration of how structural accounts can fail to respect the Principle of Equal Say, consider the Max-Mean principle. Taking our four theories described above, it would normalise them such that they would be represented as follows:



That is, Max-Mean favours Top-Heavy theories and punishes Bottom-Heavy theories. It's clear, therefore, that Max-Mean does not deal even-handedly between these two classes of theories. Exactly analogous arguments apply to Mean-Min.

What, though, of Variance Voting? If we treat the variance of choice-worthiness as the same across all four theories, they would be represented as follows:



Because Top-Heavy and Bottom-Heavy are inverses of one another, they have the same variance. So, on Variance Voting, the magnitudes of choice-worthiness between options are treated as the same, only with opposite sign. This is the result we wanted, doing better than Max-Mean or Mean-Min. But it also does better than the PEMT in

terms of how it treats Bipolar compared with Outlier: because Bipolar places most of its options at the top or bottom of its choice-worthiness function, in order to make the variance equal with Outlier, its range must be comparatively smaller than Outlier. Again, that was the result we wanted. So the consideration of particular cases seems to motivate Variance over its rivals.

These examples are suggestive, but hardly constitute a knockdown argument. Perhaps there are other normalisation methods that do as well as Variance does on the cases above. Perhaps there are other cases in which Variance does worse than the other methods I've mentioned. So it would be nice to provide a more rigorous argument in favour of Variance. The next two sections do exactly that. I'll suggest two different ways of making the idea of equal say formally precise.[95] I find the second precisification more compelling but I show that, either way, normalising at equal say means normalising at variance. In so doing, I thereby produce a non-arbitrary justification for normalising at variance rather than the range or any other features of a theory's choice-worthiness functions: the variance normalisation is the normalisation that best captures the principle of equal say.

[95] The following two sections draw very heavily on two results within social choice theory that can be found in (Cotton-Barratt ms), available at: http://users.ox.ac.uk/~ball1714/Variance%20normalisation.pdf. These results were initially motivated by the problem of moral uncertainty, arising out of conversation between us, though I had very little input on the proofs. However, they are interesting results within social choice theory, too.

IV. Two arguments for Variance Normalisation

Distance from the uniform theory

My first idea is to think of 'say' in terms of 'credit'. Every theory's baseline is the uniform theory, according to which all options are equally choice-worthy. Every move away from the uniform theory's choice-worthiness assignment costs that theory a proportionate amount of credit. Giving every theory equal say means giving them an equal amount of credit to distribute over options. In this section I'll spell this suggestion out, explain the motivation for it, and demonstrate that Variance Voting is the only account that gives every theory equal say, so understood.

Let's begin by considering different theories that *are* intertheoretically comparable. It should be clear that a completely uniform theory, according to which all options are equally choice-worthy, has no say at all: it never affects what it's appropriate to do. We'll say that it gives all options value 0 (though we could have just as well said it gives all options value 1, or value 9, or any other number, as long as the number is the same for all options). Next, consider a theory, T_1, which differs from the uniform theory only insofar its choice-worthiness function gives one option, option A, a choice-worthiness of *n*. There are two ways in which a theory T_2 might have more say than T_1. First, it could have the same choice-worthiness ordering as T_1, but its choice-worthiness function could give A a higher numerical value (remembering that, because we are talking about theories that are intertheoretically comparable, this is a meaningful difference between these two theories). If it gave A a numerical value of 2**n*, so that the choice-worthiness

difference between A and any other option is twice as great according to T_2 than according to T_1, then T_2 would have twice as much 'say' as T_1.

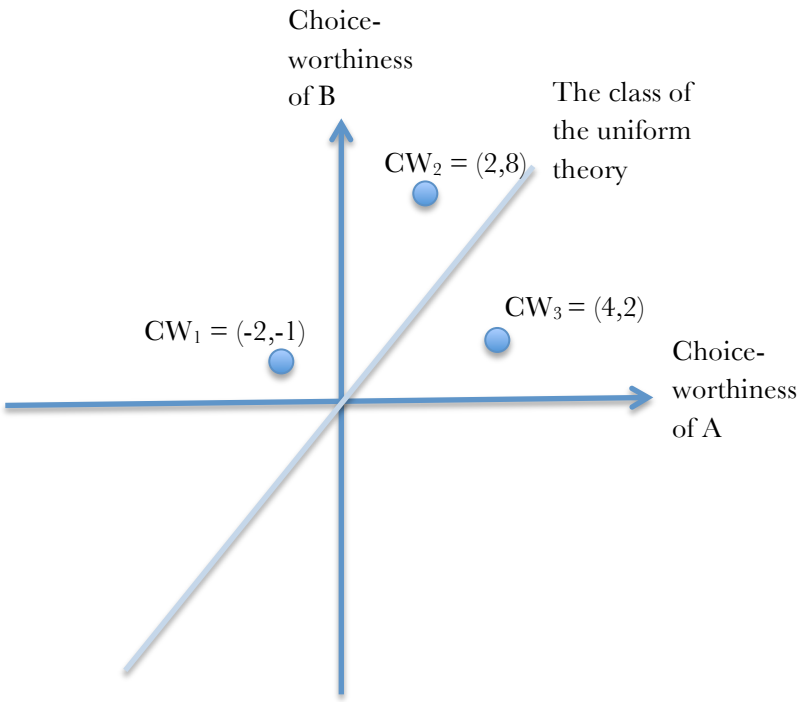
A second way in which a theory could have more 'say' than T_1 is if it also gave A value *n* (agreeing with T_1 on the difference in choice-worthiness between A and the choice-worthiness of options on the uniform theory), but also assigned non-zero numerical values to other options, too. Then it would have equal say with respect to A, but would have a greater say with respect to the other options that it assigns non-zero values to.

But what does 'moving away' from the uniform theory mean? We can take this idea beyond metaphor by thinking of choice-worthiness functions geometrically. While doing this, we have to be careful to distinguish between a *choice-worthiness function* and the *class* of informationally equivalent choice-worthiness functions (for cardinal and incomparable theories, which is what we are discussing, the informationally equivalent class of a choice-worthiness function CW_i is the class of choice-worthiness functions CW_j such that $CW_j = aCW_i + b$, with $a > 0$). As we stated before, a choice-worthiness function gives an assignment of numbers to options such that an option receives a higher number than another if it is more choice-worthy than the other, and an equal number as another if it is equally choice-worthy as the other. For cardinally measurable but non-comparable theories, if some choice-worthiness function CW_i represents a specific choice-worthiness ordering, then any positive affine transformation of CW_i represents that theory just as well. The class of a choice-worthiness function is the class of all choice-worthiness functions that represent that same choice-worthiness ordering.

To see how we can represent choice-worthiness functions geometrically, suppose, to begin with, that there are only two possible options, A and B, and three theories, T_1, T_2 and T_3. These may be represented by the following table:

	T_1	T_2	T_3
A	-2	2	4
B	1	8	2

Using the choice-worthiness of A as the x-axis and the choice-worthiness of B as the y-axis, we may represent this geometrically as follows:



Representing choice-worthiness functions geometrically allows us to visualise the distinction between choice-worthiness functions and classes of choice-worthiness functions. Any point on the graph is some choice-worthiness function.

The grey line is the class of informationally equivalent choice-worthiness functions that represent the uniform theory, that is: the class of CW_i such that, for all *A, B*, $CW_i(A) = CW_i(B)$. Every point on the half-plane to the left of the line of the uniform theory is a choice-worthiness function in the same class of choice-worthiness functions, the class that represents $A \succ B$. (So CW_1 and CW_2 represent the same choice-worthiness ordering. This can be confirmed by considering that, for all *A*, $CW_2(A) = 6*CW_1(A) + 14$.) Every point on the half-plane to the left of the line of the uniform theory is a choice-worthiness function in the same class of choice-worthiness functions, the class that represents $A \succ B$.

A choice-worthiness function gives an assignment of a real numbers to every option, so if there are *n* options a choice-worthiness function can be represented as a collection of

n real numbers. Just as pairs of real numbers give us Cartesian coordinates in the plane, and triples give us coordinates in three-dimensional space, so we can interpret this collection as the coordinates of a point in *n*-dimensional Euclidean space. This is enough to give us a definition of distance between choice-worthiness functions. If $\mathrm{CW}_1 = (u_1, u_2, u_3, \ldots, u_n)$ and $\mathrm{CW}_2 = (v_1, v_2, v_3, \ldots, v_n)$, then the distance between them is $\sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$. So the distance between the choice-worthiness functions CW_1 and CW_2 is $\sqrt{97}$ and the distance between CW_1 and CW_3 is $\sqrt{45}$.

This definition of distance allows us to make sense of our precisified notion of 'equal say'. Giving each theory 'equal say' means choosing a (normalised) choice-worthiness function for each theory such that, for every choice-worthiness function, the distance from that choice-worthiness function to the line that represents the class of uniform theories is the same.

To work out the distance from any choice-worthiness function to the uniform theory, let us consider some choice-worthiness function $\mathrm{CW}_1 = (u_1, u_2, u_3, \ldots, u_n)$. Let $\mathrm{CW}_1' = (u_1 - m, u_2 - m, u_3 - m, \ldots, u_n - m)$, where m is the mean of CW_1. CW_1' is a linear shift of CW_1 parallel to the line that represents the uniform theory, so CW_1' is the same distance from the uniform theory as CW_1. However, for CW_1' the closest choice-worthiness function that represents the uniform theory lies at CW_u = (0,0,0, … 0). The distance between these two points is $\sqrt{\sum_{i=1}^{n}(u_i - m)^2}$, which is the standard deviation of CW_1. So treating all choice-worthiness functions as lying at the same distance from the uniform theory means treating them such that they have the same standard deviation, or, alternatively, variance.

It should be noted that there aren't decisive reasons for preferring to understand distance between choice-worthiness functions in terms of Euclidean geometry rather than the taxicab geometry, according to which the distance between two points equals the sum of the absolute differences between the co-ordinates.[96] Even though Euclidean geometry is more familiar, I find both geometries to generate intuitive notions of distance between choice-worthiness functions. If we use the taxicab geometry, then to treat choice-worthiness functions such that they are equidistant from the uniform theory is to treat them such that the average distance from the median option is the same. This proposal would also get the correct intuitive results in the cases given in the previous section. So I think that the argument in this section should merely narrow down the contenders to Variance Voting and the average-distance-to-the-median view. The next argument, however, singles out Variance Voting uniquely.

The Expected Choice-Worthiness of Voting

The previous argument cashed out the idea of 'equal say' as 'equal distance from the uniform theory'. In the second argument, I'll borrow a concept from voting theory and utilize the notion of *voting power*. An individual's voting power is the *a priori* likelihood of her vote being decisive in an election, given the assumption that all the possible ways for other people to vote are equally likely. It is normally used for elections with just two candidates, but the concept is perfectly general. On a first pass attempt of cashing out the intuition of equal say, theories that are equally probable are given equal voting

[96] This is discussed further in (Cotton-Barratt ms), who also discusses (and rejects) using the l^{∞} metric, according to which the distance from point a to point b is the size of the greatest co-ordinate difference between point a and point b.

power. In the context of decision-making under normative uncertainty, this would mean treating the choice-worthiness differences across each each theory in such a way that, for each theory and for some randomly selected decision-situation, the likelihood that that theory will be decisive in that decision-situation is in proportion with how much credence the decision-maker has in that theory.

However, this isn't quite right. What a normative theory cares about is not just whether it is decisive in some particular decision-situation. It also cares by *how much* it wins. Getting its way in a decision between whether to prick someone with a pin matters a lot less, for utilitarianism, than getting its way in a decision about whether to let a million people die. If we are normalising at 'equal say', we should take that into account as well. That is, rather than voting power, we should use the *expected choice-worthiness of voting*: the *a priori* expected choice-worthiness that a theory gets from participating in the process of decision-making under normative uncertainty, rather than being ignored.

So a second way of making precise the principle of equal say is that every theory's expected choice-worthiness of voting should be in proportion with how much credence the decision-maker assigns to that theory. And it has been proved that Variance Normalisation is the only way of giving every choice-worthiness function an equal expected choice-worthiness of voting.[97]

Given that we have found two independent plausible ways of cashing out the principle of equal say that both lead to the same conclusion, I think it is warranted to think of the Variance Voting as strongly supported by that principle.

[97] The proof is too long to be given here, but can be found in (Cotton-Barratt ms), available at http://users.ox.ac.uk/~ball1714/Variance%20normalisation.pdf.

V. How to aggregate uncertainty in varying information conditions

In the first chapter we discussed how to take normative uncertainty into account in conditions where theories' choice-worthiness is cardinally measurable and intertheoretically comparable. In the previous chapter we discussed how to take normative uncertainty into account in conditions where theories give merely ordinal choice-worthiness. And in this chapter we discussed how to take normative uncertainty into account in conditions where theories give cardinal choice-worthiness but are intertheoretically incomparable. But how should we put these different criteria together? I will call the view that the correct metanormative theory is sensitive to the different amounts of information that different theories give the *Hybrid View*.

I used to think that the Hybrid View should take the form of multi-step procedure.[98] The idea is as follows. At the first step, aggregate all sets of cardinal and mutually intertheoretically comparable theories. For each set, you produce a new choice-worthiness function R_i (R_j … R_n), where R_i assigns numbers to options that represent each option's expected choice-worthiness (given the theories in that set). R_i is given a weight equal to the sum total of the credence of all the theories within the set. At the second step, you aggregate all the new choice worthiness functions R_i-R_n with every cardinal but non-comparable choice-worthiness function using Variance Voting, producing another new choice-worthiness function S. S is weighted by the sum of the decision-maker's credences in all cardinal theories. Then, at the third and final stage, you aggregate S and all merely ordinal theories using the Borda Rule.

[98] The following is very similar to the account I defended in (Crouch 2010).

That proposal suffers from the following fatal problem. Consider a decision-maker with the following credence distribution:[99]

4/9 credence in T_1: A≻B≻C.

2/9 credence in T_2: $CW_2(A) = 20$, $CW_2(B) = 10$, $CW_2(C) = 0$.

3/9 credence in T_3: $CW_3(A) = 0$, $CW_3(B) = 10$, $CW_3(C) = 20$.

Where T_1 is merely ordinal, while T_2 and T_3 are cardinal and comparable. If we use the multi-step procedure, then at the first step, we aggregate T_2 and T_3 to get the following output ordering:

5/9 credence in R_1: C≻B≻A

At the second step, we aggregate T_1 and R_1 using the Borda Rule, which gives option C as the winner. However, this seems like the wrong result. In particular, consider the following credence distribution:

4/7 credence in T_1: A≻B≻C.

0 credence in T_2: $CW_2(A) = 20$, $CW_2(B) = 10$, $CW_2(C) = 0$.

3/7 credence in T_3: $CW_3(A) = 0$, $CW_3(B) = 10$, $CW_3(C) = 20$.

In this decision-situation, using the multi-step procedure would give 'A' as the most appropriate option. So having *lower* credence in T_2 makes the appropriateness ordering better by the lights of T_2. This violates my *Updating* condition given in the last chapter. The reason this happens is because, in the first decision-situation, though T_2's and T_3's

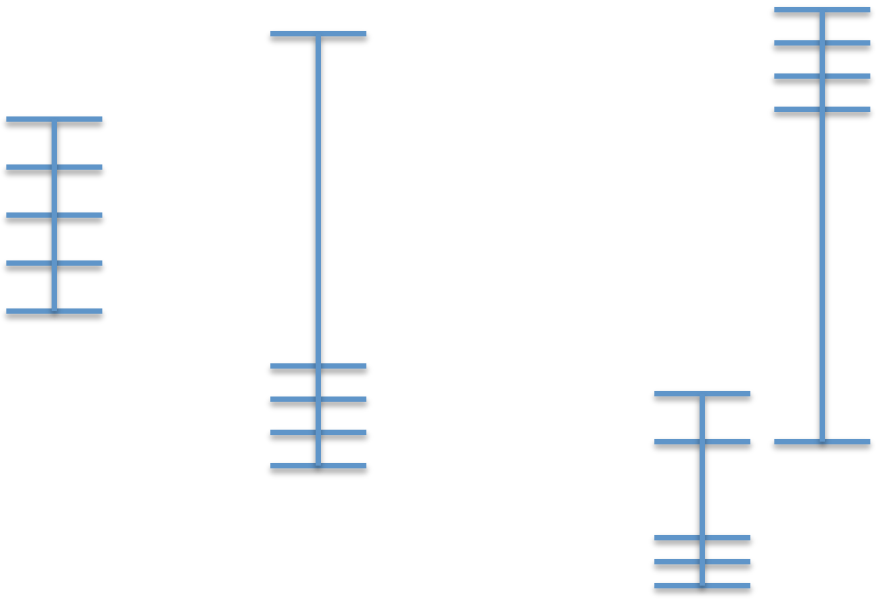
[99] This problem was first raised by Owen Cotton-Barratt; the specific case was given to me by Toby Ord.

choice-worthiness orderings cancel each out to some extent, the multi-step procedure washes this fact out when it pits the aggregated ordering R_1 against the ordinal theory T_1.

So I now, though I still endorse the *Hybrid View*, I prefer a one-step metanormative theory. Because the Borda Rule assigns scores to each option, we can treat the variance of the cardinal theories' choice-worthiness functions and variance of the ordinal theories' Borda Scores as the same.[100] But when we are normalising them with cardinal and comparable theories, we need to be careful. We can't normalise all individual comparable theories with non-comparable theories at their variance. If we were to do so, we would soon find our equalization of choice-worthiness-differences to be inconsistent with each other. Rather, for every set of cardinal and comparable theories, we should treat the variance of the choice-worthiness values of all options on that common scale as the same as the variance of every individual non-comparable theory.

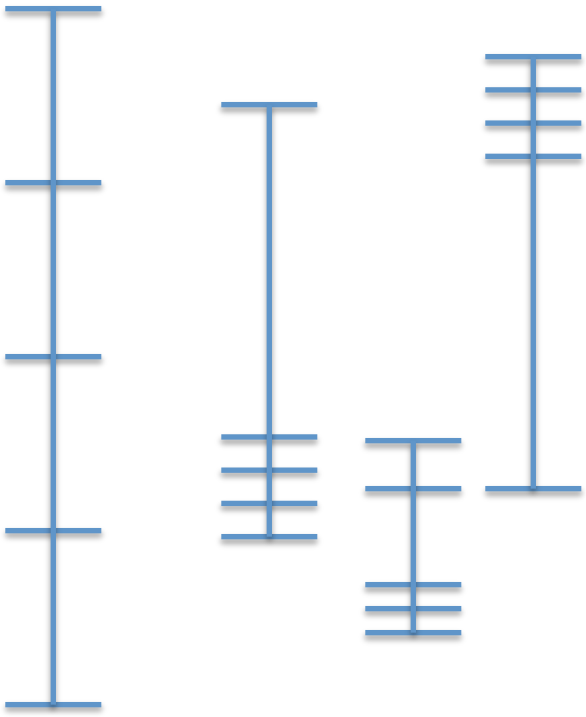
An example helps to illustrate the proposal. Consider four theories, T_1-T_4, in order from left to right:

[100] Doing this does not alter the Borda Rule as presented in chapter 2 when each theory has a strict choice-worthiness ordering over options. However, it *does* make a difference when some theories rate some options as equally choice-worthy to one another. When taking the variance of each theory's Borda Scores to be the same, a theory that ranks A~B≻C~D will weigh comparatively more heavily against D≻C≻B≻A than it would under the precise account I defended in the previous chapter. However, in that previous chapter I defended this specific way of giving Borda scores to tied options with recourse to the principle of equal say, implicitly invoking average-distance-to-the-median as the correct account of equal say. Now that we have seen that normalizing at the variance is the best account of equal say, we should be happy to correct that previous argument.



T_1 is a merely ordinal theory. The diagram illustrates the Borda scores that T_1 assigns to options. T_2 is a cardinal but non-comparable theory. T_3 and T_4 are cardinal and comparable with each other. What the single-step procedure does is to treat the variance of T_1's Borda scores as equal with the variance of T_2's choice-worthiness function and as equal with the variance of the choice-worthiness of options across *both* T_3 and T_4. As should be clear from the diagram, the variance of T_3 is smaller than the variance of T_4. But if T_3 and T_4's variances were each individually normalised with T_2, then the variance of T_3 and T_4 would be the same. So we should not normalise T_3 and T_4 individually with T_2. Rather, it's the variance of the distribution of choice-worthiness on T_3 and T_4's common scale that we treat as equal with other theories.

With their variances treated as equal in the correct way, the theories would look approximately as follows:



Then, once we have done this, we can simply maximise expected choice-worthiness.

This single-step aggregation method wouldn't be possible if we didn't use a scoring function as our voting system in the situation involving merely ordinal theories. Because in my previous work I defended a Condorcet extension (which is therefore not a scoring function) as the correct way to take into account normative uncertainty over merely ordinal theories, I was forced to endorse the multi-step procedure. But, as I said, the objection to the multi-step procedure given above looks fatal. So I take this to provide additional support in favour of the use of a scoring function to aggregate merely ordinal theories, rather than a Condorcet extension.

As a final comment on this, it's nice to see some convergence between two lines of thought. Often, in responses to my work on taking into account normative uncertainty over merely ordinal theories, people make the following objection.[101] They claim that we know that under empirical uncertainty, that expected utility theory or some variant is the correct decision theory. And we should treat normative uncertainty in the same way as empirical uncertainty. So if we encounter a merely ordinal theory, over which one cannot take an expectation, we should either ignore it or we should force some cardinalisation upon it. To this objection I replied that, under empirical uncertainty we rarely or never face merely ordinal choice-worthiness. This is a genuine disanalogy with empirical uncertainty. And to simply force merely ordinal theories to fit into the framework of expected utility theory, rather than to consider how to aggregate merely ordinal theories, is simply not to take one's credence in those merely ordinal theories sufficiently seriously.

Because of my view on this matter, I wished to work out how best to aggregate uncertainty across merely ordinal theories. And it turned out that the Borda Rule was the best account. But, in combination with the one-step aggregation procedure I have defended above, this account should also please my objectors, because, ultimately, one *is* taking an expectation over all moral theories. So, even though I initially thought that the problems of merely ordinal theories and intertheoretic incomparability were reasons to reject MEC as a *general* metanormative theory, we ultimately end up with a sort of vindication of MEC.

[101] I have heard this objection, in some form or other, at least from Michael Tooley, Hilary Greaves, Nick Beckstead, Toby Ord, and to some extent from John Broome.

VI. Broad versus Narrow

Earlier in this chapter I mentioned that Lockhart's PEMT was defined such that it treated the range of all theories' choice-worthiness as the same *within a particular decision-situation*, rather than across all possible decision-situations. Exactly the same distinction can be made with respect to the Borda Rule and Variance Voting. Should they be *Broad* (defined over all conceivable options) or *Narrow* (defined over only the options in a particular decision-situation)?[102]

Most of Sepielli's criticisms of the PEMT arose from the fact that it is Narrow, rather than the fact that it normalises at the range of the theory's choice-worthiness function. His four criticisms are as follows.[103] First, the view cannot make sense of the idea that some decision-situations are higher-stakes for some theories than for others. Second, it generates inconsistent assignments of moral value (claiming that one P is as valuable as one Q in one decision-situation, but that one P is as valuable as two Qs in another decision-situation). Third, it violates Contraction Consistency. Fourth, it makes the appropriateness relation cyclical across decision-situations.

However, in the context of my project these objections lose some of their force. First, we are using the Borda Rule and Variance Voting not as accounts of how theories actually compare, but as a way of coming to a principled decision in the face of incomparable theories. So there isn't a fact of the matter about some decision-situations being higher stakes for some of these theories rather than others. And these accounts aren't generating inconsistent assignments of choice-worthiness, because they aren't

[102] The terminology of 'broad' and 'narrow' for this distinction comes from Amartya Sen (1997, 186).
[103] (Sepielli 2012).

pretending to make claims about how choice-worthiness actually compares across theories. Rather, they are simply showing what it's appropriate to do given that choice-worthiness doesn't compare across theories.

Second, in chapter 5 I will conclude that, in order to handle problems arising from incomplete theories, one has to endorse the idea that the appropriateness relation is cyclical across decision-situations. And in the previous chapter I gave some considerations to show why violations of Contraction Consistency are to be expected in this context.

Moreover, there is a positive reason in favour of preferring Narrow accounts, which is that they are more action-guiding. If I use the Broad Borda Rule, then, for any option I face, I have simply no idea what Borda Score it should receive. But I can come to at least a rough approximation of the options facing me. So I should be able to actually use Narrow methods, at least approximately.

For these reasons, I think it's unclear whether my account should be Broad or Narrow. Ultimately, I tentatively endorse the Broad version. I understand my project as still within the domain of ideal rationality, and of giving a criterion of rightness rather than a decision-procedure. So whether or not the criterion I give is practically useful for decision-makers is not that important. And I take both violations of contraction consistency and cyclicity of appropriateness across decision-situations to be a mark against a theory, even if not a decisive one. The fact that a Narrow account would violate these conditions even if all considered theories were complete still looks like a problem to me. And, most importantly, the principle of equal say seems to be best

cashed out when we use the Broad formulations of both the Borda Rule and Variance Voting. So I think, though am not certain, that we should prefer the Broad versions of these accounts to the Narrow versions.

Conclusion

In this chapter I considered how to take normative uncertainty into account in the situation where the decision-maker has non-zero credence in only cardinal theories that are intertheoretically incomparable. Arguing that the Borda Rule is unsatisfactory in this context, and arguing against Lockhart's PEMT among others, I argued in favour of Variance Voting, on the basis that it best respects the principle of equal say. I then showed how one should aggregate one's uncertainty in varying informational conditions, ultimately giving a vindication of MEC.

This concludes my account of a general metanormative theory. However, we don't yet know much about when theories are comparable and when they are not, nor do we know what makes theories comparable, if and when they are comparable. The next chapter tackles these issues.

Chapter 4: How to Make Intertheoretic Comparisons

Introduction

An intertheoretic comparison of differences of choice-worthiness is a claim that some choice-worthiness difference between two items, *A* and *B*, on one theory T_i is greater, smaller, or the same size as, some choice-worthiness difference between two items, *C* and *D*, on a different theory T_j.

A number of philosophers have questioned whether such comparisons are ever possible. In the first statement of the problem in the modern literature, James Hudson considers two axiological theories: a hedonistic theory, and one according to which only self-realisation is of value. He says (p.225):[104]

> What is the common measure between hedons and reals [the unit of self-realisation]? 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. Clearly this second-order hedging is impossible.

In a follow-up article, Edward Gracely echoed the sentiment:[105]

> Comparisons of the relative weights given to right and wrong choices by different systems are essentially meaningless. I maintain that it is part of the very nature of a moral system that it presents a way of viewing reality, and that the differing visions of different systems cannot be directly compared… So the question becomes, is a small loss of utility as seen by [person-affecting

[104] (Hudson 1989, 225)
[105] (Gracely 1996, 328, 321)

> utilitarianism] 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!)

Similarly, when Jacob Ross considers the problem, he responds to an interlocutor who denies that such comparisons are even meaningful:[106]

> [Ross's argument] presupposes that value differences can be compared across ethical theories. But such comparisons are unintelligible. It is only possible to compare value differences within a theory, not across theories.

John Broome mentions the same problem, though in a slightly weaker form:[107]

> We then encounter the fundamental difficulty. Each different theory will value the change in population according to its own units of value, and those units may be incomparable with one another… 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 kilometers, and a speed, whose unit is kilometers per hour. Most theories of value will be incomparable in this way. Expected value theory is therefore rarely able to help with uncertainty about value.

The suggestion of total or highly widespread incomparability between theories has been the most common presentation of the problem, but it is not the most compelling. First, the arguments for the view are weak, typically considering just one or two examples before moving to a general conclusion, or simply asserting that different theories have different 'conceptions' of value. Second, we have a body of examples where there is an intuitively obvious way to make the comparison. Consider the following propositions:

> (1) James thinks that eating meat is somewhat wrong, but Jane thinks it's *really* wrong.

[106] (Ross 2006, 761)
[107] (Broome 2012, 122)

(2) If Peter Singer is right, then my duties to the global poor are much stronger than I had thought.

(3) Breaking a deathbed promise is much more wrong according to standard non-consequentialist theories than it is according to hedonistic utilitarianism.

On their face, each of these propositions makes an intertheoretic comparison. But these propositions seem clearly meaningful. In fact, they seem clearly to be true. Without a strong argument for the opposing view, I think that it's reasonable to take these impressions at face value. In which case, the question of whether intertheoretic comparisons are *ever* possible is easily answered.

However, there are still deep philosophical questions regarding intertheoretic comparisons. The two most important, in my view, are as follows. The *metaphysical* problem is about what *grounds* intertheoretic comparisons.[108] That is: in virtue of what are intertheoretic comparisons true, when they are true? But, even if we answer this question, there is a further problem to be resolved. This is the *epistemic* problem: how can we tell which intertheoretic comparisons are true, and which are false?

Answering the first problem would tell us about the nature of intertheoretic comparisons — what makes intertheoretic comparisons true. Answering the second problem would enable us, at least to some extent, to more confidently make intertheoretic comparisons: to more confidently know how two theories compare, when they do compare; and to more confidently know whether two theories are comparable at all.

[108] See (Fine 2012) for discussion of the idea of grounding.

The majority of this chapter will be taken up with discussion of the metaphysical problem. However, I take it as a virtue of the account that I ultimately propose that it enables progress to be made on the epistemic problem as well.

There has been limited work done on this problem, and, in the literature that does exist, it's often difficult to know what exactly is at stake when comparing different views. So, in the first section of this chapter, I introduce a scheme for classifying different possible solutions into three distinct categories. In sections II, III, and IV, I argue against solutions that have been proposed in the literature. In section V, I draw on work from the metaphysics of quantity literature to provide a sketch of a solution, arguing for what I call a 'Universal Scale' account.

I. A Classificatory Scheme

The attempt to solve the problems related to intertheoretic comparisons is still in its infancy. So it is useful to impose a classificatory scheme. I suggest we divide classes of accounts into three categories: (i) Structural; (ii) Common Ground; and (iii) Universal Scale.

I define a *Structural* account as claiming that intertheoretic comparisons are true in virtue merely of features each theory's choice-worthiness function (such as the choice-worthiness function's mean choice-worthiness, or maximum choice-worthiness, or variance). Lockhart's PEMT, if understood as an account of intertheoretic comparisons (rather than a proposal about what to do in the absence of intertheoretic comparisons) is an example of a structural account. I discussed accounts like these in the previous

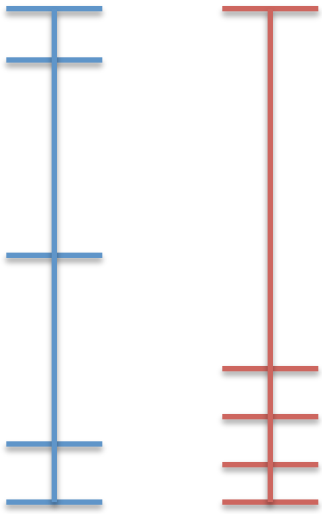
chapter. However, in that chapter I used them solely to provide a principled way of taking into account uncertainty over theories that are incomparable. In this chapter I consider whether structural accounts could be the correct accounts of what makes intertheoretic comparisons true. This is a much stronger claim than the one I discussed last chapter.

According to what I call *Common Ground* accounts, intertheoretic comparisons are true in virtue of different theories' having parts that are shared between them. Ross and Sepielli have proposed Common Ground accounts.[109]

According to *Universal Scale* accounts, intertheoretic comparisons are true in virtue of the fact that choice-worthiness scale is the same across different theories. Ross and Sepielli have both proposed Universal Scale accounts, in addition to the Common Ground accounts that they have proposed. I defend a different Universal Scale account later in this chapter.

We can represent these different accounts diagrammatically. Let us consider two theories, T_1 and T_2:

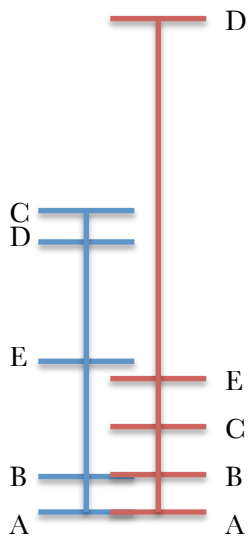
[109] (Ross 2006, 764–5; Sepielli 2009)



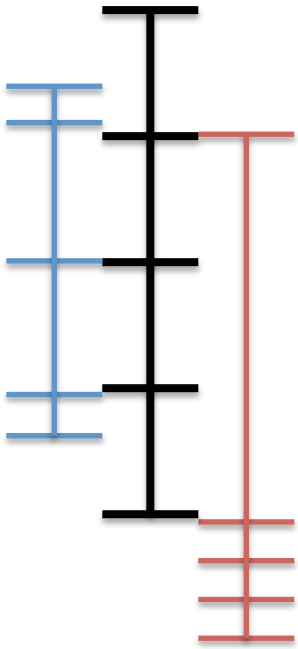
Structural accounts normalise with respect to some features of each theory's choice-worthiness function. In the diagram above, I have normalised them with respect to the range of the choice-worthiness function. The key question for Structural accounts is at which features of each theory's choice-worthiness function to normalise (such as the range, or the variance, or the maximum choice-worthiness minus the mean choice-worthiness).

Common Ground accounts attempt to find some choice-worthiness-differences between specific options that are agreed up by both theories. As opposed to Structural accounts, Common Grounds accounts require us to be able to identify options across theories (rather than merely identifying them by position in the choice-worthiness function). The key questions for Common Ground accounts are (i) to elucidate what it means for a theory to 'share parts' and (ii) to identify the options *A* and *B* whose choice-worthiness difference the two theories under consideration agree upon. In the diagram below, I

have supposed that the two theories agree on the choice-worthiness difference between A and B.



According to Universal Scale accounts, the two theories are already plotted on some shared scale, represented in black in the diagram below.



The key question for Universal Scale accounts is to explain the nature of this shared scale, and give reasons for thinking that this shared scale exists.

Intuitively, these three categories seem to me to be exhaustive (though not necessarily mutually exclusive), but I cannot prove that this is so. Any of these accounts may be understood as a purported solution to either the metaphysical problem or the epistemic problem. In what follows I will discuss accounts that have been proposed in the literature, understanding them as attempts to address the metaphysical problem. However, I will suggest, in section V, that the Common Ground accounts that have been proposed can be understood, in conjunction with my Universal Scale account, as an attempt to answer the epistemic problem. Let us now turn to accounts that have been proposed in the literature.

II. Against Structural Accounts

In the last chapter I discussed structural accounts as ways of taking into account uncertainty over theories that are cardinal but incomparable. But might a structural account actually be the best account of intertheoretic comparisons? This is one way of reading Lockhart, and has been endorsed by several people in conversation. However, I think that this proposal has severe weaknesses. In what follows, I make two arguments against this proposal: first, that it doesn't get the intuitions right; and, second, that it cannot make sense of *amplified* theories.

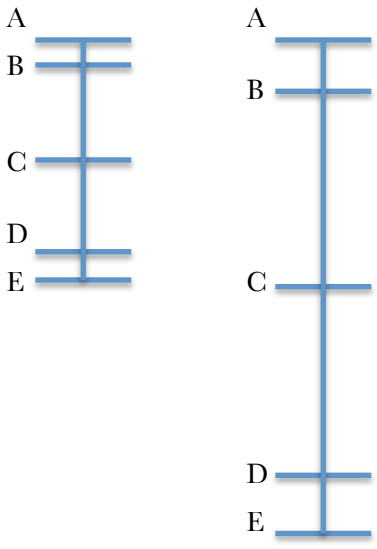
First, let us consider some core examples where we have intuitions about how theories compare. The cases of intertheoretic choice-worthiness comparisons where we have the most clear-cut intuitions are cases where the two theories are very similar, in that they agree in every respect except on the extension of the class of things that are fundamental bearers of value. Consider, for example, a utilitarian theory according to which only humans are of value, and second, a utilitarian theory according to which all sentient creatures are of value. The natural and obvious way to compare these two theories is to suppose that the value of humans is the same according to both theories. But structural accounts have to deny that: they have to claim that the value of humans is smaller according to the second theory than according to the former. They have to deny this because, if they are to satisfy the principle of equal say, they must claim that no theory can differ in the amount of value it posits in general, even though it seems natural to suppose that there is more value in general according to the theory that posits that there are a greater number of bearers of value.

Structural accounts also can't account for the intuitively plausible idea that there could be *low-stakes* theories: theories according to which decisions in general just aren't that important, either morally or prudentially. Suppose for example, that Meursault originally adheres to common-sense ethics, but then reads more and more nihilist literature. He becomes progressively convinced that nihilism is true. However, the way he becomes convinced is not that he increases his credence in nihilism and decreases his credence in common-sense ethics. Rather, he progressively realises that certain things he used to think were of value are not of value. First, he realises that art and literature and non-personal goods are of no value in itself. Then he realises that there are no other-regarding reasons, and retreats to egoism. At each step, Meursault becomes more despondent. Finally, he realises that even his own happiness is of no value, and he comes to accept full-blown nihilism.

The natural and intuitive way to understand this is that the ethical viewpoints that Meursault adheres to become progressively closer and closer to nihilism. Meursault progressively thinks that there is less and less value in the world, until eventually he thinks there is no value at all in the world. However, this is not the way that structuralist accounts would understand Meursault's progression in beliefs. According to structuralist accounts, when Meursault rejects the value of art, his other beliefs compensate and he comes to believe that personal moral reasons were much more important than he had previously thought; when Meursault rejects moral reasons, he must also come to believe that his own happiness is much more important than he had previously thought. The amount of value Meursault thinks exists, in general, is the same, right up until the point when he embraces nihilism. At that point, there is a stark

discontinuity, and suddenly Meursault thinks there is no value in the world at all. But that seems to mischaracterise what has happened. So, insofar as structural accounts misrepresent what is a perfectly normal and uncontroversial aspect of our moral lives – namely, coming to believe that there is more or less of value than one had previously thought – we have another reason against accepting structural accounts.

My second argument is that structural accounts can't account for a possible way in which two theories might be related. Introducing some new terminology, let us say that two theories T_i and T_j have the same *cardinal structure* iff for all *A*, $CW_i(A) = kCW_j(A) + c$, where $k > 0$. And let us say that T_i is an *amplified* version of T_j iff, for all *A*, $CW_i(A) = kCW_j(A) + c$, where $k > 1$. The following diagram (which is to scale) represents this idea:



Let's suppose that A-E are the only possible options. T_1 and T_2 agree that the difference between B and C is four times the difference between A and B. But the

difference between A and B, according to T_2, is twice the difference between A and B, according to T_1. So T_2 is an amplification of T_1.

With this on board, I can state my argument, as follows:

(P1) Amplified theories are possible.

(P2) If amplified theories are possible, then all structural account are false.

(C3) Therefore, all structural accounts are false.

(P2) is uncontroversial. If structural accounts are correct, we can only appeal to information concerning the theory's choice-worthiness function, which is unique only up to a positive affine transformation. So, on structural accounts, all theories with the same cardinal structure must agree on the ratios of choice-worthiness differences between options. So providing an example of two theories, one of which is an amplified version of the other, thereby shows that structural accounts are not correct. Here I provide an example of such a pair of theories:

Sophie's Change of View

Sophie initially believes in a partialist form of utilitarianism, which posits both impartial and agent-relative value. Though she thinks that human welfare is of value in and of itself, she also thinks that the presence of certain relationships between her and others confers additional value on those with whom she has the relationship. For that reason, she believes that the welfare of her friends and family is more valuable than that of distant strangers, though she thinks that both have value.

> Sophie then revises her belief, and comes to believe the idea that the welfare of all humans is of equal value. However, she realises that there are two ways in which she could come to hold this view. First, she could come to believe that there's no such thing as agent-relative value; no relationships confer additional value on the welfare of others. In which case the value of the welfare of distant strangers would be the same as she had previously thought, but the value of the welfare of close friends and family would be less than she had previously thought. Second, she could come to believe that, morally, she should 'be a brother to all', and she should regard her relationship with all other humans as being morally valuable in just the same way that she had thought that blood relationships and friendships were morally valuable. In which case, the welfare of her friends and family would be just as valuable as she had always thought; it's just that the value of the welfare of distant strangers is greater than she had thought. She is unsure which she should believe.

Let's call the first view that Sophie considers *Benthamite utilitarianism* and the second view *kinship utilitarianism*. Intuitively, it seems perfectly meaningful to think that Sophie could be uncertain between these two views. And it also seems meaningful for her to think that her relationships would have been downgraded in value, if Benthamite utilitarianism were true, but that the value of distant strangers would have increased in value, if kinship utilitarianism were true.

Moreover, in order to further explain the meaningfulness of amplified theories we can point to four distinctions between the two theories. I take these four differences to undermine one salient argument against the possibility of amplified theories: the

argument that it would *make no difference* whether or not one amplified theory or another were true.

First, Benthamite utilitarianism and kinship utilitarianism differ on the grounding of choice-worthiness: they disagree on facts concerning in virtue of what certain actions are wrong. Benthamite utilitarianism would claim that saving a human life is good because saving that life would increase the sum total of human welfare. On Benthamite utilitarianism, there is just one fact in virtue of which saving a human life is a good thing. In contrast, kinship utilitarianism would claim that saving a human life is good both because saving that life would increase the sum total of human welfare, and also because one has a certain sort of relationship to that person. On kinship utilitarianism, there are two facts in virtue of which saving a human life is a good thing. That is, Benthamite utilitarianism and kinship utilitarianism disagree on what the *right-makers* are.

In general, there is often more to a moral theory than a choice-worthiness function: there is also a metaphysical account of why that choice-worthiness function is correct. This shows that the argument for intertheoretic comparability with which I opened section I of chapter three, which appealed to the idea that there is nothing more to a normative theory than its choice-worthiness function, was mistaken. Theories differ in their metaphysics, and, intuitively, that metaphysical account can make a difference to the amplification of a theory. On Benthamite utilitarianism, one does *one* wrong thing by killing another person (namely, reducing the amount of welfare in the world), whereas, on kinship utilitarianism, one does *two* wrong things (one reduces the amount of welfare in the world, and one violates an obligation that arises out of a special relationship that one has). Committing both wrong X and wrong Y is worse than

committing just wrong X. So it's more wrong to kill, according to kinship utilitarianism, than it is according to Benthamite utilitarianism.

Second, it seems plausible that which attitudes it is fitting for Sophie to have, given revision of her initial belief, depends on which amplification of utilitarianism she comes to believe. If she comes to believe Benthamite utilitarianism, it seems fitting for her to be disappointed: she has lost something of value, as her friends and family are merely as valuable as distant strangers. In contrast, the same is not true if she comes to believe kinship utilitarianism. Perhaps, instead, it would be fitting for her to feel a sense of wonder and new connectedness with those whom she doesn't know.[110]

Third, it seems plausible to me that the epistemological facts can differ depending on which theory we are discussing, and that they can differ in virtue of the amplification of the theory. Perhaps the idea of downgrading the value of her friends and family seems abhorrent to her; or perhaps she finds the idea that certain relationships should confer additional value on welfare metaphysically spooky. Either of those views seem reasonable, and either one would mean that she'd find one of the two theories more plausible than the other.

Fourth, facts about what it's appropriate to do under normative uncertainty can differ depending on which amplification of utilitarianism Sophie has credence in. If she has 20% credence in kinship utilitarianism and 80% credence in non-consequentialism then if she follows MEC she will more often act in accordance with utilitarianism than if she

[110] Note that I use the term 'fitting' rather than 'ought'. Utilitarianism rejects the idea that the fittingness of certain reactive attitudes is normatively relevant. But that doesn't mean that there aren't facts about which attitudes are fitting. Analogously, one might reject the idea that the requirements of etiquette are normatively relevant while still acknowledging that it's against the requirements of etiquette to eat with one's elbows on the table.

has 20% credence in Benthamite utilitarianism and 80% credence in non-consequentialism. This is because things are higher-stakes in general for kinship utilitarianism than for Benthamite utilitarianism.

One might complain that I have only given one example, and that we shouldn't trust our intuitions if they pertain to merely one case. But I could give more examples. Consider Thomas, who initially believes that human welfare is ten times as valuable as animal welfare, because humans have rationality and sentience, whereas animals merely have sentience. He revises this view, and comes to believe that human welfare is as valuable as animal welfare. He might now think that human welfare is less valuable than he previously thought because he has rejected the idea that rationality confers additional value on welfare. Or he might now think that animal welfare is more valuable than he previously thought, because he has extended his concept of rationality, and thinks that animals are rational in the morally relevant sense. Or consider Ursula, who initially believes that wrong acts are ten times as wrong as wrong omissions, but then comes to believe that acts and omissions are on a par. Does she come to believe that wrong omissions are worse than she had thought, or does she come to believe that wrong acts aren't as wrong as she had thought? If the former, then it might be fitting for her to feel horror at the idea that, insofar as she had let others die, she had been doing things as bad as murder all her life. If the latter, then it might be fitting for her to feel less blame towards those who had killed others. In exactly the same way as with Sophie, we can explain the distinction between these pairs of amplified theories by looking at differences in rightmakers, differences in fitting attitudes, differences in epistemological reasons, and differences in facts about what it is appropriate to do.

So we should reject structural accounts. Let's see if other accounts can do better.

IV. Against two Common Ground Accounts

A Common Ground account is suggested by both Ross and Sepielli.[111] The idea is to look at "cases in which, for some pair of options, we know that the difference between their values is the same according to both ethical theories."[112] We then can use that difference to define one unit of choice-worthiness that is comparable across both theories.

The trouble with this account is that neither Ross nor Sepielli give an explanation of what it is for some choice-worthiness difference to be 'shared' between two options. Sepielli is clearest: he takes agreement between theories to consist in the fact that two theories agree where some part of their choice-worthiness functions have the same cardinal structure. More precisely, Sepielli's view is as follows. For some three particular options A, B, and C:

$$\text{If } \frac{\mathrm{CW}_i(\mathrm{A}) - \mathrm{CW}_i(\mathrm{B})}{\mathrm{CW}_i(\mathrm{B}) - \mathrm{CW}_i(\mathrm{C})} = \frac{\mathrm{CW}_j(\mathrm{A}) - \mathrm{CW}_j(\mathrm{B})}{\mathrm{CW}_j(\mathrm{B}) - \mathrm{CW}_j(\mathrm{C})} \text{ then}$$

$$\mathrm{CW}_i(\mathrm{A}) - \mathrm{CW}_i(\mathrm{B}) = \mathrm{CW}_j(\mathrm{A}) - \mathrm{CW}_j(\mathrm{B}).$$

[111] (Ross 2006, 764–5; Sepielli 2009)
[112] (Ross 2006, 764)

But this account is internally inconsistent.[113] Consider the example of utilitarianism and square-root prioritarianism that I gave in the previous chapter.

> *Utilitarianism and Prioritarianism - I*
>
> Annie and Betty have both lived for 16 years, and will live for a further 9 years if a certain drug is given to them. The decision-maker's options are as follows:
>
> A: Save both Annie and Betty
>
> B: Save Annie only
>
> C: Save Betty only
>
> D: Save neither of them

Both utilitarianism and prioritarianism agree that the ratio of differences in choice-worthiness between A and B and B and D are the same. (They both agree that the difference in choice-worthiness between A and B is the same as the difference in choice-worthiness between B and D). So, according to this Common Ground view, the difference in choice-worthiness between A and B, on utilitarianism, is the same as the difference in choice-worthiness between A and B, on prioritarianism. According to the prioritarian, the choice-worthiness difference between A and B is $\sqrt{25} - \sqrt{16}$, which equals 1. According to the utilitarian, the difference is 25 – 16, which equals 9. So, according to Ross and Sepielli's view, 1 unit of choice-worthiness on prioritiarianism

[113] This fact was first noticed by Toby Ord, though the example is my own. Sepielli recants this view, because of this objection, in (Sepielli 2010).

equals 9 units of choice-worthiness on utilitarianism. But now consider a variant on the above case:

Utilitarianism and Prioritarianism - I

Charlotte and Doreen have both lived for 64 years. They each will live for a further 9 years if a certain drug is given to them. The decision-maker's options are as follows:

E: Save both Charlotte and Doreen

F: Save Charlotte only

G: Save Doreen only

H: Save neither of them

Both utilitarianism and prioritarianism agree that the ratio of differences in choice-worthiness between E and F and F and H are the same. (They both agree that the difference in choice-worthiness between E and F is the same as the difference in choice-worthiness between F and H). So, according to this Common Ground view, the difference in choice-worthiness between E and F, on utilitarianism, is the same as the difference in choice-worthiness between E and F, on prioritarianism. According to the prioritarian, the choice-worthiness difference between A and B is $\sqrt{73} - \sqrt{64}$, which equals ~0.5. According to the utilitarian, the difference is 74 – 64, which equals 9. So, according to Ross and Sepielli's Common Ground view, 1 unit of choice-worthiness on prioritiarianism equals approximately 18 units of choice-worthiness on utilitarianism. But this is different from what we concluded with respect to Annie and Betty. So Ross

and Sepielli's account is generates inconsistent pronouncements about how choice-worthiness compares across two theories. So their account should be rejected.

Another Common Ground account that Sepielli briefly suggests[114] is that there might be 'paradigms' of morally okay actions, and paradigms of morally heinous actions, which are definitive of choice-worthiness. So just as one might think that the Standard Kilogram defines what it means to have 1kg of mass, so one might think that the difference in choice-worthiness between some two particular, extremely well-specified options (listening to music, and killing for fun, for example), defines one unit of choice-worthiness. So all moral theories must agree that that unit is the same.

The problem with this account is just that there is far too much disagreement among moral theories for this to be a plausible general view. According to ethical egoism, the difference in choice-worthiness between listening to music and killing for fun will very different compared to the difference in choice-worthiness between listening to music and killing for fun, according to utilitarianism. And the same will be true for any pair of options.

III. Against two Universal Scale Accounts

The discussion of amplified theories made some suggestions about ways in which we can tell the difference between two theories with the same cardinal structure. This idea motivates some different accounts of intertheoretic comparisons.

[114] (Sepielli 2010, 186)

I mentioned that the amplification of a theory can make a difference to facts concerning what it's appropriate to do under normative uncertainty. So perhaps it's those facts that make it the case that a certain intertheoretic comparison holds. This is a view suggested by Ross.[115]

As I understand this suggestion, the claim is that facts about how choice-worthiness differences compare across theories are determined by facts about what it is appropriate to do in light of uncertainty between those theories. If I face options A and B, and have 10% credence in T_1, according to which $CW_1(A) > CW_1(B)$, and 90% credence in T_2, according to which $CW_2(B) > CW_2(A)$, and it is appropriate for me to do A, then, *because* it is appropriate to do A in this situation, $(CW_1(A) - CW_1(B))$ is at least 9 times greater than $(CW_2(B) - CW_2(A))$.

The obvious objection to this account is that it puts the cart before the horse. Consider Kate, who has 80% credence in common-sense views about how she should spend her money, and 20% credence in Singer's view that she has strong obligations to donate much of her money to alleviate extreme poverty. In this case, intuitively it's appropriate for her to donate the money. But we have that intuition because it seems clear how the choice-worthiness differences compare across the two normative views in which she has credence. It's not that we have the intuition that it's appropriate for Kate to donate part of her income, and thereby infer what the respective choice-worthiness differences between the common-sense view and Singer's view are. Ross's proposal therefore seems to get the order of explanation the wrong way around.

[115] (Ross 2006, 763). It is also endorsed by (Riedener 2013), and by John Broome in conversation.

A different sort of meta-scale account is suggested by Sepielli.[116] He wishes to use degrees of blameworthiness as the scale by which choice-worthiness difference may be compared. The exact nature of his proposal is unclear. But I think that his principal initial proposal is that a decision-maker believes that $(CW_i(A) - CW_i(B)) = (CW_j(C) - CW_j(D))$ iff the strength of the decision-maker's disposition to blame for doing *A* rather than *B*, conditional on T_i, is the same as the strength of the decision-maker's disposition to blame for doing *C* rather than *D*, conditional on T_j. It should be fairly clear that this isn't the right account. The decision-maker might just have the sort of personality where she wouldn't be terribly disposed to blame, if some very demanding normative theories were true. Or it might be that she would be deeply depressed, if one particular theory were true, and therefore her dispositions to do anything would be weaker than they ordinarily are. But these factors don't seem to affect how choice-worthiness differences compare across the different theories in which she has credence.

One might try to tweak Sepielli's account by claiming that choice-worthiness differences are measured by how one disposed to blame one *ought* to be. But that account would suffer from problems as well. On utilitarianism, how disposed to blame one *ought* to be is not perfectly correlated (indeed, is sometimes highly uncorrelated) with the degree of wrongness of a particular action. So this account would misrepresent choice-worthiness differences according to utilitarianism.

Instead, the best account in this area, I think, is that choice-worthiness differences are measured by the degree to which is it *fitting* to blame for a certain action (or, as I will use the term, the degree to which an action is *blameworthy*). More precisely: $(CW_i(A) -$

[116] (Sepielli 2010, 183ff.)

$CW_i(B)) = (CW_j(C) - CW_j(D))$ iff the blameworthiness of the decision-maker for doing *B* rather than *A*, conditional on T_i, is the same as blameworthiness of decision-maker for doing *D* rather than *A*, conditional on T_j. The idea behind this account is that facts about fitting attitudes are constitutive of the nature of concepts like 'ought' and 'choice-worthy'; they are therefore not something about which different moral theories can disagree.

I think that this account has at least something going for it: in my discussion of amplified theories, I suggested that there is a link between the amplification of a theory and which attitudes it is fitting to have. The principal question, again, however, is whether choice-worthiness differences should be explained in terms of fitting attitudes, or the other way around. And this fitting-attitude account suffers from the following problem, which is that it cannot explain where cardinally measurable degrees of blameworthiness come from.[117] It cannot, for example, use probabilities to provide the cardinal measure. To do so would require making claims such as: "S is equally blameworthy for choosing (i) A and the guarantee that T_1 is true, as she is for choosing (ii) a 50% probability of B and T_2 being true, and a 50% probability of C and T_2 being true." (which would show that the difference in choice-worthiness between T_2(B) and T_1(A) is the same as the difference between T_1(A) and T_2(C)). But if 'probability' in that sentence means objective chance, then it doesn't make any sense, because there can't be objective chances about which theories are true (except 1 and 0). If 'probability' means either 'subjective credence' or 'rational credence', then the account becomes extremely similar to Ross's "facts about appropriateness" account, which, as we saw, got the order

[117] I thank (Riedener 2013) for this point.

of explanation the wrong way around. So I don't think that this account is satisfactory, either. Let's turn to a different account that I think does better.

V. A Universal Scale Account

I believe that we can make progress on understanding intertheoretic comparisons by learning from work that has been done in the literature on the metaphysics of quantity. This debate around the metaphysics of quantity addresses questions such as: "In virtue of what is this object more massive than this other object?" or "In virtue of what is it true that this object is 2kg and this object is 4kg?"

There are two classes of answers. *Comparativists* answer that it is the mass-relations ("*x* is more massive than *y*") that are fundamental, and claims about intrinsic mass properties ("*x* is 4kg") are grounded in mass-relations. *Absolutists* answer that it is the intrinsic mass properties of objects that ground mass-relations. For absolutists, the fact that *x* is heavier than *y* is true in virtue of facts about the intrinsic properties of the objects themselves; for comparativists, it is the other way around.

Though work on the metaphysics of quantity has, so far, entirely focused on scientific quantities ('mass', 'size', 'temperature', etc), we can ask just the same questions about the metaphysics of quantities of value, or of choice-worthiness. We can ask: If it is true that the difference in choice-worthiness between A and B is twice as great as the difference in choice-worthiness between B and C, is that true in virtue of the fact that A, B and C each have an intrinsic property of a certain degree of choice-worthiness? Or is the metaphysical explanation the other way around? Moreover, in the same way as the

possibility of amplified theories is a crucial issue in the debate concerning intertheoretic comparisons, the possibility of a world in which everything is identical except insofar as everything is twice as massive is a crucial issue in the debate between absolutists and comparativists.

Within the metaphysics of quantities literature, it is generally recognised that absolutism is the more intuitive position. Yet it seems to me that all the discussion of intertheoretic comparisons so far has assumed comparativism about quantities of value or choice-worthiness. If we reject that assumption, then we can provide a compelling metaphysical account of intertheoretic comparisons. In what follows, I'll first present comparativism about mass, then present Mundy's elegant absolutist account of mass, then explain how something like this account could be applied to value and choice-worthiness.

The standard comparativist account of mass is to analyse mass in terms of the relation "*x* is more massive than *y*", and the concatenation operator "*x* and *y* are together equally as massive as *z*". Three things are important to note about standard comparativist accounts. First, the account is first-order: the variables, *x*, *y*, and *z* are variables over *objects* (rather than over properties, which would make the account second-order). For this reason, the account is nominalist: it gives an account of mass without any reference to the properties of objects. And, third, the account is empiricist: attempting to give an analysis of mass solely in terms of observable mass-relations. (So, for example, both '*x* is more massive than *y*' and '*x* and *y* are equally as massive as *z*' can be defined operationally, identifying them with the behaviour of those objects on scales: *x* is more massive than *y* iff, when *x* and *y* are placed on opposite sides of the scale,

the scale will tip in *x*'s direction; *x* and *y* are together equally as massive as *z* iff, when *x* and *y* are placed on one side of the scale, and *z* on the other side, then the scale will not tip in ether direction). Using those two relations, and several axioms,[118] it can be shown that "*x* is more massive than *y*" relation can be represented using numbers, where M(*x*) > M(*y*) iff *x* is more massive than *y*, where the numerical representation is unique up to a similarity transformation ($f(x) = kx$).[119]

In contrast, Mundy's account is second-order, defined over properties as well as objects. Letting *X* refer to the mass of *x* and *Y* refer to the mass of *y* (etc), the fundamental mass relations, on Mundy's account, are "*X* is greater than *Y*" and "*X* and *Y* are equal to *Z*". That is, the fundamental mass-relations are defined over the mass-properties of objects, rather than over those objects themselves. It is therefore clearly realist rather than nominalist: it posits the existence of properties (which are abstract entities), over and above the existence of objects. And it is Platonist rather than empiricist, because properties are abstract entities that can exist without being instantiated. Using this framework, Mundy is able to give a full formal account of quantities of mass; he then argues that there are significant *empirical* reasons for preferring it to the traditional, first-order, comparativist accounts. In particular: the traditional comparativist account of

[118] I'll use the axiomatisation given in (Suppes and Zinnes 1963). Let A be the set of all objects, 'Rxy' mean 'x is either less massive or equally as massive as y'. Let $x^\circ y$ refer to a binary operation from A x A to A: the 'concatenation 'of x and y (where 'concatenation' of x and y may be defined as, for example, placing x and y on the same side of a scale). The axioms are as follows:
1. Transitivity: if Rxy and Ryz, then Rxz.
2. Associativity: $(x^\circ y)^\circ cRx^\circ(y^\circ c)$
3. Monotonicity: If Rxy then $R(x^\circ z)(z^\circ y)$
4. Restricted Solvability: If not Rxy, then there is a z such that $Rx(y^\circ z)$ and $R(y^\circ z)x$
5. Positivity: Not $x^\circ yRx$
6. Archimidean: If Rxy, then there is a number n such that $Ry(nx)$ where the notation (nx) is defined recursively as follows: $1x = x$ and $nx = (n-1)x^\circ x$
As Suppes and Zines note, axiom 5 in conjunction with the order properties of *R* and the definition of ° imply that the set A is infinite.
[119] See (Krantz et al. 1971).

mass needs to assume that, for any two objects, there is an actual third object that is equal in mass to those two objects. But the universe may well be finite, and if so then this assumption would be false. But it seems very plausible that objects have mass-quantities *whether or not* the universe is finite.

There is considerable debate between absolutists and comparativists. The key issue, however, when it comes to quantities of value or of choice-worthiness, is that absolutism about quantities of choice-worthiness can neatly solve the problem of intertheoretic choice-worthiness comparisons.

Consider the issue of whether there could be a world w_1, where all the relations between objects are the same as in world w_2, but where all objects are twice as massive in w_1 as they are in w_2. It is generally regarded as a problem for comparativism that it cannot make sense of the idea that w_1 and w_2 could be distinct worlds: the mass-relations between all objects in w_1 are the same as the mass-relations in w_2, so, according to comparativism, there is no difference between those two worlds. In contrast, absolutism is able to explain how those two worlds are distinct. Properties necessarily exist; so the two worlds differ in the intrinsic properties that objects in those two worlds instantiate. Note, also, that, if w_1 and w_2 are distinct worlds, then we have conclusive evidence for the existence of inter-world mass relations: we can say that object *x* in w_1 is twice as massive as it is in w_2.

Similarly, now, consider the issue of whether there could be two theories T_1 and T_2, where T_1 which has the same cardinal structure as T_2, but where the choice-worthiness differences between all options are twice as great on T_1 as they are on T_2. In my

argument against structural accounts of intertheoretic comparisons, I argued that this is a genuine possibility. But, if so, then we have a good argument against comparativism about choice-worthiness, according to which the only fundamental facts about choice-worthiness are facts about choice-worthiness relations between options. (One could try to explicate this idea in comparativist terms using Ross's Universal Scale account; but we saw that that account was unsatisfactory, getting the order of explanation the wrong way around.) In contrast, if we endorse absolutism about choice-worthiness, then we have an explanation of how T_1 and T_2 could be distinct theories. The same choice-worthiness quantities exist in many different epistemically possible worlds, so we can use them as the measuring rod to compare the choice-worthiness of *A* in the world in which T_1 and the choice-worthiness of *A* in the world in which T_2 is true. Moreover, we have an answer to the question of grounds: the choice-worthiness difference between *A* and *B* on T_1 is different from the difference in choice-worthiness between *A* and *B* on T_2 in virtue of the fact that A and B instantiate different intrinsic choice-worthiness quantities in the world in which T_1 is true than in the world in which T_2 is true.

In general, if $(CW_i(A) - CW_i(B)) = (CW_j(C) - CW_j(D)$ is true, then it is true in virtue of the fact that the difference in the magnitude of the property of choice-worthiness that *A* instantiates and the magnitude of the property of choice-worthiness that *B* instantiates in the epistemically possible world in which T_i is true, is the same as the difference in the magnitude of the property of choice-worthiness that *C* instantiates and the magnitude of the property of choice-worthiness that *D* instantiates in the epistemically possible world in which T_j is true. In fact, as long as we know to take the following second-order claim at face-value, rather than analyse it in comparativist terms, we can

state this claim in very natural language, namely: if $(CW_i(A) - CW_i(B)) = (CW_j(C) - CW_j(D)$ is true, then it is true in virtue of the fact that the difference between the choice-worthiness of *A* and the choice-worthiness of *B,* in the epistemically possible world in which T_i is true, is the same as the difference between the choice-worthiness of *C* and the choice-worthiness of *D* in the epistemically possible world in which T_j is true.

Absolutism about choice-worthiness is something of a flat-footed response to the problem of intertheoretic comparisons. It takes statements about choice-worthiness at face value: as ascribing an intrinsic property to an option. And once we allow the existence of necessarily existent choice-worthiness properties, then we have the resources to explain how intertheoretic comparisons are possible. The absolutist about choice-worthiness mimics the absolutist about mass in this respect: the absolutist about mass takes statements about the mass of objects at face value (as ascribing an intrinsic property of mass to an object), and then uses this to explain how inter-world mass relations are possible (as in the mass-doubled world case).

One objection to this account is as follows. An absolutist about mass needs only to defend the idea that quantitative mass-properties exist across all *metaphysically* possible worlds. In contrast, in order to solve the problem of intertheoretic comparisons, I must say that quantitative choice-worthiness properties exist across all *epistemically* possible worlds. But that's a very strong claim — in fact, too strong to be plausible. Whatever the merits of absolutism about choice-worthiness, one should not be certain in that view.

In response to this objection, I agree with its main point. Even though I think that absolutism about choice-worthiness is the most plausible view, I do not think that one

should be certain in that view. And, insofar as I've successfully argued against the other accounts of intertheoretic comparisons that are the table, it's plausible to me that, conditional on comparativist about choice-worthiness being true, one should think that different normative theories genuinely are incomparable with each other (though of course one should not be certain in this, either.)

So my all-things-considered view is that the decision-maker should 'divide and conquer'. The majority of her credence should go to absolutism about choice-worthiness being true. And, conditional on absolutism, she can make intertheoretic comparisons and apply MEC when making decisions. But she should still have some credence that absolutism about choice-worthiness is not true. In which case, she should believe that different normative theories are incomparable, and she should use Variance Voting to take into account her normative uncertainty. In effect, for every pair of theories, she will have some credence that they are comparable in such-and-such a way, and some credence that they are incomparable. In which case, her uncertainty should be taken into account in the manner prescribed by the Hybrid View, as I argued at the end of the previous chapter.[120]

Before concluding this chapter, I will make two further comments on this account. The first comment is that it is not ad hoc to side with absolutism about choice-worthiness, rather than comparativism. This is for two reasons. First, I argued above that amplified theories are possible and that, in general, intertheoretic comparisons are clearly possible. Insofar as absolutism can give a natural and plausible explanation of that, whereas

[120] There might be other cases where a decision-maker believes that two theories are comparable, but simply has no idea about how they compare. In such a case, it might be that the decision-maker should use Variance Voting across those two theories as well.

comparativism seemingly cannot, we have reason to prefer absolutism about choice-worthiness. Second, the principal reason for rejecting absolutism about quantities of mass (and other scientific quantities) is a worry about needing to posit abstract entities such as properties in one's ontology, and the desire to posit the existence only of relations that are observable. Whether or not this argument is successful in general, it is considerably weaker in the case at hand. If we are assuming normative realism, then though not an inconsistent combination of views, it certainly seems like an *odd* combination of views to be happy with the existence of normative facts, but be sceptical of the existence of properties.

The second comment is that this account gives us the resources to at least partially answer the epistemic question. Given absolutism about choice-worthiness, for any theory T_i and for any real number n we can make sense of another theory T_j whose choice-worthiness function is n times that of theory T_i. That is: every possible amplification of T_i is itself another theory. So when we ask: "How, if at all, do utilitarianism and this rights-based non-consequentialist theory compare?" we're really asking which, of the infinitely many different theories that have the same cardinal structure as utilitarianism, and which, of the infinitely many different rights-based non-consequentialist theories, should we have most credence in?

To take an earlier example, consider Sophie, who initially believed partialist utilitarianism, but then became unsure between that view and the view according to which all persons have equal moral weight. The question about how to make intertheoretic comparisons between those two views reduced to the question of which, of all infinitely many theories within the class of classical utilitarian theories (including

what I called kinship utilitarianism and Benthamite utilitarianism) she should come to have credence in. If she was moved to classical utilitarianism because it is a simpler theory, then it seems plausible that she should come to have credence in Benthamite utilitarianism. If she was moved to classical utilitarianism by reflecting on the fact that there is a deep arbitrariness in who she has special relationships with, then it seems plausible that she should come to have credence in kinship utilitarianism. Either way, we can explain why, as is intuitive, she should come to have credence in one of those theories, but not a different theory (according to which, perhaps, the value of distant strangers' welfare is one million times as great as it is on the partialist theory). Basic epistemic conservatism suggests that she should alter her beliefs as little as possible in order to accommodate new evidence (in this case, new arguments). Having partial belief in partialist utilitarianism, and partial belief in anything other than kinship or Benthamite utilitarianism, would be oddly incoherent.[121]

If the account I have given is correct, this is an exciting development for first-order normative ethics. Moral theories, when they have been given, have really been *classes* of moral theories. And different views within this class can me more or less plausible than other views within this class. So there may be scope to revisit old ethical theories, and assess which specific versions of those theories are most plausible.[122]

---

[121] This account also, I believe, explains what Ross and Sepielli were getting at when they suggested that that parts of two different theories can 'agree' with each other. The agreement consists in the fact that the options instantiate the same choice-worthiness properties.

[122] For example, Frances Kamm (1992) and Thomas Nagel (2008) claim that utilitarianism is implausible because it does not posit the existence of rights, and therefore means that humans do not possess the value of dignity that can only be conferred by the possession of rights. But Kamm and Nagel do not distinguish between two different versions of utilitarianism. According to the first, no-one has any rights, and so humans are indeed of less value. According to the second, people do have rights not to be killed (for example), but they also have equally strong rights to be saved. Both have the same cardinal structure.

Conclusion

In this chapter I have providing a way of categorising different accounts of intertheoretic comparisons of value. I argued against the whole class of Structural accounts, and against the Common Ground and Meta-Scale accounts that have been defended in the literature. I then argued that invoking absolutism about choice-worthiness is the best way to answer the problem.

This concludes the first part of this thesis. I have defended a general metanormative theory, and elucidated the nature of intertheoretic comparisons of choice-worthiness. Now let us turn to some further issues that are affected by consideration of normative uncertainty.

But, according to the latter for of utilitarianism, humans *do* have the value of dignity that can only be conferred by having rights. So Kamm and Nagel's argument would not go through.

Part II: Further Issues

Chapter 5: Infectious Incomparability

Introduction

In this chapter, I'll introduce a new problem for MEC: what I call the 'infectious incomparability' problem.[123]

So far, I have been assuming that the decision-maker only has positive credence in theories whose choice-worthiness ordering is complete. But that's clearly a false assumption for almost all real-life decision-makers. Almost all plausible moral theories give an incomplete choice-worthiness ordering, whether because they think the choice-worthiness relation is vague, or because they think exists incomparability between different sorts of value. In this chapter, I'll argue that the fact that one should have positive credence in theories whose choice-worthiness ordering is radically incomplete poses a serious problem for accounts of decision-making under moral uncertainty, but that it can be overcome.

I begin by considering Jacob Ross's argument that it is appropriate to 'reject' nihilism, and arguing that it fails. In section I, I define nihilism and show why 'expected choice-worthiness' reasoning can't be used to reject nihilism; this also serves to introduce the infectiousness problem. In section II, I consider Ross's dominance principle, which he uses to argue that we can reject nihilism. I give two arguments to show why Ross's principle is false. In section III, I consider two responses that Ross has made in

[123] Sections I-IV of this chapter are very closely based on (MacAskill 2013).

correspondence, and argue that they fail. In section IV, I consider a rejoinder, but argue that that rejoinder fails.

In sections V and VI, I look at whether a fuller account of how one ought to act in the face of incomparability. I argue that there is a solution, but one that is only available to a certain class of theories about how to respond to incomparable values.

I. Jacob Ross on nihilism

In 'Rejecting Ethical Deflationism' Jacob Ross argues that a decision-maker is permitted to *reject* certain moral theories: that is, assume that such theories are false for the purposes of practical reasoning.[124] Consider, for example, a 'uniform' theory, according to which all options are equally choice-worthy, and suppose that a decision-maker is less than certain in such a theory. Every time that such a decision-maker encounters a decision-situation, the portion of her credence devoted to this uniform theory will give no reasons in favour of any option; whereas the portion of her credence in positive moral theories will give reasons in favour of some options over others. In every decision-situation, therefore, the same actions will be appropriate and inappropriate as if she had zero credence in the uniform theory. So the decision-maker need never worry about the portion of her credence devoted this uniform theory.

What's striking about Ross's conclusion is that it holds no matter how confident you are in that uniform theory, as long as that confidence doesn't amount to full certainty.

[124] (Ross 2006)

Even if you are 99.9% certain that the uniform theory is correct, you should still reject that theory and act in the way determined by the remaining 0.1% of your credence.

Ross extends his argument to other theories. He discusses certain sorts of relativistic theories, and certain 'relatively deflationary' theories, and argues that they can be rejected too. I will not focus on these arguments. Rather, I will focus on his argument that one may reject nihilism. Insofar as nihilism is a widely held view, Ross's argument that we can reject nihilism seems particularly important.

Ross defines nihilism as:[125]

> the view that the notions of good and bad and of right and wrong are illusions and that, objectively speaking, no option or state of affairs is better than any other, nor are any two options or states of affairs equally good. Thus, while uniform theories assign the same value to all of our options, nihilistic theories don't assign values to any of our options.

Notice that nihilism, on this definition, is very different from the 'uniform' theory mentioned above. Ross states the views in terms of *value*, but, in accordance with my framework, I will consider the views as applied to choice-worthiness; nothing will hang on this. According to the uniform theory, a positive choice-worthiness relation obtains between all options: every option is equally as choice-worthy as every other option. According to nihilism, *no* positive choice-worthiness relation obtains between any two options. That is, the choice-worthiness of every option is undefined.

However, if the choice-worthiness of every option is undefined, according to nihilism, then, for any decision-maker with non-zero credence in nihilism, there's a big problem

[125] (Ross 2006, 748). He notes (p.749 fn.3) that his argument only applies (in my terminology) to nihilism about first-order normative claims. A more thoroughgoing form of nihilism would hold that, as well as there being no facts about first-order normative claims, there are also no facts about what it is *appropriate* to do. It's highly unclear how one ought to take uncertainty over that sort of theory into account.

for her if she attempts to incorporate moral uncertainty into her reasoning about expected choice-worthiness. The problem is as follows. If we try to take an expectation over possible theories, and the choice-worthiness of an option is undefined according to one theory in which the decision-maker has non-zero credence, then the expectation as a whole is undefined. Because, according to nihilism, the choice-worthiness of *every* option is undefined, for a decision-maker with non-zero credence in nihilism, the expected choice-worthiness of every option is undefined, too. Non-zero credence in nihilism is therefore sufficient to infect practical reason, resulting in there being no facts about which options are more appropriate than which other options. [126]

This is pretty terrifying. Nihilism isn't just a speculative hypothesis. It's a view that's defended by many intelligent philosophers. It's difficult to see how it could be logically false, or somehow conceptually confused. So it seems exceedingly plausible that we ought to have non-zero credence in the view. But if so, then, for each and every one of us, the expected choice-worthiness of all options is undefined.

II. Using dominance reasoning to reject nihilism

Ross, however, doesn't use expected value reasoning in order to argue that one ought to reject nihilism. Rather, he uses dominance reasoning. He supposes, as an illustration, that he has credence in two theories only: '$\mathcal{T}_L$', which is a positive moral view, and '$\mathcal{T}_n$', a nihilist theory. He argues:[127]

[126] This 'infectious incomparability' problem is formally very similar to what Broome calls 'greedy incomparability' (Broome 2004, 169–171).
[127] (Ross 2006, 748)

> According to T_L, it would be better for me to send the trolley to the left than to send it to the right. And so my credence in T_L gives me *pro tanto* subjective reason to send the trolley to the left. The only way this could fail to be the most rational option would be if my credence in T_n gave me a sufficiently strong countervailing subjective reason to send the trolley to the right. But T_n implies that there would be nothing valuable or disvaluable about either alternative. And so my credence in T_n gives me no subjective reason to favor either alternative. Hence the *pro tanto* subjective reason to send the trolley to the left is unopposed, and so this is the rational option.

That is, Ross appeals to something like the following principle (again, translating his claims into my terminology):

> *Dominance over Theories (DoT):* If some theories in which you have credence claim that A is more choice-worthy than B, and no theories in which you have credence claim that B is more choice-worthy than A, then it is appropriate to choose A over B.[128]

Note that DoT is not a strong enough principle to warrant Ross's conclusion that one can always reject nihilism. It merely warrants the rejection of nihilism in cases where one has credence in only one theory other than nihilism, and only a small number of cases in which one has credence in more than one non-nihilist theory. We could easily formulate a stronger principle, which would give Ross the result he wants. But I'll focus on this weaker principle for ease of exposition; all the same considerations I raise would apply *mutatis mutandis* to the stronger principle. I'll give two arguments to show why DoT is false.[129]

---

[128] Ross endorsed a very similar principle in correspondence, but then after discussion suggested a modified principle, which is discussed in section III.

[129] These arguments assume that rational choice should not be cyclical. This assumption is discussed in section III.

Argument 1. DoT conflicts with more plausible metanormative principles

Suppose, for example, that the decision-maker has credence in the following two theories only. First, a specific person-affecting view of population ethics. The basic idea behind this view is that you can only have reasons to make people better off (or to avoid making people worse off), and that one cannot be made better off by being brought into existence, so there is no reason *simpliciter* to bring someone into existence, even if that person would have a happy life. According to this view, population X is better than population Y iff the two populations have exactly the same people in them and there is a greater average wellbeing in population X than in population Y. If the populations do not have exactly the same people in them, then it is undefined whether X is better than Y (or, in other words, X is incomparable in value to Y). If X and Y are entirely distinct populations, then there is not even a *pro tanto* reason to bring about X over Y, or vice versa, because there's no-one for whom X is better than Y, or for whom Y is better than X.

The second theory in which the decision-maker has credence is egalitarian utilitarianism, which places value on both average wellbeing, and on the equality of the distribution of that wellbeing, and according to which these two values can be weighed against each other. The theory gives a complete ordering of populations in terms of choice-worthiness.

Now consider three populations one could bring about: populations A, B and C. A and C have the same people in them. C has a much greater average wellbeing than A, but the wellbeing in A is distributed much more equally. Population B is larger in size than populations A and C, and no person in population B exists in population A or C.

Population B has the same average wellbeing as A, but a slightly less egalitarian distribution of wellbeing. According to egalitarian utilitarianism, A is more choice-worthy than B, which is more choice-worthy than C (i.e. A≻B≻C). According to the person-affecting view, A and B are incomparable, B and C are incomparable, and C is more choice-worthy than A (i.e. A⋡B and B⋡A; B⋡C and C⋡B; C≻A).

Now let's suppose the decision-maker finds the person-affecting view much more likely: she has 99% credence in the person-affecting view, and only 1% credence in egalitarian utilitarianism. And let's suppose that there's a huge amount at stake, in the choice between A and C, according to the person-affecting view, whereas not much is at stake in the choice between A and C, according to the egalitarian utilitarian view. Let's suppose that that the choice-worthiness difference between C and A, according to the person-affecting view, is 100 times greater than the choice-worthiness difference between A and C according to the egalitarian utilitarian view. Doing the math, the expected choice-worthiness of A is 0.01*1 = 0.01 and the expected choice-worthiness of C is 0.99 * 100 = 99.

Intuitively, it's appropriate to choose C over A: the situation is both much higher stakes according to the person-affecting view, and the decision-maker has much higher credence in the person-affecting view. But, by DoT, it's appropriate for the decision-maker to choose A over B, because egalitarian utilitarianism gives some reason in favour of choosing B over A, and the person-affecting view gives no reason in favour of choosing B over A. And, by exactly similar reasoning, DoT implies that it's appropriate for the decision-maker to choose B over C. Therefore, assuming acyclicity, it's not also

appropriate to choose C over A. So DoT conflicts with our intuitive appraisal of the above situation.

In the above case, both the size of the credence in the person-affecting view and the magnitude of the ratio of the choice-worthiness-difference between A and C across the two theories were arbitrary. So let's take that into account, and generalise the above objection into a principle, which we can call the Weak Comparativism Principle:

> *WCP*: If one has credence p in T_1, and credence $(1-p)$ in T_2 (where $1>p>0$), and *A* is more choice-worthy than *B* according to T_1 and *B* is more choice-worthy than *A* according to T_2, and the difference in choice-worthiness between *A* and *B* is n times greater, according to T_1, than it is according to T_2, then, for some p and n, one should choose *A* over *B*.

WCP is in conflict with DoT. But WCP is pretty darn plausible: it's about the weakest non-dominance principle that I can think of. It's *much* weaker, for example, than the principle that one should prefer *A* over *B* if *A* has higher expected choice-worthiness than *B*. It seems much more plausible than DoT.[130] So we should reject DoT.

Argument 2. DoT generates appropriateness-cycles

To see this, consider the following case. Suppose that you are the leader of the World Government, and you are to decide what policies to pursue. The principal questions facing you are whether to pursue a policy of population control, and to what extent you

[130] This seems especially true when we consider, as we will in argument 3, that we can formulate a principle that is very similar to DoT that does not suffer from the same conflict.

should mine natural resources, improving technological and economic progress but at the cost of environmental preservation.

You have many options available to you, but here are three. Option A involves population control, and moderate use of scarce resources. If so, then the future population will be one billion people at wellbeing level 4, one billion people at wellbeing level 2, and moderate environmental destruction. Option B involves population control, but the preservation of scarce resources. This will result in two billion people all living at wellbeing level 2, with no environmental destruction. Option C involves no population control, and the preservation of scarce resources. This will result in five billion people living at wellbeing level 2, and five billion people living at wellbeing level 1.

You are uncertain about what the true moral theory is. You split your confidence between three theories only. The first, T_1, is a person-affecting view. According to this view, when comparing two distributions of wellbeing between populations that have the same number of people, one should choose the distribution with the highest average wellbeing. Populations involving different numbers of people, on the other hand, are absolutely incomparable in value. On this view, environmental preservation is of no value. So, according to this theory, A is better than B, and C is incomparable with A and B (i.e. $A>B$; $B\sim C$; $A\sim C$). (i.e. $A>B$; $B\not\geq C$ and $C\not\geq B$; $A\not\geq C$ and $C\not\geq A$).

The second, T_2, is an environmentalist theory. On this theory, average wellbeing and environmental preservation are both of value, so one should choose policies that increase average wellbeing and choose policies that preserve the environment. But, on

this view, average wellbeing and environmental preservation are absolutely incomparable in value: there are no facts about how one should weigh average wellbeing and environmental preservation. So, according to this theory, B is better than C, and A is incomparable with B and C ($B \succ A$; $A \not\succcurlyeq C$ and $C \not\succcurlyeq A$; $A \not\succcurlyeq B$ and $B \not\succcurlyeq A$).

The third theory, T_3, is an egalitarian theory. It values both the sum total of wellbeing, and the level of equality of wellbeing among people. But, on this theory, the value of total wellbeing and the value of equality are absolutely incomparable: there are no facts about how one should weigh one against the other. So, according to this theory, C is better than A, and B is incomparable with C and A (i.e. $C \succ A$; $B \not\succcurlyeq C$ and $C \not\succcurlyeq B$; $A \not\succcurlyeq B$ and $B \not\succcurlyeq A$).

In the above case, DoT entails that one ought to have cyclical preferences. The above example is a bit complex to think through, so I lay out the theories in the following table, which makes things easier to see. Each theory has two columns underneath it. If one option is above another in the same column, then one has greater reason to choose the higher option than the lower option. If one option is in a different column than the other option, then the two options are incomparable.

	T_1		T_2		T_3
A	C	A	B	C	B
B			C	A	

According to DoT, it's appropriate to choose A over B, B over C, and C over A. That is, DoT entails that sometimes appropriateness is cyclical.

III. Responding to rejoinders

In response to these arguments, one might be tempted by one or both of two responses. In this section I'll argue that these responses fail.

1. Modify DoT

First, one might be tempted to modify DoT. In correspondence, Ross has suggested the following modified principle in place of the principle suggested in his article:

> *Modified Dominance over Theories (MDoT)*: For any action ϕ, rationality requires that if, among the theories in which you have positive credence, some imply that you have most objective reason to ϕ, and none imply that you have any objective reason not to ϕ, then you intend to ϕ.

The key change to this principle is that, while DoT appealed to the idea that one's credence in a theory might not give one subjective reason overall for preferring *x* to *y* over vice-versa, MDoT appeals to the idea that a theory might entail that there are *no* reasons against choosing a particular option.[131] To illustrate the importance of this

[131] Two other changes — that the principle is now explicitly wide scope, and that we are now talking about rational intention rather than rational choice — are not relevant to the arguments here discussed. The fact that the principle now only refers to an option's having most reason in favour of it, rather than being a comparative relation between two options, means that the cyclicity generated is over option-sets, rather than within an option-set. Given the arguments that follow, this also isn't an issue. Though these changes distract from the main issue somewhat, I wished to be true to Ross's new formulation.

distinction, consider that there are two ways in which incomparability can arise. First, it can arise because there are reasons in favour of one option, and reasons in favour of another option, but that those competing reasons cannot be weighed against one another, and so the two options are incomparable. A theory might, for example, claim that there is some reason to become a lawyer (because becoming a lawyer will benefit others), and some reason to become a clarinettist (because becoming a clarinettist furthers one's aesthetic ideals), but that these reasons cannot be weighed against each other. In which case having credence in that theory would not give one subjective reason for preferring to become a lawyer over a clarinettist or vice versa. But it would still be the case that there is some objective reason, according to that theory, not to become a lawyer, and some objective reason, according to that theory, not to become a clarinettist.

Second, incomparability can arise because some precondition of two options' being comparable is not met. For example, if nihilism is true then, though becoming a lawyer and becoming a clarinettist are incomparable, that is not because there are reasons in favour of each, and those reasons cannot be weighed against each other. Rather, the options are incomparable, according to nihilism, because there are no objective reasons at all. So credence in nihilism gives one *no* subjective reason to prefer becoming a lawyer to becoming a clarinettist or vice-versa. But, unlike in the previous case, this is because there's no objective reason against becoming a lawyer, and no objective reason against becoming a clarinettist.

Most instances of incomparability are similar to the former example rather than the latter, and so MDoT, while allowing one to reject nihilism, will run into trouble less

often than DoT because the kind of dominance it defines kicks in only in when the dominated theories offer no reason at all pertaining to the options in question. But it still gets into trouble. Consider the example from Argument 1, except now suppose that each pair of options is presented subsequently. When A and B are the available options, according to MDoT it's appropriate to choose A over B: the egalitarian utilitarian theory gives objective reason to choose A; whereas on the person-affecting view there is no objective reason against choosing A nor any objective reason against choosing B (because in neither case would choosing that option make anyone worse off than they could have been). Similarly, when B and C are the available options, according to MDoT it's appropriate to choose B over C. But, as before, intuitively it's appropriate to choose C over A: C is much more likely to be right, and there is much more at stake if it is right. So MDoT runs into just the same troubles as before DoT did, and for just the same reason.

2. Embrace cyclicity

Second, either instead of modifying the principle, or in addition to it, one might argue that, sometimes, it's appropriate to choose in a cyclical manner.[132] Against Argument 1, one might claim that, in the case I give, one ought to choose A over B, B over C, and C over A. Against Argument 2, one might claim that, in the case I give, one ought to choose A over B, B over C, and C over A. One might then develop an account of how

[132] Ross's view, in correspondence, is that one should *both* modify the principle in the way suggested and accept any resultant cyclicity.

one should overcome the difficulties that a decision-maker with cyclical appropriateness would face: how such a decision-maker could avoid money-pumps, and so on.

If one were willing to accept the theoretical cost of cyclical appropriateness, then one would escape my argument. However, accepting that one's metanormative theory can lead to cyclical appropriateness is a cost that should be borne only if one has a good reason for accepting such cyclicity. But, despite appearances, Ross's principle is not well-motivated.

To see this, we should consider analogues of DoT in other domains. Consider, for example:

> *Dominance over Times (DoTi)*: If, at some times, *x* is better off than *y*, and at no times is *x* worse off than *y*, then *x* is better off than *y* overall.

This principle looks prima facie compelling. But the principle is false, as can be seen from the following case:

> Abraham: (Ω, 1, 2, 3, 4, Ω, Ω, Ω)
>
> Bethenel: (Ω, Ω, 1, 2, 3, 4, 5, Ω)

In this example, the numbers represent how well off the person is at a time, and the omegas represent that the person is not alive at that time. Assume that it is never true to say that someone is better, worse, or equally as well off, at a time, as someone who is not alive at that time.[133] If so, then Abraham is better off at some times than Bethenel is, and is worse off at no times. So DoTi tells us that Abraham is better off overall than

[133] For a defence of that view, see (Broome 2004, 67).

Bethenel, even though Bethenel has a greater sum total of wellbeing, a greater average wellbeing, a greater peak wellbeing, and a better end of life, than Abraham. So the above example shows that DoTi should be rejected.

As a second example, consider the analogous dominance principle over people.

> *Dominance over People (DoP)*: If, for some people, one population *x* is better than another population *y*, and for no-one is population *x* worse than population *y*, then *x* is better than *y* overall.

Again, this principle seems plausible. But only if we assume that one population can't be incomparable with another from the perspective of one person. Consider, for example, the following case:

> Population A: (2, Ω, 1)
>
> Population B: (1, 2, Ω)
>
> Population C: (Ω, 1, 2)

In this example, the numbers represent how well off the person is, and the omegas represent that the person is not alive. Assume that it is meaningless to say that someone who exists could be better, worse, or equally as well off, in a world in which they do not exist.[134] In which case, according to DoP, Population A is better than Population B; Population B is better than Population C; and Population C is better than Population

[134] (Broome 2004, 67).

A. So, according to DoP, A is better than B, B is better than C, and C is better than A. DoP makes betterness cyclical.[135]

So we get formally analogous problems if we attempt to use dominance over incomparability in other domains as we do in the case of using dominance over theories. In these other domains, these problems mean that we should reject using dominance over incomparability. We must distinguish carefully between two formulations of dominance principles. For example, if we wish to formulate a dominance principle over people, rather than DoP we should write:

> *Genuine Dominance over People (GDoP)*: If, for some people, one population *x* is better than another population *y*, and, for all other people, *x* and *y* are *equally good*, then *x* is better than *y* overall.

If we want to formulate a dominance principle over theories, we should write:

> *Genuine Dominance over Theories (GDoT):* If some theories in which you have credence claim that *x* is more choice-worthy than *y*, and all other theories in which you have credence claim that *x* and *y* are *equally* choice-worthy, then it is appropriate to choose *x* over *y*.

GDoP and DoP appear very similar. But only the former is an acceptable principle of aggregation over people — the latter has apparent plausibility only because of its superficial similarity to the former. In exactly the same way, it is only GDoT that is an

[135] Note that, though I gave the analogue of Argument 1 against DoTi and the analogue of Argument 2 against DoP, that the analogues of *both* arguments can be run against both principles (assuming, at least, that a person can go out of existence, for a time, and come back into existence). To see this, one simply needs to switch round and relabel the examples I give against each principle.

acceptable metanormative principle — DoT's apparent plausibility only arose because it is so easily confused with GDoT. Once this distinction is made, it should be clear that DoT has no theoretical plausibility. So we have no reason to endorse DoT, and therefore we have no reason to endorse the consequent cyclicity that DoT generates.

IV. The infectious nihlism problem

As a rejoinder, one might motivate MDoT (or DoT) on the grounds of sheer necessity. If we have non-zero credence in nihilism and use expected choice-worthiness reasoning, we get infected and the expected choice-worthiness of all options is undefined. Using MDoT and accepting the chance of cyclicity is a small price to pay, so one might argue, in order to avoid the expected value of all options being undefined. So it is rational to use a dominance principle like MDoT to reject nihilism.

I have sympathy with this line of reasoning. Ultimately, I will in fact endorse accepting the cyclicity of appropriateness across decision-situations. Unfortunately, however, the reasoning does not work as a defence of MDoT. The reason why is because the problem of infectious nihilism is just the tip of a much larger iceberg. The expected choice-worthiness of almost all options is undefined *even if* we accept MDoT and the resulting cyclicity. Moreover, options are undefined for precisely the same reason as before: because of the fact that incomparability is infectious under uncertainty.

Consider, for example, an 'incomparabilist' moral theory according to which there are two values, V_1 and V_2, but that these values are absolutely incomparable: it is never

true, on this theory, that one ought to sacrifice some amount of V_1 for a large enough gain in V_2, or vice-versa. As a result, on this theory, for any two options, *A* and *B*, if *A* produces more of V_1 than *B*, but less of V_2 than *B*, then *A* and *B* are incomparable. The value of *A* is undefined relative to *B*.

Now suppose that our decision-maker has non-zero credence in this view. If so, then there's trouble, as follows. As noted before, if the choice-worthiness of an option is undefined according to one theory in which the decision-maker has credence, then expected choice-worthiness is undefined. So, because, according to this incomparabilist moral view, the choice-worthiness of *A* relative to *B* is undefined, for a decision-maker with non-zero credence in this moral view, the expected choice-worthiness of *A* compared with *B* is undefined, too. This is true even if one has very small credence in the incomparabilist view.

We can go further. For the expected choice-worthiness of *A* relative to *B* to be undefined, it need not be the case that option *A* certainly involves producing more of V_1 and producing less of V_2 than *B*. Just one state of nature in which the value of *A* is undefined relative to *B* is sufficient for the expectation of *A* relative to *B* to be undefined. So non-zero credence in the incomparabilist moral view combined with the mere empirical possibility that *A* would produce more of V_1 than *B* and less of V_2 than *B* is sufficient for the expected value of *A* relative to *B* to be undefined.

Again, this is pretty terrifying. For almost any option and any pair of values, we should have some credence that that option involves increasing one value but decreasing the second value. And the idea that there might exist absolute incomparability between

values is not just a speculative hypothesis; it's a definite epistemic possibility.[136] The idea that we should have zero credence in such views is highly implausible. In which case, even if we accept MDoT, if we try to incorporate moral uncertainty into our expected choice-worthiness reasoning, then, on the standard understandings of incomparability and expected choice-worthiness maximisation, for every one of us the expected choice-worthiness of almost all options is undefined.

This should make us suspicious of using anything as simple as Rossian dominance-style reasoning in order to reject nihilism. Infectious nihilism is simply one instance of the more general problem of infectious incomparability. In order to understand why we should reject nihilism, we need an understanding of how in general it is rational to act given that one has credence in theories that posit extensive incomparability. So let's now look at some such accounts.

V. Theories of rational reponses to incomparability of value

We may divide theories of rational action in the light of incomparability of value into two broad categories: what I'll call *Permissive* accounts, and *Hard-line* accounts. The two categories are distinguished by their attitude to dynamic choice. *Hard-line* accounts claim that, if nothing changes other than the option-set, it's irrational for a decision-maker to make the following pattern of choices: choose A over B if {A,B} were the

[136] For example, the idea that different people's wellbeings are absolutely incomparable is a standard view in classical economics. It's also one way of understanding the 'separateness of persons' argument against utilitarianism. Similarly, the idea that very different goods, such as the value of the environment and the value of hedonic states, are absolutely incomparable, is not *so* implausible that one could justifiably assign zero credence to it.

option-set, choose B over C if {B,C} were the option-set; and choose C over A if {A,C} were the option-set. According to *Permissive* accounts, such a pattern of choices is not irrational. Indeed, according to *Permissivists* it might be irrational *not* to make such a sequence of choices.

Prima facie, however, neither seems to be able to adequately handle the infectious incomparability problem. The most natural permissivist response is as follows.[137] Let us call a *completion* of a theory's choice-worthiness ordering to be a complete choice-worthiness ordering that is consistent with every positive relation in the incomplete theory's choice-worthiness ordering. According to this account:

> (i) *A* is more appropriate than *B* iff *A* has a greater expected choice-worthiness than *B* on all possible completions of every moral theory in which the decision-maker has credence.
>
> (ii) Both *A* and *B* are appropriate iff *A* has a greater expected choice-worthiness than *B* on some possible completions of every moral theory in which the decision-maker has credence, and *B* has a greater expected choice-worthiness than *A* on some possible completions of every moral theory in which the decision-maker has credence.

One might think that this solves the infectious incomparability issue rather neatly. Suppose, for example, that Sophie has 99% credence in Kant's ethics, and 1% credence in nihilism. She is faced with two options:

> A: tell a grave lie in order to mildly benefit herself.

[137] A very similar account is suggested by (Hare 2010), who calls it 'prospectism'.

B: tell the truth.

On Kant's ethics, B≻A. On nihilism A⊁B and B⊁A. The possible completions of nihilism are: A≻B; B≻A; and A~B. Kant's ethics, let us suppose, only orders A and B ordinally. Sophie faces only ordinally measurable moral theories, so, as I have argued, the Borda Rule is the correct way to aggregate her uncertainty. Using the Borda rule, B is the appropriate option on all completions of nihilism. So, using the supervaluationist approach, nihilism's incomparability is not disastrously infectious. If Sophie's credence in nihilism were higher (greater than 50% in nihilism, and less than 50% in Kant's ethics), then both A and B would be appropriate. But that is a much less worrying problem, if it is a problem at all, than the previous infectiousness problem, where the expected choice-worthiness of all options was undefined no matter how small one's credence in nihilism.

However, this account doesn't work. In the above example, we only considered *ordinal* completions of nihilism's choice-worthiness ordering. But that seems unwarranted. Consider another example.

Suppose, that Terry has 99% credence in utilitarianism, and 1% credence in nihilism. She is faced with two options:

A: save the lives of one billion happy people.

B: let those one billion happy people die.

If we consider only ordinal completions of nihilism's choice-worthiness ordering then, using the metanormative theory I have argued for, A would again be the most

appropriate option on all completions of nihilism. But there is no reason to only consider ordinal completions of nihilism. According to utilitarianism, not only is it true that A is more choice-worthy than B: A is *much* more choice-worthy than B. But if we allow cardinal completions of nihilism, then one could extend nihilism such that B is more choice-worthy than A and the difference in choice-worthiness between B and A is equal to the value of letting one trillion happy people die. Using this completion of nihilism, B would be the most appropriate option. If we allow cardinal completions of nihilism, then, in the above case, A is neither more appropriate than B nor vice-versa. The infectious incomparability problem still remains.

I have said that there is no reason to only allow merely ordinal completions of nihilism. One might claim that the infectious incomparability issue is a good reason for doing this. But that would only be true if there were no other more principled way to avoid the problem. I will soon suggest a principled way of avoiding the problem. But, first, let's see if the hard-line approach can do any better.

The most natural hard-line approach is as follows: for every theory, choose *any* completion of that theory's choice-worthiness ordering, and then stick with that completion.[138]

The implications of this proposal are just as absurd as the implications for the supervaluationist approach. Consider Terry again. If she is permitted to choose *any* completion of each incomplete theory in which she has credence, then she may choose the completion of nihilism according to which B is more choice-worthy than A and the

[138] This account has been suggested in conversation by Toby Ord.

choice-worthiness difference between B and A is as great as the difference in choice-worthiness between saving one trillion happy lives and doing nothing.

The implications of this hard-line account are a bit different from the implications of the supervaluationist account. For the supervaluationist account, in every decision-situation involving positive credence in nihilism, all options are rationally permissible. For the hard-line account, after having chosen some completion of the incomplete theories in which she has credence, there will be a determinate answer about which option or options are rationally permissible, and which are impermissible. But there will be a great deal of arbitrariness in this: the pattern of options that are permissible or impermissible will be determined entirely by the decision-maker's initial decision about how to extend nihilism. So this hard-line account does not provide a satisfactory response to the infectiousness problem.

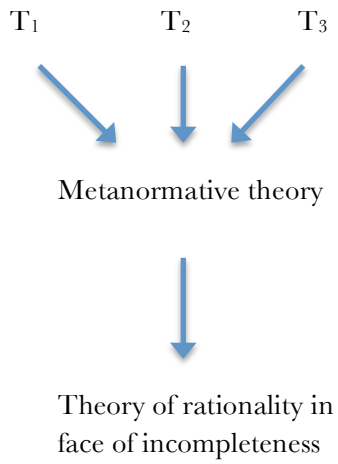
VI. Solving the infectious incomparability problem

There is, I believe, a good response to the infectiousness problem — a response that I suspect is only available to the permissivist.

In both of the above accounts, we presupposed one particular order of aggregation. At the first stage, we moved directly from a distribution of moral theories (some of which are incomplete) to an appropriateness ordering. (Permissivist accounts move from incomplete moral theories to an incomplete appropriateness ordering. Hard-line

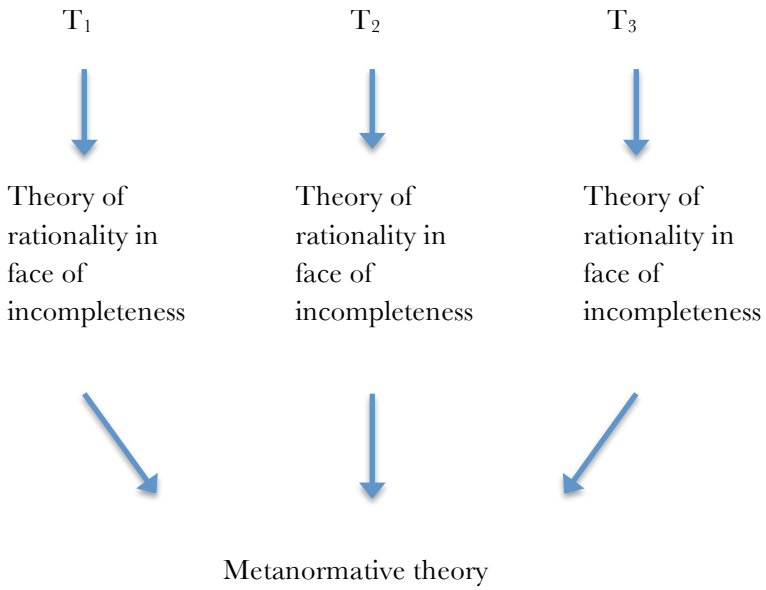
accounts move from incomplete moral theories to a complete appropriateness ordering.) At the second stage, we applied an account of what it's rational to do in the face of an incomplete appropriateness ordering. (In both cases assuming that what it's rational to do is to choose any maximally appropriate option.)

Visually, this looks as follows:



The solution I suggest is to reverse the order of aggregation. At the first stage, we should apply, to each moral theory, an account of what it's rational to do in the face of incompleteness. Then we aggregate the output orderings of each of those accounts into a final appropriateness ordering.

Visually, this looks at follows:



Consider how this would work with the supervaluationist account. And consider Terry again, who has 99% credence in utilitarianism, and 1% credence in nihilism. She is faced with two options:

A: save the lives of one billion happy people

B: let those one billion happy people die

Utilitarianism, we are assuming, is complete. So the rational choice-worthiness ordering we move to would be the same as the utilitarianism's choice-worthiness ordering. But the rational choice-worthiness ordering for nihilism would not be the same. On the supervaluationist account, both A and B will be rationally permissible, on nihilism.[139] In which case, the aggregation at the next stage is easy. Terry is 99% certain that she should rationally choose A over B. And she's 1% certain that both A and B are

[139] Remembering, as I noted in fn.125, that that the version of nihilism we are considering is only nihilistic about first-order normative claims, and not about facts about what it's appropriate or rationally permissible to do.

rationally permissible. So it's more appropriate to choose A over B. This was the result we wanted.

In fact, this account will always allow one to rationally reject nihilism. On supervaluationism, the rational response to total incompleteness is for all options to be rationally permissible. So, when aggregating in the new way I suggest, every option will be rationally permissible, on nihilism, in every decision-situation. Nihilism will never affect what it's appropriate to do. So you will be able to rationally reject nihilism. Moreover, this alternative way of aggregating avoids the infectious incomparability issue in general. On the supervaluationist account, theories that give radically incomplete choice-worthiness orderings generate extensive rational permissibility. But rational permissibility does not 'infect' the ultimate appropriateness ordering. So permissivists are able to dodge the infectious incomparability issue. The account has the implication that appropriateness is cyclical across decision-situations; but cyclicity across decision-situations was already something that the permissivist was happy with. So the permissivist is on relatively good footing.

However, it seems that this move is much harder to make for hard-line accounts, because hard-line accounts would have to move directly from incomplete theories to a complete appropriateness ordering. Consider the hard-line account suggested above. If we switched the order of aggregation, we would still have exactly the same problem that the arbitrary choice of completion of nihilism, at the first stage, would entirely determine which options are appropriate and which are not. In fact, I can't think of any way for hard-line accounts to make the move that the permissivist account made.

Considerations relating to incompleteness under moral uncertainty therefore have important more general implications. Insofar as permissivist accounts seem to be able to avoid infectious incomparability, whereas hard-line accounts cannot, consideration of moral uncertainty given credence in incomplete theories provides an argument in favour of permissivist accounts over hard-line accounts.

Conclusion

In this chapter I considered how to handle uncertainty over incomplete moral theories, including radically incomplete moral theories such as nihilism. I argued against Jacob Ross's argument for the conclusion that we can reject ethical nihilism, and argued that 'infectious incomparability' is a serious problem for MEC. I then suggested a solution to the infectious incomparability problem, and showed how it generated a new argument against hard-line accounts of rationality in the face of incomparability.

Chapter 6: Smokers, Psychos, and Decision-Theoretic Uncertainty

Introduction

Recently, there has been significant debate about the nature of decision theory: whether the correct theory is evidential, causal, or something else again. The principal problem in this debate is that powerful counterexamples seem to have been raised to all the major views.

In response to this, some philosophers have expressed pessimism. Rachael Briggs argues that "no decision rule can do everything that we want";[140] Andy Egan regretfully asserts that he "[does] not have… a theory to offer"[141] that is able to get the intuitions right in the cases that have been given in the literature. Others instead have tried to develop new decision theories that satisfy the intuitions,[142] but these generally come at the cost of considerable theoretical inelegance, or suffer from counterexamples of their own.

In this chapter I suggest that metanormativism is a powerful tool in the debate between causal and evidential decision theory, in two ways. First, metanormativism can provide a novel explanation of why our intuitions sometimes favour evidential decision theory (as in the Psychopath Button) and sometimes favour causal decision theory (as in the

[140] (Briggs 2010, 1)
[141] (Egan 2007, 93, 113)
[142] For example: (Arntzenius 2008; Wedgwood 2013; J. E. Gustafsson 2011; Price 2012).

Smoking Lesion). Second, metanormativist considerations considerably strengthen the case for causal decision theory over evidential decision theory.

The structure of my argument is as follows. After quickly describing Newcomb's problem and the causal/evidential distinction in section I, in section II I introduce and motivate metanormativism about decision theory: that is, the idea that there is an important sense of 'ought' (though certainly not the only sense of 'ought') according which a decision-maker ought to take decision-theoretic uncertainty into account. I call any metanormative theory that takes decision-theoretic uncertainty into account a type of *meta decision theory*.[143] In section III, I show how the most natural formulation of meta decision theory gets the right intuitive results in both *The Smoking Lesion* and *The Psychopath Button*. Moreover, it can convincingly explain *why* we get the intuitions we do, and it can provide this explanation in a way that is far more theoretically elegant than other accounts that have been proposed in the literature.

In section IV, I show that decision-theoretic uncertainty undermines the intuitive case for evidential decision theory over causal decision theory. In section V, I show that decision-theoretic uncertainty gives us the resources to construct a counterexample to the "Why Ain'cha Rich?" argument in favour of evidential decision theory. In section VI, I consider and respond to the objection that metanormativism faces a vicious regress.

[143] In contrast to a metanormative view according to which there are norms that are relative to moral and prudential uncertainty, but not relative to decision-theoretic uncertainty.

I. Newcomb's Problem

Newcomb's problem is typically introduced through the following case:

Standard Predictor[144]

> You have two boxes in front of you, Box A and Box B. Box A is opaque; box B, transparent. You have the option to take either box A only, or both B and A. You can see that Box B contains $1000. Box A either contains $1million or $0. Moreover, someone ("The Predictor") with an amazing ability to predict other people's actions had control over the boxes. If the Predictor predicted that you would choose Box A only, then he put $1 million into Box A. If the Predictor predicted that you would choose both boxes, then the Predictor put nothing into Box A. What should you do?

Representing the decision-problem in a table, we have:

	Money in both boxes	Money in one box only
Take one box only	$1,000,000	$0
Take both boxes	$1,001,000	$1,000

There are two distinct but each seemingly compelling lines of reasoning available here. First, I could reason that if I take Box A only, then I'm almost certainly going to get

[144] In all the cases I give in this paper, I stipulate that the relevant correlations (e.g. between the Predictor and the money in the box, and between smoking and having a lesion) are perfect correlations. I discuss whether this aspect of my examples is problematic in section III.

$1million. In contrast, if I take both boxes, I'm almost certainly going to get only $1000. So I should take Box A only. Such reasoning motivates *evidential decision theory (EDT)*. According to EDT, one should choose the option with the maximal evidential expected value, where the evidential expected value of an action is defined as the sum, over all possible outcomes, of the value of the outcome, given that you perform that action, multiplied by the probability of the outcome conditional on you performing that action.[145] According to this account, in *Standard Predictor* you should one-box. The precise formalization of this view will not matter for the purposes of this paper, but one simple way to formalize the view is as follows:

$$EEV(A) = \sum_{i=1}^{n} C(O_i|A)V(O_i \,\&\, A)$$

Just as stated in the Introduction, in the above equation, C is the decision-maker's credence function, and A, B, C (etc) are actions that are available to the decision-maker. O_1, O_2 … O_n are propositions that describe a way the world might be. $\mathcal{O}$ is the set of such propositions. The members of $\mathcal{O}$ form a partition (with respect to C) over the state of the world, so the decision-maker is certain that one, but only one, of O_1, O_2, … O_n is true. V is the decision-maker's *value function*: for any outcome O_i, $V(O_i)$ takes a real number that measures how valuable O_i is to the decision-maker.[146]

The above gave a line of reasoning that favoured evidential decision theory. But there is an alternative line of argument. I could reason that the Predictor has already put the

[145] In what follows I will talk about 'value' rather than 'choice-worthiness' simply to stay in line with the decision-theoretic literature. I don't think that anything of importance hangs on this.
[146] In this chapter for simplicity I assume that the decision-maker is certain in her value-function over outcomes.

$1million in Box A, or decided against doing so. My choosing both boxes can't change that. And, no matter what amount of money is in Box A, I'll get an additional $1000 if I take both boxes. So I should take both boxes. Such reasoning motivates *causal decision theory* (CDT). According to CDT, one should choose the option with the highest causal expected value, where the causal expected value (CEV) of an action is defined as the sum, over all outcomes, of the value of that outcome multiplied by the probability of the outcome counterfactually conditional on one's action. There are very many ways to formalize CDT, but these won't matter for my purposes, so I will use the following simple formulation: [147]

$$CEV(A) = \sum_{i=1}^{n} C(A \Rightarrow O_i)V(O_i \,\&\, A)$$

In this equation, '⇒' denotes the counterfactual conditional: that is, a conditional of the form "if I were to perform A, O_i would happen". According to this account, in *Standard Predictor* you should two-box.

Different people's intuitions vary strongly in response to the *Standard Predictor*. So, in attempting to adjudicate between causal and evidential decision theory, other cases are normally used. But before moving on to them, I'll introduce and motivate the idea of meta decision theory.

[147] A very similar formulation is given, for example, in (Ahmed 2013).

II. Meta Decision Theory

Given the trenchant disagreement between intelligent and well-informed philosophers, it seems highly plausible that one should not be certain in either causal or evidential decision theory. In light of this fact, Robert Nozick briefly raised an interesting idea: that perhaps one should take decision-theoretic uncertainty into account in one's decision-making.[148] He noticed that our intuitions in Newcomb problems seem to be *stakes-sensitive*. That is, it seems that we can generate clear counterexamples to both EDT and CDT simply by playing around with the *Standard Predictor* case. By altering the stakes, we can alter our intuitions. Consider, first, the following case:[149]

> *High-Stakes Predictor I (HSP-I)*
>
> Box A is opaque; Box B, transparent. If the Predictor predicts that you choose Box A only, then he puts one wish into Box A. With that wish, you'd save the lives of 1 million terminally ill children. If he predicts that you choose both Box A and Box B, then he will put nothing into Box A. Box B — transparent to you — contains a stick of gum. You have two options only: choose Box A, or choose both Box A and Box B.

Representing this in a table:

[148] (Nozick 1994, 43–50). Toby Ord and I in conversation independently came up with this idea before discovering that Nozick had written on this.

[149] This example and the next are structurally the same as examples given by Nozick — I've just altered them a little bit to make the case even stronger.

	Wishes in both boxes	Wishes in one box only
Take one box only	1,000,000 lives	Nothing
Take both boxes	1,000,000 lives + gum	Gum

In this case, intuitively, should you one box, or two box? Though it can be difficult not to let theory cloud one's judgment, my intuitive view is clearly that if someone two-boxes in that case, they made the wrong decision. So do we have a slam-dunk argument in favour of EDT? Unfortunately not. Consider the following case:

> *High-Stakes Predictor II (HSP-II)*
>
> Box C is opaque; Box D, transparent. If the Predictor predicts that you choose Box C only, then he puts one wish into Box C, and also a stick of gum. With that wish, you save the lives of 1 million terminally ill children. If he predicts that you choose both Box C and Box D, then he will put nothing into Box C. Box D — transparent to you — contains an identical wish, also with the power to save the lives of 1 million children, so if one had both wishes one would save 2 million children in total. However, Box D contains no gum. One has two options only: choose Box C only, or both Box C and Box D.

Representing this in a table:

	Wishes in both boxes	Wishes in one box only
Take one box only	1,000,000 lives + gum	Nothing
Take both boxes	2,000,000 lives + gum	1,000,000 lives

In this case, intuitively, should you one box, or two box? My intuitive view is clear: if someone one-boxes in the above case, they made the wrong decision.

What's going on in these two cases? From one perspective, they are structurally identical. In both cases, EDT recommends one-boxing, because one-boxing has the higher evidential expected value. In both cases, CDT recommends two-boxing, because two-boxing has the higher causal expected value (and, indeed, dominates one-boxing). From another perspective, however, they are very different. In *HSP-I*, one's decision is of huge consequence, according to EDT. From its perspective, the difference in value between one-boxing and two-boxing is the difference in value between saving a million innocent lives and getting a free stick of gum. For CDT, however, one's decision in *HSP-I* is fairly trivial. The decision about whether to one-box or two-box is merely the decision about whether to get a free stick of gum or not. In contrast, in *HSP-II*, the decision is of huge consequence for CDT. The decision between one-boxing and two-boxing is the decision about whether to save a million innocent lives. Whereas, the decision in *HSP-II* is fairly trivial for EDT: it merely concerns whether to get a free stick of gum or not.

As Nozick noticed, this sort of stakes-sensitivity is suggestive of the idea that our intuitions are governed at least in part by uncertainty over both CDT and EDT. We

feel the force of both sorts of decision theory, and so we have credence in both of them. And then, when making decisions, we hedge our bets, going with CDT when the relative stakes are sufficiently high for CDT, and going with EDT when the relative stakes are sufficiently high for EDT.[150] I call this idea *Meta Decision Theory* (MDT).[151] According to any version of MDT, one should maximize *meta expected value*, where the meta expected value (MEV) of an action is defined as the sum, over all decision theories, of the probability of that decision theory multiplied by the value of that action, conditional on that decision theory. Or, formally (and again with the caveat that there are many possible ways to formalize this idea):

$$MEV(A) = \sum_{i=1}^{n} C(D_i)\, D_i(A)$$

In this formula, D_1, D_2, … D_n each refers to a decision theory, and $D_i(A)$ is the value that D_i assigns to A. In section V, I distinguish between the causal version of MDT and the evidential version of MDT. However, until that point the distinction will not matter for my purposes so I state MDT simply in terms of unconditional probabilities.[152]

[150] Of course, this is not the only possible explanation for why our intuitions switch in the two cases. In sections II and III, I consider and ultimately reject alternative explanations of this phenomenon.

[151] A terminological clarification: I'll use "Meta Decision Theory" (capital letters) or MDT to refer to the specific view that one ought to maximize expected choice-worthiness over decision theories. I'll use "meta decision theory" to refer to any decision theory that claims that what one ought to do (in the relevant sense) is determined in part by one's credences in first-order decision theories.

[152] A couple of other notes on this. First, we should of course have non-zero credence in decision theories other than CDT and EDT, such as Benchmark Theory (Wedgwood 2013), and so uncertainty about these other theories will also have to be taken into account. In order to keep things simple however I will leave these alternative decision theories to the side. Second, one might worry whether meta decision theory suffers from the problem of intertheoretic comparisons. However, the problem of intertheoretic comparisons is substantially easier in the case of EDT and CDT than it is between different moral theories. EDT and CDT both agree on what a decision-maker should do in all the many cases where $Cr(A \Rightarrow O_i) = Cr(A|O_i)$. We can use this agreement to normalize the two theories.

On the reasonable assumption that we have at least small positive credence in each of EDT and CDT, MDT would make sense of the stakes-sensitivity suggested above. Because *HSP-I* is so much higher-stakes according to EDT than according to CDT, even very small credence in EDT would make one-boxing have the higher MEV. The same is true vice-versa for *HDP-II*.

These high-stakes predictor cases make me think that there is a sense of 'ought' that is relative to decision-theoretic uncertainty. However, oddly Nozick himself ultimately *rejects* that idea, in favour of a subtly different idea. He says:

> I suggest that we go further and say not merely that we are uncertain about which *one* of these two principles, [CDT] and [EDT], is (all by itself) correct, but that both of these principles are legitimate and each must be given its respective due. The weights, then, are not measures of uncertainty but measures of the legitimate force of each principle. We thus have a *normative* theory that directs a person to choose an act with maximal decision-value.[153]

And also:

> Theorists of rationality have been intent upon formulating the one correct and complete set of principles to be applied unreservedly in all decision situations. But they have not yet reached this—at any rate, we do not have complete confidence that they have. In this situation, won't a prudent and rational individual hedge her bets? I want to say more, namely, that no one of the principles alone is wholly adequate— it's not simply that we have yet to find the knockdown argument for the one that is correct.[154]

[153] (Nozick 1994, 45).
[154] (Nozick 1994, 46–47).

That is, as I understand him, Nozick rejects what I call meta decision theory in favour of what might be called *decision-theoretic pluralism* (DTP).[155] Whereas MDT is not a rival to CDT or EDT, DTP *is* a rival first-order theory.

What's odd about Nozick's suggestion is that, even though MDT seems to be the natural explanation of our stakes-sensitive intuitions, he gives no argument for preferring DTP to MDT (apart, perhaps, from the cryptic suggestion that MDT wouldn't be "genuinely normative"). But we already know that we're decision-theoretically uncertain, and that expected utility theory in general the best way to handle uncertainty. This is enough to make MDT plausible, and MDT is enough to explain our stakes-sensitive intuitions. There therefore seems to be nothing to gain by suggesting that DTP is true. So DTP seems unmotivated.

Moreover, DTP is not merely unmotivated: it also has two major problems that MDT lacks. First, DTP has multiple explanatory gaps. Why weigh EDT against CDT in one way rather than another? MDT has a principled answer to this — namely, that the weights are one's credences — whereas DTP does not. And why should the values EDT and CDT assign to acts be additively separable? Again, MDT has an explanation of this — that taking an expectation requires values across states to be additively separable — whereas DTP does not. And, finally, why even think that there would be different sorts of "decision-theoretic value"? Decision-theoretic pluralism is very different from other sorts of pluralism about value: typical pluralist theories make sense of different values

[155] The analogy is with pluralist moral theories. Someone who maximizes expected choice-worthiness under uncertainty about whether only wellbeing, or both knowledge and wellbeing, are of value looks a lot like someone who is conforming with a first-order moral theory that assigns both wellbeing and knowledge value. In the same way, someone who follows MDT looks a lot like someone who is conforming with a first-order decision theory that gives casual expected value and evidential expected value weight.

because different values supervene on different sorts of *stuff*. In contrast, the different decision-theoretic values that Nozick suggests arise merely out of how uncertainty is taken into account. So Nozick's account does not gain plausibility from the plausibility of pluralism about value in general.

Second, DTP misrepresents what's going on in the stakes-adjusted Newcomb cases. To see this, consider a variation on his cases.[156]

> *Four-box Predictor*
>
> Box A and Box C are opaque; Box B and Box D, transparent. The Predictor has a 100% success rate at predicting which box or boxes you'll choose. You have the following four options:
>
> (1): Take A and C only
>
> (2): Take A, B and C
>
> (3): Take A, C and D
>
> (4): Take A, B, C and D
>
> If the Predictor predicts you will take Box B, he will put nothing in Box A. If he predicts you will not take Box B, he will put into Box A a wish with value of 1 million children's lives.

[156] I thank Toby Ord for this suggestion.

If the Predictor predicts you will take Box D, he will put nothing in Box C. If he predicts you will not take Box D, he will put into Box C a wish with the value of 1 million children's lives, plus a stick of gum.

Box B — transparent to you — contains a stick of gum. Box D — also transparent to you — contains a wish with the value of 1 million children's lives but no gum.

Representing this in a table:

	Wish in neither A nor C	Wish in A, but not C	Wish in C, but not A	Wish in both A and C
Take A and C	Nothing	1 million lives	1 million lives + 1 bit of gum	2 million lives +1 bit of gum
Take A, B and C	1 bit of gum	1 million lives + 1 bit of gum	1 million lives + 2 bits of gum	2 million lives + 2 bits of gum
Take A, C and D	1 million lives	2 million lives	2 million lives + 1 bit of gum	3 million lives + 1 bit of gum
Take A, B, C and D	1 million lives + 1 bit of gum	2 million lives + 1 bit of gum	2 million lives + 2 bits of gum	3 million lives + 2 bits of gum

The astute might have noticed that someone in a *Four-box Predictor* situation is just someone who faces both *HSP-I* and *HSP-II* at the same time. The very astute might have noticed that this therefore constitutes a "Jackson Case" under decision-theoretic

uncertainty: a case in which one ought (in some sense) to do something that one knows one ought (in some other sense) not to do.[157]

According to CDT, one ought to perform act (4), two-boxing with respect to both A and B, and C and D. According to EDT, one ought to perform act (1), one-boxing with respect to both A and B, and C and D. But we should think that, in at least some sense of 'ought', what the decision-maker ought to do is act (3). By performing (4) rather than (3), one risks losing the opportunity to save 1 million children for the sake of a stick of gum. (This was the motivation for one-boxing in *HSP-I*). By performing (1) rather than (3), again one risks losing the opportunity to save 1 million children for the sake of a stick of gum (this was the motivation for two-boxing in *HSP-II*). And if one performs act (2) rather than (3), one takes both risks at the same time. So one should perform act (3), and take Boxes A, C and D: that's the only safe bet. And it's the only choice that seems consistent with our intuitions in both *HSP-I* and *HSP-II*.

In the above situation, the correct thing to say, I think, is that, in *some* sense of 'ought' (the sense that first-order decision theories are talking about), one ought to perform either act (1) or act (4), but that, in another sense of 'ought' (the sense that is relative to decision-theoretic uncertainty), one ought to perform act (3). That's the appraisal that MDT gives of the situation. But that's not the appraisal that Nozick's view gives. According to Nozick's view, all there is to say is that one ought to perform (3), because that's what the true decision theory (that is, DTP) claims: it's simply false, in any sense, that one ought to choose either (1) or (4). And that seems to misrepresent what's really going on in our appraisal of the *Four-Box Predictor*.

[157] Where "Jackson case" refers to (Jackson 1991).

For these reasons, for the rest of the paper I will set Nozick's view to one side, and instead assume that MDT is the correct conclusion to draw from the stakes-sensitivity of our intuitions. Nozick quickly moved on from the suggestion, and as far as I know it has not been pursued elsewhere.[158] But that's unfortunate, because MDT has important implications that have not been noticed.

III. The Smoking Lesion and the Psychopath Button

First, MDT allows us to resolve an apparent conflict in our intuitions.[159] My suggestion is that the divergence in our intuitions across cases in the literature can be understood as hedging between EDT and CDT, in a way that is mandated by MDT.

Consider *The Smoking Lesion*:[160]

> *The Smoking Lesion*
>
> Susan is debating whether or not to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause — a lesion that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking

[158] Though (Sepielli 2013) considers the view in the context of discussing the regress problem.
[159] That is not to say that it can resolve the divergence in *all* of our intuitions. In particular, there is a class of cases that it seems to me are not about whether evidential or causal decision theory is true, but rather are about whether the usual formulation of causal decision theory accurately captures the idea of 'causing the best consequences'. I place Andy Egan's (2007) time-traveller and Oracle cases in this category, as well as Arif Ahmed's (2013) 'Nomological Gamble' example. I take these examples to violate the letter of CDT, but not its spirit (though Ahmed goes on to argue that *no* way of formalising CDT in such a way that it gets the right answer in his case is compatible with free choice. That is an interesting, but very different, argument).
[160] Here I use the formulation given in (Egan 2007, 94).

> and cancer. Susan prefers smoking without cancer to not smoking without cancer, and she prefers smoking with cancer to not smoking with cancer.

In this case, intuitively Susan should smoke. But, problematically, EDT recommends against smoking. *The Smoking Lesion* has been taken to be a fatal counterexample to EDT. However, if our intuitions are explained in part by MDT (rather than by any other features of the case, such as its causal structure, or how realistic it is), then our intuition regarding *The Smoking Lesion* should change if we alter the stakes. And it seems that it is. Consider the following case.

> *Stakes-Adjusted Smoking Lesion*
>
> The lesion is not correlated with mere lung cancer. Rather, the lesion causes people both to smoke before they are 35 and to burst into flames on their 35th birthday, enduring several hours of agony, before dying (even though smoking does not *cause* the spontaneous self-combustion). Moreover, let us suppose that Susan isn't really that fussed about smoking. She hasn't been inclined to smoke previously, but she's feeling whimsical today and so has a slight preference for smoking that cigarette. It's the day before her 35th birthday. Should she smoke?

In this case, it seems very clear to me, intuitively speaking, that Susan should not smoke, even though CDT would recommend smoking.[161] Simply by altering the stakes, we have transformed an apparent counterexample to EDT into an apparent counterexample to CDT. This stakes-sensitivity is exactly what MDT would predict.

[161] Some causal decision theorists I have spoken with have bitten the bullet in this case. But I have a very hard time believing that such a response is genuinely a basic intuition, rather than a judgment that has been tainted by one's theoretical commitments.

Next, consider *The Psychopath Button*:[162]

> *The Psychopath Button*
>
> Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. And Paul is almost certain that he is not a psychopath. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world *with* psychopaths to dying. Should Paul press the button?

In this case, intuitively Paul should not press the button. But, problematically, according to CDT Paul should press the button. Again, however, if MDT explains our intuitions, then our intuitions about *The Psychopath Button* should be stakes-sensitive. And it seems that they are. Consider the following modification of the case:

> *Stakes-Adjusted Psychopath Button*
>
> It is 1890, and that Paul knows that Baby Hitler is a psychopath, knows that only one other person (which may be him) is a psychopath, and he knows of the atrocities that will happen in the following sixty years if Hitler survives. Moreover, let us suppose that Paul has a terminal illness. He will surely die within a few hours. He wants to have those last few hours alive and, being a selfish sort of person, mildly prefers having those hours to killing Hitler. However, because it's so morally important to kill Hitler, he only has a very

[162] Again I just modify slightly the formulation given in (Egan 2007, 97). This case was initially presented in Egan's paper, but was suggested to him by David Braddon-Mitchell.

mild preference for living those few hours at the cost of Hitler's survival. Now, what should Paul do?

It seems very plausible to me that he should push the button. But, if so, then, again, our intuitions about this case have switched merely by altering the stakes involved. Again, this is exactly what MDT predicts.

We can make this explanation more precise. Let's define the *relative stakes ratio*, in two option cases, as the ratio of (EEV(right action according to EDT) − EEV(wrong action according to EDT) to (CEV(right action according to CDT) − CEV(wrong action according to CDT)). If MDT is correct, then it's how this ratio changes that should affect our intuitions.

Now, I'd personally be roughly indifferent between a guarantee of $1000 and 1% chance of $1 million, so, given my preferences, the relative stakes ratio in *Standard Predictor* is approximately 99:1. *The Smoking Lesion* is taken to be more favourable to CDT than *Standard Predictor* is. So, if our intuitions roughly track MDT's recommendations, then we should expect the relative stakes ratio to be less than 99:1. And that's what we find. If I ask myself, for example, whether Susan would be willing to take up smoking even at the cost of *causing* a 1% chance of moving from the low-risk group for lung cancer (which non-smokers belong in) to the high-risk group for lung cancer (which regular smokers belong in), I imagine her being willing to take that cost. I imagine her only becoming indifferent at around 10%, suggesting that, when presented with the case, I intuitively assess the relative stakes ratio as only being about 10:1. This is in line with MDT's prediction.

The Psychopath Button is taken to be more favourable to EDT than *Standard Predictor* is. So, if our intuitions roughly track MDT's recommendations, we should expect the relative stakes ratio in *The Psychopath Button* to be greater than 99:1. And that's what we find. If I ask myself, for example, whether Paul would be willing to kill all psychopaths even at the cost of a 1% chance of *causing* his own death (perhaps he has a gun with 99 empty chambers but one loaded chamber, and pointing the gun at his head and pulling the trigger is the only way to kill all psychopaths), I imagine him not being willing to take that risk. I certainly wouldn't do it. *Even if* I thought it was ok to murder innocents for a greater good (!), and *even if* I thought that killing all psychopaths would be a net good, I still value my own life too much to make that sort of sacrifice. But if that's correct, then I have intuitively judged the relative stakes ratio to be greater than 99:1, which is what MDT predicts.[163]

One might object that, in the 'high-stakes' and 'stakes-adjusted' cases given above, we can explain the divergence in our intuitions by appeal to empirical uncertainty. According to this explanation, in *HSP-I* we get the one-boxing intuition because we can't really imagine ourselves to be certain that the Predictor will get it right purely

[163] Egan gives another case, the *Murder Lesion*. It seems to me that, again, the reason our intuitions in this case favour EDT is because of the relative stakes. However, the relative stakes don't seem to be quite as biased towards EDT as they are in *The Psychopath Button*. This might explain why *The Murder Lesion* is not as convincing a 'counterexample' to CDT as *The Psychopath Button* is. MDT allows us to explain a couple of other puzzles as well. First, Joyce (2012, 125) says that *The Psychopath Button* 'is not original with Egan,' because structurally similar cases were given in (Weirich 1985; Gibbard 1992; Pearl 2010). Joyce takes the shared structural similarity to be that all are cases where every option (prima facie) unratifiable according to CDT. But, if this is right, then why weren't these *earlier* cases taken to be grave counterexamples to CDT, in the way that some at least have taken *The Psychopath Button* to be? The answer is with the stakes. *The Psychopath Button* is *not* similar to the earlier cases with respect to the relative stakes ratio. And it's the relative stakes ratio that gives *The Psychopath Button* its bite. Second, MDT enables us to explain is why our intuitions in *Standard Predictor* seem to favour EDT significantly more if the Predictor is *infallible*, rather than merely highly accurate (as (Sobel 1988) discusses). The answer is twofold. First, increasing the probability of the $1 million further biases the stakes in EDT's favour. Second, as a matter of psychology, we tend to overvalue a "sure thing" (which is why, for example, the Allais paradox arises): so the move from 99% certainty to 100% certainty biases the stakes in favour of EDT by considerably more than merely the value of an additional 1% chance of $1 million.

through prediction. In any situation we can imagine, so the objection goes, there will remain some residual uncertainty that choosing the one box *causes* there to be a wish in the one box, and that's how the Predictor pulls off his trick. Similarly, in *Stakes-Adjusted Smoking Lesion*, perhaps we simply can't imagine ourselves not to have some credence that smoking causes bursting into flames. In either case, if we've got even small credence in that empirical hypothesis, then both CDT and EDT will recommend one-boxing, in *HSP-I* and not-smoking, in the Stakes-Adjusted Smoking Lesion. One can attempt an analogous explanation with respect to *HSP-II* and *The Stakes-Adjusted Psychopath Button*.

Speaking personally, my intuitions are sufficiently robust that we could replace the stick of gum with the lives of ten thousand children and I would have the same view that one should one-box in *HSP-I* and two-box in *HSP-II*. (The same is true for adjustments to the stakes in *Stakes-Adjusted Smoking Lesion* and *Stakes-Adjusted Psychopath Button*.) Given this, empirical uncertainty doesn't seem to be a very good explanation of my intuitions. However, there is a stronger response: which is that these extreme cases aren't strictly necessary to the use of MDT as an explanatory hypothesis for why our intuitions favour CDT in *The Smoking Lesion* and EDT in *The Psychopath Button*. All we need to show is that the relative stakes are more heavily biased towards CDT in *The Smoking Lesion* than they are in the *Standard Predictor*, and are more heavily biased towards EDT in *The Psychopath Button* than they are in the *Standard Predictor*. And that's exactly what my discussion of the relative stakes ratio accomplished.

So MDT seems to do well in terms of giving a rational grounding for our seemingly conflicting intuitions. In fact, I think that it's the best account of our intuitions in these cases that I know of.

Consider, in contrast, the response to *The Psychopath Button* suggested by James Joyce, as part of a defense of CDT.[164] The idea is that, as Paul decides to perform one action rather than another, he immediately gains evidence about whether he is a psychopath. Given his initial credences, pushing the button has the higher causal expected value. But as soon as he begins to decide to push the button, he gains evidence that he is a psychopath, and his credences should change. And with those new credences, not-pushing the button has the highest causal expected value. But as soon as he begins to decide to not-push the button, he gains evidence that he is not a psychopath, and suddenly pushing the button has the higher causal expected value again. Eventually, his credences over whether he's a psychopath or not end up in equilibrium, with the expected causal value of both pushing the button and of not-pushing the button as the same.

This response is interesting. However, It seems to me that my explanation of our intuitions is significantly better than Joyce's.[165] This is for three reasons, presented in order of increasing importance.

First, Joyce's account can't explain our intuitions in *HSP-I* and the *Stakes Adjusted Smoking Lesion*. In order to explain our intuitions in those cases he would have to appeal to some

[164] (Joyce 2012). His view is very similar to that of (Arntzenius 2008), and both draw heavily on work by (Skyrms 1990). What I will say in response to Joyce applies fairly straightforwardly to Arntzenius's view.
[165] Of course, my explanation is not *inconsistent* with Joyce and Arntzenius's explanation. What I'm questioning is not whether their account is *true* but whether it's a satisfactory explanation of our divergent intuitions across these cases.

other explanation. In contrast, my account can both take the intuitions at face value, rather than having to offer a speculative debunking argument, and can offer one unified explanation for our varying intuitions, rather than having to offer two distinct explanations.

Second, his account doesn't get the intuitions right in very similar cases. Suppose, for example, that the button is wired up to Paul's brain, so that as soon as he begins to intend to push the button, all psychopaths are killed. His beliefs therefore aren't able to achieve equilibrium. In this case, CDT really would recommend that he intend to push the button. But it seems that this minor alteration to the case doesn't affect our intuitive appraisal of what Paul should intend to do.

Third and finally, Joyce's account doesn't capture the intuition even in the original case. Once deliberational equilibrium is reached, pushing the button has the same expected value as not-pushing the button. But that's not capturing the intuition, which is clearly in favour of it being a *mistake* to push the button, rather than it being permissible to push the button. Joyce makes some very brief suggestions, based on the heuristics and biases literature, concerning why we might the intuition is not reliable in this case, and, in general, I am perfectly happy to sacrifice fit with the intuitive data for the sake theoretical elegance. But if we have an independently motivated explanation of why those intuitions are rational, then we should prefer that explanation to the debunking explanation, unless the debunking explanation is on very strong ground indeed. So we should prefer MDT's explanation to Joyce's.

Joyce's is not the only alternative explanation in town. Ralph Wedgwood has introduced 'Benchmark Theory,' which gets the right answer in both *The Smoking Lesion* and *The Psychopath Button*.[166] But it suffers from intuitive counterexamples too.[167] Johan Gustafsson proposes a decision theory that captures the intuitions in both *The Smoking Lesion* and *The Psychopath Button* cases.[168] But that proposal comes at a cost of considerable theoretical inelegance: importing an idea of "iterated general ratifiability" that does not seem independently motivated. Another potential explanation comes from Huw Price, who suggests we should understanding causality in subjectivist terms, so that evidential probability and causal probability are, despite appearances, the same.[169] Again, however, this comes at major theoretical cost, depending on the truth of particular positions in the metaphysics of both causation and free will.[170] In general, if an alternative account explains our divergent intuitions without using such heavy philosophical machinery, as the MDT account does, then we should prefer that alternative explanation.

In general, we already know (i) that we are decision-theoretically uncertain; and (ii) that expected utility theory is in general the best way to accommodate uncertainty. So, even independently of its ability to explain our divergent intuitions, we should think that there is an argument for thinking that MDT is true. It explains our divergent intuitions

---

[166] (Wedgwood 2013)
[167] See, for example, (Briggs 2010).
[168] (J. E. Gustafsson 2011).
[169] (Price 2012).
[170] Price himself acknowledges this, when he says: "As we have seen, the EviCausalist relies heavily on the idea that the epistemic viewpoint of an agent is distinctive in certain ways. Roughly, it requires that agents see their own actions as "uncaused," at least in the midst of deliberation about those actions. This not only binds the fate of the EviCausalist, at least in some sense, to that of free will; it also means, potentially even more uncomfortably, that EviCausalism becomes a rope that bins *causation* to the fate of free will — no problem, perhaps, if these notions turn out to share the same fate, but a problem if they do not." (2012, 536 original italics).

without using any ad hoc philosophical machinery There is therefore a strong argument via Occam's razor for preferring the MDT explanation to any other explanation that does not have the same independent plausibility.[171]

So I think that MDT provides the best explanation of our apparently inconsistent intuitions. This gives further evidence in favour of MDT's truth, and that our intuitions are at least roughly in line with MDT. Now let's turn to implications of this view, and see how it gives grounds to undermine the two best arguments in favour of EDT.

IV. Undermining the intuitive argument for EDT

One way to argue in favour of EDT is via appeal to cases. EDT looks appealing for people, like me, who think that you should one-box in the standard Newcomb problem,

[171] Might not MDT have formal problems that plausible first-order decision theories lack? The short answer is 'yes'. Here is a recipe to show that it does: Take some decision theory T with bad formal property X (perhaps, for example, T violates contraction consistency in an implausible way). Concoct a story where an agent ought rationally to have high credence in T, and envisage a situation where the decision is important according to T, but not terribly important according to the other decision theories in which the agent has credence, and where T violates formal property X. In such a case, MDT's recommendations will be the same as T's, and therefore MDT will also violate formal property X. However, given the context I find this sort of violation to be unproblematic: in fact, I find it to be exactly the right recommendation to make. The problem arises not out of MDT, but instead out of the fact that this agent is in the unfortunate situation where she has high credence in T. And it seems right to me that, if an agent is in the unlucky situation where she has high credence in a theory with bad formal properties, she ought (in the meta decision theoretic sense) to act in a way that exhibits those problematic properties (violating contraction consistency, or whatever). (As an analogy, consider subjective utilitarianism, according to which one ought to maximize expected wellbeing. It is no objection to the view that sometimes it recommends pointlessly killing one's spouse, simply because there are unfortunate cases where an agent should have high credence that their spouse is the Devil, even though she isn't.) Now, one could construct cases where even small credence in T has the result that MDT violates the formal property (because the situation is high-stakes according to T but very low-stakes according to all other decision-theories). But those would, by their nature, be fringe cases. In general, because meta decision theories are less idealized than first-order decision theories it seems inappropriate to hold them up to the same formal standards as first-order decision theories, especially when the bad formal properties arise merely *out* of credence in first-order decision theories with those properties. So I do not think that this constitutes a compelling reason to reject the idea of MDT.

and for those who are particularly concerned by *The Psychopath Button*. So it looks like EDT it at least fairly well supported by the intuitive data.

Considerations of decision-theoretic uncertainty undermine this argument. The relative stakes ratio in the standard Newcomb problem depends on one's level of risk-aversion with respect to money, but for any normal agent is heavily biased in favour of EDT. For me the relative stakes ratio is approximately 99:1, so if I had only 2% credence in EDT, then, by MDT's lights, I should one-box in the standard Newcomb problem. So, far from providing an argument for thinking that the evidential approach is the best approach, the intuitions merely show that we should have at least a small credence in EDT. Indeed, because intuitions in the standard Newcomb case are unclear, and because the Smoking Lesion favours CDT even though in that case the stakes are still biased towards EDT, it seems that the credence in EDT that is warranted by appeal to intuitions about particular cases is not very large at all.

The fair way to adjudicate, on intuitive grounds, between EDT and CDT, would be to consider cases where the stakes are evenly balanced. Such a case would look as follows:[172]

	Money in both boxes	Money in one box only
Take one box only	\$20	\$0
Take both boxes	\$30	\$10

[172] I use small amounts of money, so that we can safely assume that utility is approximately linear with respect to money in this case.

Even I — who used to self-identify as a stark-raving one-boxer[173] — get almost no intuition in favour of one-boxing in this case. So EDT is not the intuitive view. In fact, I only start to get one-boxing intuitions once the amount that might be in the opaque box is twenty times as great as the amount that is certainly in the transparent box. So, as far as the argument from intuition goes, I should have no more than a small credence in EDT. So the intuitive argument for EDT is far weaker than it would have first seemed.

Appeal to intuitions has been used as one major argument in favour of EDT. The other argument is the "Why Ain'Cha Rich?" argument. Let's consider that now.

V. A Counterexample to "Why Ain'Cha Rich?"

When I introduced meta decision theory, I used unconditional credences. But we could formulate both evidential and causal versions of meta decision theory. According to causal meta decision theory (CMDT) we should maximize causal meta expected value (CMEV) where:

$$CMEV(A) = \sum_{i=1}^{n} C(A \Rightarrow D_i)\, D_i(A)$$

Again using '⇒' to denote the counterfactual conditional. It should be clear that for all A, D, $\Pr(A \Rightarrow D) = \Pr(D)$. Acting one way rather than another can't affect which decision theory is true. So nothing is lost by simply using unconditional credences:

[173] Anecdote: when I was first presented with *The Smoking Lesion* case I thought it was supposed to be an argument *in favour* of EDT.

$$CMEV(A) = \sum_{i=1}^{n} C(D_i)\, D_i(A)$$

In contrast, according to evidential meta decision theory (EMDT) we should maximize evidential meta expected value (EMEV), where:

$$EMEV(A) = \sum_{i=1}^{n} C(D_i|A)\, D_i(A)$$

These two views will almost never come apart: it's a very rare situation when acting one way or another gives you evidence for one decision theory rather than another. But it's not impossible for the two to come apart. And if we look at those admittedly rare cases we can construct a counterexample to the "Why Ain'cha Rich?" argument.

According to the "Why Ain'cha Rich?" argument, the average return of one-boxing exceeds the average return of two-boxing. Moreover, everyone can *see* that the average return of one-boxing exceeds the average return of two-boxing: so one-boxing foreseeably gives us more of what we want than two-boxing does. And, so the argument goes, a decision theory cannot be correct if it recommends an option that foreseeably gives you less of what you want than some other option does. So CDT can't be correct.

In response, the defender of CDT can say that Newcomb's cases are unusual: these are cases where a devious person has set things up to reward irrational behavior. So it isn't surprising that irrational people like those who act in accordance with EDT end up richer. However, to date the defender of CDT hasn't been able to come up with a

convincing case where one gets rewarded for *not* following EDT.[174] And that seems problematic.

However, if we are comparing CDMT and EMDT, things are different. Once we allow decision-theoretic uncertainty into the picture, we can construct a case where performing the action EMDT recommends foreseeably makes one poorer. So there is no longer an asymmetry between the causal and evidential approach, and the "Why Ain'cha Rich?" argument loses its force. Here's the case:

The Meta Newcomb Problem

Sophie faces two boxes, as follows:

	Wishes in both boxes	Wishes in one box only
One box	2 million lives[175]	0 lives
Two box	3 million lives	1 million lives

Sophie's beliefs are as follows. She has 51% credence in EDT and 49% credence in CDT. Before taking her action, she is almost certain that there are wishes in both boxes. However, *conditional* on her two-boxing, she is almost certain that there is a wish only in the transparent box.

Given these credences:

[174] (Lewis 1981) argues that it's impossible. In his introduction to decision theory, Weatherson summarizes the literature on this as follows: "it turns out to be very hard, perhaps impossible, to construct a problem of this sort for evidential decision theorists." (p.89; available at brian.weatherson.org). Arntzenius (2008) has proposed an example, but it is debatable whether the example is successful or is even coherent. See (Ahmed and Price 2012) for discussion of that case.

[175] Again using "lives saved" rather than dollars because linear value over number of lives saved is more plausible than linear value over dollars.

V_{EDT}(One Box) = ~1 * 2 million lives + ~0 * 0 lives = ~2 million lives

V_{EDT}(Two Box) = ~0 * 3 million lives + ~1 * 1 million lives = ~1 million lives

V_{CDT}(One Box) = ~1 * 2 million lives + ~0 * 0 lives = ~2 million lives

V_{CDT}(Two Box) = ~1 * 3 million lives + ~0 * 1 million lives = ~3 million lives

So the meta decision problem looks as follows:

	Value, given EDT	Value, given CDT
One box	~2 million	~2 million
Two box	~1 million	~3 million

However, Sophie places great weight, epistemically, on what people actually do in Newcomb cases (rather than what people claim their intuitions are about what they would do in such cases). She thinks that the actions of typical human agents in Newcomb cases provide very good evidence in favour of CDT or EDT. And she believes she is a typical human agent. So how she acts will affect her credences in the two decision theories. If she one-boxes, she will update in favour of EDT, and come to have 52% credence in EDT and only 48% in CDT. If she two-boxes, she will significantly update in favour of CDT, and come to have 60% credence in CDT and 40% credence in EDT.

What should Sophie do? To answer this, let's work out the expected values. We have:

V_{CMDT}(One Box) = 0.51×2 + 0.49×2 = 2

$V_{CMDT}(\text{Two Box}) = 0.51\times1 + 0.49\times3 = 1.98$

So CMDT recommends one-boxing. And we have:

$V_{EMDT}(\text{One Box}) = 0.52\times2 + 0.48\times2 = 2$

$V_{EMDT}(\text{Two Box}) = 0.4\times1 + 0.6\times3 = 2.2$

So EMDT recommends two-boxing.

So if we take into account decision-theoretic uncertainty, then the causal theory can tell you to one-box while the evidential theory tells you to two-box. In the above case, if Sophie follows EMDT she *foreseeably* ends up saving fewer lives than if she follows CMDT. So, unlike at the first order, one cannot construct a "Why Ain'cha Rich?" argument in favour of EMDT over CMDT.

This makes it seem very plausible that the correct *meta* decision theory is causal. But we can go a bit further than that. It would seem odd if the correct meta decision theory were causal while the correct first-order decision theory is evidential. It seems plausible that our views about which variety of first-order decision-theory is correct and which variety of meta decision theory is correct should be at least roughly coherent. So evidence about which meta decision theory is true seems also to give evidence about which first-order decision theory is true. So, even if we can't construct a counterexample to "Why Ain'cha Rich?" for EDT, the fact that we can construct such a counterexample for EMDT weakens, at least to some degree, the "Why Ain'cha Rich?" argument in favour of EDT.

So, as well as being able to provide an explanation of our divergent intuitions across cases, considerations relating to meta decision theory allow us to generate novel arguments against EDT. So meta decision theory (and the idea of metanormativist more generally) seems to be a powerful tool in the causalism versus evidentialism debate. In the next section, however, I'll consider what I consider to be the biggest objection to the very idea of meta decision theory.

VI. A Vicious Regress?

The previous discussion concerned what one ought to do when one is decision-theoretically uncertain. But there's an obvious problem looming. Though I think that the arguments in favour of CMDT are compelling, I still think that one shouldn't be *certain* in that view. One should still retain some credence in EMDT, and one should probably have credence in other meta decision theories as well, such as "Act in accordance with the decision-theory that you think is most likely to be right." But if so, then don't we need a metametatheory, to govern how to act given uncertainty in meta decision theories, and a sense of 'ought' that is relative to uncertainty across meta decision theories? And if we do propose such norms, a decision-maker presumably shouldn't be certain in *them*. So we need a metametametatheory… and so on *ad infinitum*.

This is clearly an important issue, but we need to be clear on what exactly the *objection* to MDT is. As I understand it, the objection is that MDT is *arbitrary*. MDT takes only first-order normative uncertainty into account. But there seems to be no reason for a decision theory to merely take first-order normative uncertainty into account, rather

than second-order normative uncertainty as well. But if so, then one should take third-order normative uncertainty into account as well. But every time one takes into account a further level of uncertainty, one's theory will look just as arbitrary as before. There is no non-arbitrary stopping point for attempts to take normative uncertainty into account. And so, the objection might go, this is a *reductio* of the whole metanormativist project.

It strikes me, however, that this line of argument proves too much. If we wish to say that meta decision theory is arbitrary, then we should also say that first-order decision theory is arbitrary: after all, it seems arbitrary for decision theory to factor in one type of an agent's uncertainty, but not another. The only genuinely non-arbitrary place to stop is to take into account *none* of the agent's uncertainty, and for rational norms to recommend simply what it's actually best for the agent to do (by her lights), whether or not she's in a position to work out what option is actually best by her lights. But then it seems to me that we've given up on decision theory altogether.

A second response to the regress worry is that there *is* a non-arbitrary place for the decision-maker to stop. At some point up the regress, the decision-maker might *know* what she ought (in the relevant sense of ought) to do. For example, consider someone who is only uncertain between CMDT and EMDT. In almost all cases, these two theories agree: it was only in recherché examples like the *Meta Newcomb Problem* where the decision-maker's conditional probabilities across first-order decision theories differ from the decision-maker's unconditional credences across first-order decision theories. And I think it's reasonable to claim that we know that, when all decision theories in which one has credence claim one ought to do A, one ought (in the relevant sense of 'ought') to do A. So, in almost all cases where a decision-maker is uncertain merely

between CMDT and EMDT, she will know what she ought to do, because the two theories are in agreement.

Now, we can dream up cases where there *isn't* convergence between the theories in which one has credence; I give such a case in the appendix. But in those cases, it seems to me exactly right to say that there's no fact of the matter about what is most rational for the decision-maker to do (though there still might be facts about what she ought *not* to do). In such cases, the rational choice-worthiness ordering will be incomplete. But that does not seem terribly problematic. We should already expect the rational choice-worthiness ordering to be incomplete if we believe that there are incommensurable values, or if we have the plausible view that there is no rational way of deciding between different varieties of the Pasadena game.[176] In this case, the incompleteness is just symptomatic of a decision-maker being in an odd epistemic situation where there's no way of aggregating her uncertainty such that one option comes out as uniquely rational (or such that some number of options come out as equally and maximally rational). Given that we're trying to give norms that apply to epistemically limited agents like ourselves, we should expect outcomes like that to occur.

VII. Conclusion

In this paper I've argued in favour of one particular brand of metanormativism: that there are decision-theoretic norms that are relative to decision-theoretic uncertainty. Moreover, MDT allows us to neatly explain the apparent divergence in our intuitions

[176] See (Nover and Hájek 2004) for a description of the Pasadena game.

between the *Standard Predictor*, *The Smoking Lesion*, and *The Psychopath Button*. It undermines both the intuitive argument in favour of EDT, and, to some extent, the "Why Ain'cha Rich?" argument as well. And it does all this without having to make invoke any complex refinements to our favoured decision-theories. Considerations of decision-theoretic uncertainty are therefore a powerful tool for use in debates between causal and evidential decision theory — a tool that gives the causal approach a significant new advantage.

Appendix: *Oscillating Rationality*

In this appendix I give an example where, as we progress up higher orders of decision-theory, what one ought (in the relevant sense) to do oscillates between two options, A and B. The key to the example is that it is risky for the decision-maker to follow the dictates of a risk-averse theory, if the decision-maker's credence in that theory is low. It's like the risk-averse theory is, to some extent, saying: 'Don't listen to me!' Here's the case.

Sophie is in a situation of uncertainty over which meta decision theory is true. She has 18/23 credence in *Risk Neutral*, according to which one ought to maximise expected choice-worthiness over all first-order decision-theories; and 5/23 in *Risk Averse*, according to which one ought to maximise the expectation, over all first-order decision-theories, of the square root of the difference in choice-worthiness between the option in question and the worst option in the option set. (The distinction between causal and evidential decision theory does not matter here.) Moreover, she has exactly the same credences in the analogous meta meta decision theories, *Risk Neutral*$_2$, according to which one ought to maximise expected choice-worthiness over all meta order decision-theories; and 5/23 in *Risk Averse*$_2$, according to which one ought to maximise the expectation, over all meta decision-theories, of the square root of the difference in choice-worthiness between the option in question and the worst option in the option set. And the same is true for meta meta meta decision theories, and so on ad infinitum.

She has two options available to her, A and B. She faces a meta decision problem, as follows:

	Risk Neutral$_1$ *(C=18/23)*	*Risk Averse*$_1$ *(C=5/23)*
A	1	0
B	0	4

B is the option with the highest expected choice-worthiness, so she ought to choose B (in the meta meta sense of 'ought'). A, however, is the option with the highest expectation of the square root of choice-worthiness. The difference between the expected choice-worthiness of B and the expected choice-worthiness of A is four times smaller than the difference between the expectation of the square root of choice-worthiness of A and the expectation of the square root of choice-worthiness of B.[177] At the second order, therefore, her decision situation looks as follows.

	Risk Neutral$_2$ *(Cr=18/23)*	*Risk Averse*$_2$ *(Cr=5/23)*
A	0	4
B	1	0

[177] The expected choice-worthiness of B is 20/23; the expected choice-worthiness of A is 18/23. The expectation of the square root of choice-worthiness of (A minus the worst option) is 18/23; the expectation of the square root of choice-worthiness of (B minus the worst option) is 10/23. That is: the difference in choice-worthiness between A and B is four times greater according to *Risk Averse*$_2$ than the difference in choice-worthiness between A and B is according to *Risk Neutral*$_2$. Even with intertheoretic comparability of units of value, transformations of each individual choice-worthiness function by an absolute value are permissible, and transformations of all choice-worthiness functions by a multiplying factor are permissible. So in the following table I have renormalized the scales, putting each theory's least choice-worthy option at 0 and simplifying the units.

The situation at this order of rationality is the mirror of the situation at the lower order of rationality. *Risk Neutral*$_2$ claims that B is more choice-worthy than A; whereas *Risk Neutral*$_1$ claimed that A was more choice-worthy than B. *Risk Averse*$_2$ claims that A is more choice-worthy than B; whereas *Risk Averse*$_1$ claimed that B was more choice-worthy than A. And the ratio of the differences between the two orderings, at this level, is the reciprocal of what it was at the previous level.

Because of this, once we move up to the uncertainty at the third order (that is, meta meta meta uncertainty), the situation flips back to a situation isomorphic with the situation that we encountered at the first order. At this order, A has the highest expected choice-worthiness, and so A is the rational option. So Sophie's decision-problem looks as follows:

	Risk Neutral$_3$	*Risk Averse*$_3$
A	1	0
B	0	4

This will generate, at order 4, a decision-situation isomorphic with the decision-situation at order 2. And so on. As we progress up the different level of rationality, what it is rational to do at that level oscillates between the two options A and B.

This example shows that there can be an infinite decision-theoretic regress, where there's no level after which you keep getting the same option as being maximally choice-worthy at higher levels.

Chapter 7: The Value of Moral Philosophy

Introduction

What's the value of moral philosophy? How important is it, for example, for large organisations to consult texts in moral philosophy before making a high-stakes decision? How important is it for an individual to study ethics before embarking upon their career? And how important is it for we, as philosophers, to continue our research, rather than to do something else with our time — or for governments to continue to fund our research, rather than spend the money on healthcare or unemployment benefits?

In this chapter, I introduce a framework for answering this question. Using this framework, I suggest that the value of both studying and doing further research into moral philosophy is greater than one might expect, and certainly greater than the low value implicitly given to it by the actions of real-world individuals.

In section I, I explain how we should assess the value of empirical information, and explain how we could extend this to the case of moral information. In section II, I give three illustrative case studies: the choice of how a large foundation should spend its resources, the choice of career for an individual, and academic research into moral philosophy. In section III, I show how the value of moral information gives one reasons to keep one's options open, using risk of human extinction as a case study. In section IV, I consider and respond to objections.

Before I begin, I should highlight that in what follows I use the unusual term 'moral information'. I use this term in the hope of remaining almost entirely non-committal on the issues of moral epistemology and moral metaphysics: as I understand it, something is a piece of moral information iff coming to have it should make one alter one's beliefs or one's degrees of belief about at least some moral proposition. So the term 'moral information' could apply to experiences, arguments, intuitions, or knowledge of moral facts themselves.

I. Value of information, empirical and moral

In this section I'll explain how one should calculate the value of information for empirical matters. One can work out the value of *perfect* information — that is, the value of coming to know some particular proposition for certain — and the value of *imperfect* information, which is the value of improving one's evidence base but not coming to know any additional proposition for certain. I'll begin by discussing the simpler concept of the value of perfect information. I'll illustrate the idea of perfect information with recourse to the 'newsboy' example, a common example from decision analysis,[178] before discussing the idea in general.

Newsboy

> Jonny sells newspapers. He sells each newspaper for $1. On each day, he has the option of either buying 50 newspapers, for $15, or buying none – he can't buy any intermediate number. The number of newspapers that he sells in any one

[178] This version is borrowed from (Eeckhoudt and Godfroid 2000).

> day, when he tries to sell newspapers, is either 0 or 50, and never anything in between, and today he doesn't know how many newspapers he'll be able to sell if he tries. We'll assume that he doesn't value his time at all: all Jonny cares about is making money. And we'll assume that the value of additional dollars for Jonny is linear over this amount. Our question is: how much should Jonny be willing to pay in order to know for certain how many newspapers he'll be able to sell if he tries?

According to the standard analysis,[179] he should answer this question as follows. First, he should work out how many newspapers he expects to sell, given his current evidence. Let's suppose that he thinks there is a 50/50 chance of selling either 0 newspapers or 50 newspapers, and no likelihood at all of something in the middle. Second, he should work out the expected value of his options, given his current evidence. In this case, the expected value of buying 0 newspapers is $0. The expected value of buying 50 newspapers is 0.5($50) – $15 = $10. The expected value of buying the newspapers is higher than the expected value of not buying newspapers. So, given his current evidence, he should buy the newspapers.

Third, he should work out the additional value of gaining the new information. If he finds out that 50 people will buy the newspapers, then the additional information has no value for him: he would not change his behaviour with this new information, and so he would have made the same amount of money even without the new information. However, if he were to find out that 0 people will buy newspapers, he would change his behaviour: he would decide against paying $15 for 50 newspapers. So, if it is the case

[179] See, for example, Raiffa (1968).

that 0 people will buy newspapers, the value for Jonny of finding that out is $15. Jonny thinks there is a 50% chance that he will find out that 50 people will want to buy newspapers (which would have no value for him), and a 50% chance that he will find out that 0 people will buy newspapers (which would be worth $15). So the expected value of gaining that new piece of information is 0.5(0) + 0.5($15) = $7.50. This gives the amount up to which he should be wiling to pay to gain that information.

When dealing with the value of information, there are some important points to note. First, as one might have noticed from the above, gaining new information is *only* valuable on our analysis if there is some chance that one will change one's behaviour. If Jonny were merely uncertain about whether 50 people or 40 people would buy newspapers, then there is no value for him in gaining additional information, because it would be rational for him to purchase the newspapers either way. Similarly, if Jonny knows that he is just pig-headed and lazy, and will buy no newspapers no matter how rational it is for him to do so, then, again, gaining new information would have no value for him. In reality, factors such as peace of mind and the intrinsic value of having more accurate beliefs can mean that it can be rational to gain new evidence even if one will not change one's behaviour. But for simplicity, and because it would only strengthen my argument if we incorporated these details, I leave these to the side in my analysis.

Second, note that the *worth* of information is very different from how much one actually has to pay for that information. Perhaps Jonny could find out the demand for newspapers merely by asking someone on the street, costing him nothing. In which case, he simply had a bargain — but the amount he had to pay does not change the fact that the information was worth $7.50 (and that, if he had no better option, that he

should have been willing to pay up to $7.49 to receive it). Third, the higher stakes a decision is, the greater the value of information. To illustrate, suppose in the case above that we multiplied all the monetary values by 10: each newspaper sells for $10, but Jonny has to pay $100 in order to buy the 50 newspapers. In which case, the value of information for Jonny would have the same proportional change, increasing to $75.

The above method for calculating the value of additional information is intuitively appealing and widely accepted within decision analysis. But, to my knowledge, it has only ever been used to work out the value of gaining new empirical information: that is, information about how the world will pan out. One particular evaluation of all possible states of the world is always presupposed. But we shouldn't be certain about how to evaluate all possible states of the world, and we should change our moral views in response to new arguments and ethical discussion. So it seems that we should be able to apply value of information analysis to changes in moral views as well as changes in empirical views. In what follows, I'll give three examples, in increasing complexity, illustrating some applications of this value-of-information analysis to moral information.

II. Three examples

A philanthropic foundation

Our first example provides the simplest illustration of the value of moral information. Let us suppose that the leader of major philanthropic foundation is deciding how to allocate her resources. She can either spend her resources on international development, which would prevent 1.5 million human lives being lost to malaria,

costing $2.25 billion. Or she can spend those resources on campaigning to improve animal welfare, which would prevent one billion chickens from being brought up in factory-farmed conditions. Now let's suppose that decision-maker is pretty sure of the value of saving one human life, and let's say that saving one human life has 1 unit of value, so that the current expected value of policy A is 1.5 million. The decision-maker is extremely uncertain about the value of benefitting chickens: being in a poor epistemic situation, she is 99% certain that benefitting chickens is of no value, but has 1% credence that preventing one chicken from living in factory farmed conditions has one-tenth the moral value of saving one life from malaria. So policy B has an expected value of 1 million. So, given her current epistemic state, she should choose policy A.

Now suppose that the decision-maker has the option of gaining perfect information about the relative value of benefitting chickens and saving humans. What's the value of this information? Well, she should think that there's a 99% chance of finding out that benefitting the chickens is of no value, so gaining this information is 99% likely not to change her behaviour, and therefore have no value (at least, within the context of this decision). But she should think that there's a 1% chance that she will learn that benefitting chickens is of value: if this happened, then the expected value of policy B would become 100 million. The additional benefit she would produce given this outcome would therefore be 98.5 million. Multiplying the value of this outcome by its probability gives us the expected value of the information, and so the expected value of gaining this information is equivalent to the value of saving 985,000 lives for certain. So she should be willing to pay $1.4 billion (that is, 985,000 × $1500) in order to gain this information before making her decision about where to spend the $2.25 billion. In the

above calculation her starting budget was not relevant: in general, she should be willing to spend about 38% of her budget in order to know how she ought to spend the remaining 62% (so if her budget was fixed at $2.25 billion, then she should be willing to spend $855 million in order to find out how to spend the remaining $1.395 billion). Though in order to construct the case I had to invent some numbers, I hope they weren't a completely unreasonable idealisation of commonly held degrees of belief, and I chose the numbers to be accurate with respect to the empirical facts.[180] So the above example shows that it's at least possible for the value of moral information to be remarkably high, which is notable given that philanthropists (or other similar actors, like governments) typically spend almost nothing on gaining new moral evidence.

Now, the above case described the value of gaining perfect moral information on a particular issue. This is of course unrealistic, as we should never become fully certain in a particular moral view. But the issues remain just the same, though slightly more complicated, if one gains imperfect information. In general, because information brings about a proportional change in the expected value of the options under consideration, if you're dealing with extremely high-stakes issues, then the value of information becomes extremely high as well.

Career choice

As well as spending money to gain new moral information, one can also spend time gaining new moral information. This is relevant, for example, to the question of how

[180] On the basis of information provided by GiveWell and Effective Animal Activism, two charity evaluators.

much time young people should be willing to spend studying ethics before choosing which career to pursue. Again, I'll give an example to illustrate. Consider Sophie. She comes from a poor family in the USA, but is very bright and hardworking, and won a scholarship to a top university. She's undecided about what career to pursue. She could become an NGO worker, and through that save the lives of 100 people in the developing world, but it would mean that she could not give back to her family at all. Or she could become a lawyer: this would not benefit those in developing countries at all, but would mean that she could pay for health insurance and better living condition for her extended family, improving the overall lives of each of 22 of her family members by 30%. She therefore realises that she can benefit those in the developing world much more than she can benefit her family. But she isn't sure how to weigh those respective benefits. She's 95% confident that it's 100x more important to benefit her family, but has 5% credence remaining that it's just as important to benefit those in the developing world as it is to benefit her family. Given this, how much time should she be willing to spend studying ethics in order to get perfect information about how to value benefits to her family compared with benefits to those in the developing world?

In what follows, I'll suppose that saving one life in the developing world, according to the partial view, is worth 1 unit of value, and that benefitting someone's life by 30% provides 0.3 times as much benefit as saving someone's life. Given her current beliefs, Sophie should choose becoming the lawyer: the expected value of doing so is $0.3\times22\times100 = 660$ units of value, whereas the expected value of becoming the NGO worker is $(0.95\times100 + 0.05\times100\times100) = 595$ units of value. But she also has the option of getting more moral information: she could take years out before university in order to

study moral philosophy. How many years should she be willing to spend studying in order to get perfect information about how to weigh benefits to her family against benefits to those in the developing world?

She should think it 95% likely that she wouldn't change her decision. But she should think it 5% likely that she would, because she should be impartial between distant strangers and her family, and that by choosing to become the NGO worker she would increase the value of her career by $100\times100 - 0.3\times22\times100 = 9340$. So the value of this information is $0.05\times9340 = 467$. So she should be willing to lose out on 467 units of value in order to gain perfect information about how to spend her 40 years. Assuming that the benefit to her family is linear over a 40-year career, she produces $(0.3\times22\times100)/40 = 16.5$ units per year. So she should be willing to spend $467/16.5 = 28.3$ years in order to get perfect information about how to spend those 40 years. In general: she should be willing to spend $28.3/(28.3+40) = 41.4\%$ of her time to gain perfect information about how to spend the remaining 58.6%. So, if she only had those 40 years to spend, she should be willing to spend a little over 16 of them studying ethics in order to get perfect information about what she should do with the remainder of her career.

Again, this example shows that the value of additional moral information *can* be extremely high: I tried to choose credences and impacts that Sophie could have that weren't at least completely unrealistic. But the thought, at least, that it could be worth anyone spending 16 years of the life studying ethics just so that they make a better decision at the end of that time is surprising. Indeed, for most non-philosophers I

imagine that the thought that one should spend *any* time studying ethics before making major life decisions would be surprising.

Of course, in the above case the conclusion is not that Sophie actually should spend 16 years studying ethics. Again, we need to distinguish the *worth* of gaining moral information from how much time it would actually take to get that information. Perhaps Sophie would learn everything she needs to after only a few years' of study. In which case she should only spend a few years studying. But that does not diminish the value of those few years of study — it just means that, for those few years, she is getting a bargain, evaluatively speaking. A second caveat, when it comes to how much of time the typical person should spend studying ethics, is that the above assumption that the benefit Sophie would produce is linear over a 40-year career will likely often be inaccurate. It seems plausible that the benefit one produces in one's career increases dramatically over the course of one's life, as one gets promoted, and becomes more experienced and more influential. In which case, insofar as studying ethics pushes back one's career, thereby taking years off the end of one's career, the cost of studying ethics is higher than the above calculation would suggest. And one can lose career options by studying ethics for too long, providing another reason against too many years of study.

But even despite these caveats, as with the previous case it seems plausible that the value of gaining new moral information is higher than one might expect. It seems perfectly plausible that being in a better epistemic state with respect to the moral facts can mean that one does ten times as much good in the rest of one's life as one would otherwise have done (e.g. perhaps one focuses on climate change mitigation rather than a domestic policy issue because one comes to believe that future people are much more

important than one had thought). In which case it would be worth spending half one's working life studying ethics in order to improve how one uses the remaining half — even if 80% of the value that one contributes to the world occurs in the latter half of one's career.

Research in moral philosophy

The third illustration is that of doing research into moral philosophy. Like the previous two examples, value-of-information reasoning suggests that the value of research into moral philosophy has the potential to be remarkably high. The numbers I use will be speculative and intended to be illustrative merely. But the basic argument for this is as follows:

1. On at least some plausible moral theories, astronomical amounts of value lie in the future.

2. Of those moral theories, different theories disagree on how to obtain those astronomical amounts of value.

3. So the importance of making sure that we act on the right theory, rather than the wrong theory, is itself astronomical.

4. Moral philosophy has a non-negligible chance of making sure that we act on the right theory, rather than the wrong theory.

5. So the value of moral philosophy is very high.

In support of (1) and (2), consider that different conceptions of value can differ radically in terms of how they envisage the best possible future. According to hedonism, the best long-run outcome for the world might involve people minimising their working time so that they can spend the majority of their lives taking bliss-inducing hallucinogens. According to objective list theories of the good, the best long-run outcome for the world might involve discovering the fundamental truths about reality and producing great works of art. In the objective list theorist's eyes the hedonist's perfect world might be of very little value, compared to objective list's theorist's perfect world. And the same might be true vice-versa for the hedonist. So if humanity converges on the wrong account of value, then almost all the potential value of the future might be lost.

The above reasoning is at least suggestive of the idea that governments of the world should sometimes be willing to spend a significant proportion of their resources on doing and promoting new ethical research.[181] A harder question is what the value is for someone to go into moral philosophy and do research. One might naturally think that the chance of additional research in moral philosophy making the difference to whether we achieve a high-value future is so small, given the world as it is, that there is little value of producing additional research in moral philosophy. But it's at least unclear that that fact is strong enough to diminish the value of new research into moral philosophy. Moral philosophy has had its successes in the past, especially over the long run.[182] And, moreover, the chance of individual success does not need to be large in order for the

[181] Again with the note that what they should be willing to spend is not they same as what they should spend, depending on how much it costs to gain such information.

[182] Fairly uncontroversial cases include Locke's influence on the American Revolution, Mill's influence on the woman's suffrage movement; Marx's influence on the rise of socialism and communism; and Singer's influence on the animal rights movement. If we broaden our horizons, and include Aristotle, Confucius and Buddha in our comparison class, as I think we should, then it's hard to deny that the work of moral philosophy has shaped millennia of human history.

expected value to be extremely high. The situation is analogous to voting: in pursuing the activity, one has a very small chance of making any difference; but the activity is still of great value because the difference one does make, if one makes the difference, is huge. If we at least know about the true moral theory, there is some chance that our descendants will act in accordance with it. But if no-one ever discovers the true moral theory, the chance is of us achieving something close to the best possible future is close to zero. So I think that there is a reasonable, though certainly not indisputable,[183] case to be made that the information-value of producing additional moral philosophy, even given the way the world currently is, is very significant indeed.

Summing up on these examples

In each of the above examples, we have found the value of gaining new moral information to be remarkably high. The reason is as follows. First, out decisions are often highly sensitive to changes in our moral assumptions: slightly different moral assumptions can radically alter the value of the options available to us. But, second, at least some of these assumptions typically are not particularly robust: there is a good chance that further ethical enquiry or research will give us information that should make us change our view. Finally, the value of information grows with the stakes of the decision, and moral issues are often greatly relevant to very high-stakes decisions. So when the stakes are as high as how one should spend one's working life, how a

[183] In particular, one might reasonably have worries about the applicability of expected value reasoning when it comes to very small probabilities. See, for example (Bostrom 2009).

foundation should spend its money, or how the future of the human race should progress, the value of moral philosophy gets very high indeed.

These examples considered the value of 'purchasing' moral information, whether through studying ethics or doing research directly. But value-of-information analysis has another subtle but important implication.

IV. Option Value

This implication is the value of *keeping options open*. Illustrating what I think is the most important application of this idea, let's consider the question of how to evaluate the possible extinction of the human race.

The human race might go extinct from a number of causes: asteroids, supervolcanoes, runaway climate change, pandemics, nuclear war, and the development and use of dangerous new technologies such as synthetic biology, all pose risks (even if very small) to the continued survival of the human race.[184] And different moral views give opposing answers to question of whether this would be a good or a bad thing. It might seem obvious that human extinction would be a very bad thing, both because of the loss of potential future lives, and because of the loss of the scientific and artistic progress that we would make in the future. But the issue is at least unclear. The continuation of the human race would be a mixed bag: inevitably, it would involve both upsides and downsides. And if one regards it as much more important to avoid bad things

[184] Scientific discussion of such risks is provided in (Leslie 1998; Posner 2004; Rees 2003).

happening than to promote good things happening then one could plausibly regard human extinction as a good thing.

For example, one might regard the prevention of bads as being in general more important that the promotion of goods, as defended historically by G. E. Moore,[185] and more recently by Thomas Hurka.[186] One could weight the prevention of suffering as being much more important that the promotion of happiness. Or one could weight the prevention of objective bads, such as war and genocide, as being much more important than the promotion of objective goods, such as scientific and artistic progress. If the human race continues its future will inevitably involve suffering as well as happiness, and objective bads as well as objective goods. So, if one weights the bads sufficiently heavily against the goods, or if one is sufficiently pessimistic about humanity's ability to achieve good outcomes, then one will regard human extinction as a good thing.[187]

However, *even if* we believe in a moral view according to which human extinction would be a good thing, we still have strong reason to prevent near-term human extinction. To see this, we must note three points. First, we should note that the extinction of the human race is an extremely high stakes moral issue. Humanity could be around for a very long time: if humans survive as long as the median mammal species, we will last another two million years.[188] On this estimate, the number of humans in existence in

[185] (1903, 212, 222).
[186] (2010).
[187] There are other grounds, too, for being uncertain about the value of human extinction. For example, John Broome has suggested that, on the grounds of his critical-level population ethics, he at least does not know whether human extinction would be good or bad (2010).
[188] This way of estimating future human population is suggested by (Matheny 2007). Other speculative estimates of future population are given by (Bostrom 2003; Nick Bostrom and Cirkovic 2008; Adams 2008).

the future, given that we don't go extinct anytime soon, would be 2×10^14.[189] So if it is good to bring new people into existence, then it's *very* good to prevent human extinction.

Second, human extinction is by its nature an irreversible scenario. If we continue to exist, then we always have the option of letting ourselves go extinct in the future (or, perhaps more realistically, of considerably reducing population size). But if we go extinct, then we can't magically bring ourselves back into existence at a later date.

Third, we should expect ourselves to progress, morally, over the next few centuries, as we have progressed in the past. So we should expect that in a few centuries' time we will have better evidence about how to evaluate human extinction than we currently have.

Given these three factors, it would be better to prevent the near-term extinction of the human race, *even if* we thought that the extinction of the human race would actually be a very good thing. To make this concrete, I'll give the following simple but illustrative model. Suppose that we have 0.8 credence that it is a bad thing to produce new people, and 0.2 certain that it's a good thing to produce new people; and the degree to which it is good to produce new people, if it is good, is the same as the degree to which it is bad to produce new people, if it is bad. That is, I'm supposing, for simplicity, that we know that one new life has one unit of value; we just don't know whether that unit is positive or negative. And let's use our estimate of 2×10^14 people who would exist in the future, if we avoid near-term human extinction. Given our stipulated credences, the expected benefit of letting the human race go extinct now would be (.8-.2)×(2×10^14) = 1.2×(10^14).

[189] Assuming ten billion lives per century for two million years.

Suppose that, if we let the human race continue and did research for 300 years, we would know for certain whether or not additional people are of positive or negative value. If so, then with the credences above we should think it 80% likely that we will find out that it is a bad thing to produce new people, and 20% likely that we will find out that it's a good thing to produce new people. So there's an 80% chance of a loss of 3×(10^10) (because of the delay of letting the human race go extinct), the expected value of which is 2.4×(10^10). But there's also a 20% chance of a gain of 2×(10^14), the expected value of which is 4×(10^13). That is, in expected value terms, the cost of waiting for a few hundred years is vanishingly small compared with the benefit of keeping one's options open while one gains new information.

In general, when one has the choice between two options, one of which is irreversible, and one expects to make moral progress, then option value gives one additional reason in favour of choosing the reversible option. This is true even if one thinks it likely that the irreversible option will have the better consequences. When the decision is very high stakes, this option value can be very high indeed. In the above, I gave the extinction of the human race as a case study of this, and it is probably the most important application of the idea. But there are other applications too: option value gives a reason not to consume non-renewable resources such as fossil fuels; to preserve irreplaceable works of art and historic monuments; and in general to grow the economy's resources rather than to spend them on consumables.

Before I conclude, let me consider two important objections to my argument.

V. Objections

Moral philosophy gives us very little new moral information, and certainly never gives us certainty when we didn't have certainty before.

In the above examples I assumed that we'd be able to achieve certainty in the moral facts of the matter. But that's unrealistic: we should never end up with certainty about some controversial moral view. So, in our decision-analytic language, we should be thinking about imperfect information — which improves our epistemic state — rather than perfect information, which gives us certainty.

Imperfect information doesn't change the underlying argument. But it makes the maths more difficult. Take, for instance, the decision whether to study ethics for 6 months. To get a crude approximation of the value of imperfect information, one could ask oneself: after that time period, how likely am I to have changed my moral view? And, given that I change my view, what is the difference in value between the decision I'd make then and the decision I'd make now? This procedure would approximate the value of information, but it wouldn't be quite satisfactory. Really, you'd want to provide a probability distribution over all the possible ways in which you could change your view, and the gain in value for all of those possibilities. The value of imperfect information would be the integral of the gains in value with respect to that probability distribution.

How could you even guess at likelihoods of changing one's view? A simple way would be to use induction from past experience: if one has already spent a fair bit of time doing ethical research, one could look at how many months one had spent doing the

research, how many times one had changed one's view on the topic, and how big a difference to the expected value of one's decisions those changes made. This would give one some amount of data by which to make a guess about how likely it is for one to change one's view given additional research. And if one hasn't done research in the past, then one could use information about the likelihood of change from those who have.

The value of perfect information functions as an upper bound on the value of imperfect information. So the value of gaining moral information in the real world will be lower than the values I suggested in my examples. But it might still get very high, especially as one approximates certainty. In my own case, for example, on the basis of philosophical arguments I moved from very high credence that eating factory farmed non-human animals for pleasure is permissible to very high credence that it is seriously morally wrong. In this case for me the value of the imperfect information I have was not very different from the value of perfect information.

In other cases, even small changes to our credences can be of highly significant value. Consider again the evaluation of human extinction. The crucial point does not involve gaining certainty. Rather, the crucial point is that the number of people who will live in the next few hundred years, as a proportion of the number of lives whose existence we can affect, is very small (e.g. 0.02% change) compared to the chance of us changing our moral view in that time (e.g. upward of a 1% chance). Because of this, the gain in expected value from increasing our credence in the true moral theory and changing our moral view dwarfs the loss in value from losing the ability to decide whether or not the

people in the next few hundred years exist or not, even if we do not expect to achieve anything close to certainty.

The framework is premised on moral realism. But moral realism is false.

In the above I have argued for the high value of moral philosophy on the basis that it helps us to make better decisions. But one might question the argument on the basis that it presumes that there's a notion of 'improving' with respect to the moral truth. And that sounds pretty robustly realist.[190]

However, the meta-ethical view that is required is realist only in a minimal sense: as long as one can make sense of a notion of moral proposition's being true or false, and of one having better or worse evidence with respect to those propositions, then one can makes sense of it being important to gain new moral information. And very many meta-ethical views can make sense of that. Sophisticated subjectivist moral views certainly can: it's certainly non-obvious, for example, what one would desire oneself to desire if one were fully rational;[191] and one can certainly improve one's evidence on the question of what such desires would look like. And the sorts of non-cognitivist views that are defended in the contemporary literature[192] want to capture the idea that one's moral views can be correct or incorrect, and that one can have greater or lesser credence in different moral views.

[190] Using 'robust realist' to refer to a position that denies non-cognitivism, error theory, and subjectivism (Rosen 1994). In what follows, when I say 'anti-realist' I contrast that with robust realism.
[191] Using Smith's (1994) view as an example.
[192] Such as Blackburn (1999) and Gibbard (2003).

It's true that the likelihood that one places on changing one's view might vary depending on the meta-ethical view one endorses. If one is robustly realist, then the idea that common sense has got things radically wrong generally becomes more plausible than if one is some flavour of anti-realist. But it seems to me that anti-realist views actually support my argument rather than detract from it. If one is a subjectivist, one should be optimistic about the likelihood of finding the moral truth — as finding the moral truth is ultimately just about working out what one values. The subjectivist should therefore think it more likely that she will change her view in light of further study and reflection than the robust realist, and that makes the value of information higher.[193]

Moreover, even if one endorsed a meta-ethical view that is inconsistent with the idea that there's value in gaining more moral information, one should not be certain in that meta-ethical view. And it's high-stakes whether that view is true — if there are moral facts out there but one thinks there aren't, that's a big deal! Even for this sort of anti-realist, then, there's therefore value in moral information, because there's value in finding out for certain whether that meta-ethical view is correct.

Conclusion

In this chapter I have argued that the value of gaining new moral information has the potential to be high, and is likely to be much higher than practitioners attribute to it.

[193] Though one might place less worth on doing moral philosophy oneself in order to influence others, as the view that best fits with one's own fundamental desires might be very different from the view that best fits with other people's.

Philanthropic foundations should at least sometimes be willing to spend a large proportion of their finances synthesising and learning from ethical research; young people should at least sometimes be willing to spend a substantial period of time studying ethics before making their career choice; we, as a society, should at least sometimes be willing to spend a decent proportion of our resources on ethical research and dissemination; and we have strong reasons to prevent the near-term extinction of the human race.

It's important, though, to be careful to draw the correct conclusion from this line of argument. It provides an argument that the value of moral philosophy can be high, but it doesn't motivate the claim that researching moral philosophy is the best thing that one could be doing with one's time. It also doesn't provide an argument for doing first-order research into moral philosophy oneself rather than, say, taking a lucrative career and sponsoring other people to do research in your stead. And, finally, it really shows that being a moral philosopher has the potential to be a *high-stakes* career. Just as one might be dong something of immense value by increasing the chance of humanity converging on the correct moral view, if one is actually retarding that progress — by propounding seductive but poor arguments, or by generating more confusion than insight — then one might be doing something incredibly bad indeed.

Conclusion

In this thesis I first argued for a general metanormative theory. Ultimately, I argued that it's appropriate to maximise expected choice-worthiness, where cardinal but non-comparable theories are normalised with each other at their variance, and where ordinal theories are normalised with cardinal theories at the variance of their Borda Scores. I suggested that, when intertheoretic comparisons of choice-worthiness are possible, they are possible in virtue of options' instantiating intrinsic choice-worthiness properties that exist across epistemically possible worlds.

I then explored further issues relating to decision-making under normative uncertainty. I introduced the infectious incomparability problem for MEC, and suggested a solution. I discussed the implications of decision-theoretic metanormativism for the causal/evidential decision theory debate, and drew out one implication of MEC for practical ethics.

Metanormativism has been remarkably underexplored, and I hope that this thesis has begun the project of rectifying that. I hope I have shown, in the first part of this thesis, that what have been considered to be devastating problems for MEC — the problems of ordinal theories and of intertheoretic comparisons — are in fact solvable. And I hope I have shown, in the second part of this thesis, that metanormativism provides fertile ground for producing insights into other philosophical debates.

Bibliography

Adams, Fred C. 2008. "Long-Term Astrophysical Processes." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic. Oxford University Press.

Ahmed, Arif. 2013. "Causal Decision Theory: A Counterexample." *Philosophical Review*.

Ahmed, Arif, and Huw Price. 2012. "Arntzenius on 'Why Ain'cha Rich?'." *Erkenntnis* 77 (1): 15–30.

Allais, M. 1953. "Le Comportement de l'Homme Rationnel Devant Le Risque: Critique Des Postulats et Axiomes de l'Ecole Americaine." *Econometrica* 21: 503–46.

Arntzenius, Frank. 2008. "No Regrets, or: Edith Piaf Revamps Decision Theory." *Erkenntnis* 68 (2): 277–97.

Barry, Christian, and Patrick Tomlin. ms. "Uncertainty, Permissibility, and Compromise."

Beckstead, Nick. ms. "Recklessness, Timidity, and Fanaticism."

Blackburn, Simon. 1999. *Ruling Passions: A Theory of Practical Reasoning*. Oxford University Press, USA.

Blackorby, Charles, David Donaldson, and John A. Weymark. 1984. "Social Choice with Interpersonal Utility Comparisons: A Diagrammatic Introduction." *International Economic Review* 25 (2): 327–56.

Bordes, Georges, and Nicolaus Tideman. 1991. "Independence of Irrelevant Alternatives in the Theory of Voting." *Theory and Decision* 30 (2): 163–86.

Bostrom, Nick. 2003. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (03): 308–14.

———. 2009. "Pascal's Mugging." *Analysis* 69 (3): 443–45.

Briggs, Rachael. 2010. "Decision-Theoretic Paradoxes as Voting Paradoxes." *Philosophical Review* 119 (1): 1–30.

Brink, David. 1993. "The Separateness of Persons, Distributive Norms, and Moral Theory." In *Value, Welfare, and Morality*, edited by R.G. Frey and C.W. Morris. Cambridge University Press.

Broome, John. 1995. *Weighing Goods*. Wiley-Blackwell.

———. 2004. *Weighing Lives*. Oxford University Press.

———. 2010. "The Most Important Thing About Climate Change." In *Public Policy: Why Ethics Matters*, edited by Jonathan Boston, Andrew Bradstock, and David Eng, 101–16. ANU E Press.

———. 2012. *Climate Matters: Ethics in a Warming World (Norton Global Ethics Series)*. W. W. Norton & Company.

Buchak, Lara. 2013. *Risk Aversion and Rationality*. Oxford University Press.

Bykvist, Krister, and Jonas Olson. 2009. "Expressivism and Moral Certitude." *The Philosophical Quarterly* 59 (235): 202–15.

———. 2012. "Against the Being For Account of Normative Certitude." *Journal of Ethics and Social Philosophy*.

Cotton-Barratt, Owen. ms. "Geometric Reasons for Normalising Variance to Aggregate Preferences." http://users.ox.ac.uk/~ball1714/Variance%20normalisation.pdf.

Crouch, William. 2010. "How to Act Appropriately in the Face of Moral Uncertainty". BPhil Thesis, Oxford University.

Easwaran, Kenny. 2014. "Regularity and Hyperreal Credences." *Philosophical Review* 123 (1): 1–41.

Eeckhoudt, Louis, and Philippe Godfroid. 2000. "Risk Aversion and the Value of Information." *The Journal of Economic Education* 31 (4): 382–88.

Egan, A. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116 (1): 93–114.

Fine, Kit. 2012. "Guide to Ground." In *Metaphysical Grounding: Understanding the Structure of Reality*, 37–80.

Gibbard, Allan. 1992. "Weakly Self-Ratifying Strategies: Comments on McClennen." *Philosophical Studies* 65 (1-2): 217–25.

———. 2003. *Thinking How to Live*. Harvard University Press.

Gracely, Edward J. 1996. "On the Noncomparability of Judgments Made by Different Ethical Theories." *Metaphilosophy* 27: 327–32.

Guerrero, Alexander A. 2007. "Don't Know, Don't Kill: Moral Ignorance, Culpability, and Caution." *Philosophical Studies* 136: 59–97.

Gustafsson, Johan E. 2011. "A Note in Defence of Ratificationism." *Erkenntnis* 75 (1): 147–50.

Gustafsson, Johan, and Tom Torpman. forthcoming. "In Defence of My Favourite Theory." *Pacific Philosophical Quarterly*.

Hájek, Alan. 2003a. "Waging War on Pascal's Wager." *The Philosophical Review* 112 (1): 27–56.

———. 2003b. "What Conditional Probability Could Not Be." *Synthese* 137 (3): 273–323.

Handfield, Toby. 2012. *A Philosophical Guide to Chance: Physical Probability*. Cambridge University Press.

Hare, Caspar. 2010. "Take the Sugar." *Analysis* 70 (2): 237 –247.

Harman, Elizabeth. 2014. "The Irrelevance of Moral Uncertainty." In *Oxford Studies in Metaethics*.

Hudson, James L. 1989. "Subjectivization in Ethics." *American Philosophical Quarterly* 26: 221–29.

Hurka, Thomas. 2010. "Asymmetries In Value." *Noûs* 44 (2): 199–223.

Jackson, Frank. 1991. "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101 (3): 461–82.

Jaynes, E. T. 1957. "Information Theory and Statistical Mechanics." *Physical Review* 106 (4): 620–30.

Joyce, James M. 2012. "Regret and Instability in Causal Decision Theory." *Synthese* 187 (1): 123–45.

Kamm, F. M. 1992. "Non-Consequentialism, the Person as an End-in-Itself, and the Significance of Status." *Philosophy & Public Affairs* 21 (4): 354–89.

Krantz, D., R. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of Measurement, Vol. 1*. New York: Academic Press.

Leslie, John. 1998. *The End of the World: The Science and Ethics of Human Extinction*. Taylor & Francis.

Lewis, David. 1981. "`Why Ain'cha Rich?'." *Noûs* 15 (3): 377–80.

Lockhart, Ted. 2000. *Moral Uncertainty and Its Consequences*. Oxford University Press.

MacAskill, William. 2013. "The Infectiousness of Nihilism." *Ethics*.

Matheny, Jason G. 2007. "Reducing the Risk of Human Extinction." *Risk Analysis* 27 (5): 1335–44.

Mongin, Philippe. 2000. "Does Optimization Imply Rationality?" *Synthese* 124 (1-2): 73–111.

Moore, G. E. 1903. *Principia Ethica*. Cambridge University Press.

Moulin, Hervé. 1988. "Condorcet's Principle Implies the No Show Paradox." *Journal of Economic Theory* 45 (1): 53–64.

Nagel, Thomas. 2008. "The Value of Inviolability." In *Morality and Self-Interest*, edited by Paul Bloomfield.

Nick Bostrom, and Milan M. Cirkovic. 2008. "Introduction." In *Global Catastrophic Risks*. Oxford University Press.

Nover, Harris, and Alan Hájek. 2004. "Vexing Expectations." *Mind* 113 (450): 237–49.

Nozick, Robert. 1994. *The Nature of Rationality*. Princeton University Press.

Oddie, Graham. 1995. "Moral Uncertainty and Human Embryo Experimentation." In *Medicine and Moral Reasoning*, edited by K. W. M. Fulford, Grant Gillett, and Janet Martin Soskice. OUP.

Okasha, Samir. 2011. "Theory Choice and Social Choice: Kuhn versus Arrow." *Mind* 120 (477): 83–115.

Pearl, Judea. 2010. "The Curse of Free-Will and the Paradox of Inevitable Regret."

Posner, Richard A. 2004. *Catastrophe: Risk and Response*. Oxford University Press.

Price, Huw. 2012. "Causation, Chance, and the Rational Significance of Supernatural Evidence." *Philosophical Review* 121 (4): 483–538.

Rabinowicz, Wlodek. 2000. "Money Pump with Foresight." In *Imperceptible Harms and Benefits*, edited by Michael J. Almeida, 123–54. Library of Ethics and Applied Philosophy 8. Springer Netherlands.

Raïffa, Howard. 1968. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. McGraw-Hill Education.

Rees, Martin. 2003. *Our Final Hour*. Basic Books.

Regan, Donald. 1980. *Utilitarianism and Co-Operation*. Clarendon Press.

Riedener, Stefan. 2013. "Maximising Expected Value under Axiological Uncertainty". BPhil Thesis, Oxford University.

Rosen, Gideon. 1994. "Objectivity and Modern Idealism: What Is The Question?" In *Philosophy in Mind*, edited by M. Michael and J. O'Leary-Hawthorne, 60:277–319. Philosophical Studies Series. Springer Netherlands.

Ross, Jacob. 2006. "Rejecting Ethical Deflationism." *Ethics* 116: 742–68.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Harvard University Press.

Schick, Frederic. 1986. "Dutch Bookies and Money Pumps." *The Journal of Philosophy* 83 (2): 112–19.

Schulze, Markus. 2010. "A New Monotonic, Clone-Independent, Reversal Symmetric, and Condorcet-Consistent Single-Winner Election Method." *Social Choice and Welfare* 36 (2): 267–303.

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

Sen, Amartya Kumar. 1970. *Collective Choice and Social Welfare*. Holden-Day.

———. 1997. *Choice, Welfare and Measurement*. Harvard University Press.

Sepielli, Andrew. 2009. "What to Do When You Don't Know What To Do." In *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau, 350. Oxford University Press.

———. 2010. "Along an Imperfectly Lighted Path': Practical Rationality and Normative Uncertainty". PhD Thesis, Rutgers University.

———. 2012. "Moral Uncertainty and the Principle of Equity among Moral Theories." *Philosophy and Phenomenological Research*.

———. 2013. "What to Do When You Don't Know What to Do When You Don't Know What to Do…." *Noûs*.

———. 2012. "Normative Uncertainty for Non-Cognitivists." *Philosophical Studies*: 1–17.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1: 229–43.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. Harvard University Press.

Smith, Michael. 1994. *The Moral Problem*. Wiley.

———. 2002. "Evaluation, Uncertainty and Motivation." *Ethical Theory and Moral Practice* 5 (3): 305–20.

Sobel, Jordan Howard. 1988. "Infallible Predictors." *The Philosophical Review* 97 (1): 3–24.

Solomonoff, R.J. 1964. "A Formal Theory of Inductive Inference. Part I." *Information and Control* 7 (1): 1–22.

Suppes, Patrick, and Joseph Zinnes. 1963. "Basic Measurement Theory." In *Handbook of Mathematical Psychology*, 1–1. John Wiley & Sons.

Tideman, Nicolaus. 1987. "Independence of Clones as a Criterion for Voting Rules." *Social Choice and Welfare* 4 (3): 185–206.

———. 2006. *Collective Decisions and Voting*. Ashgate.

Von Neumann, John, Oskar Morgenstern, Ariel Rubinstein, and Harold William Kuhn. 2007. *Theory of Games and Economic Behavior*. Greenwood Publishing Group.

Weatherson, Brian. 2002. "Review of Ted Lockhart's 'Moral Uncertainty and Its Consequences.'" *Mind* 111: 693–96.

———. 2014. "Running Risks Morally." *Philosophical Studies* 167 (1): 141–63.

Wedgwood, Ralph. ms. "Objective and Subjective 'Ought.'"

———. 2013. "Akrasia and Uncertainty." *Organon F.* 20 (4).

———. 2013. "Gandalf's Solution to the Newcomb Problem." *Synthese*: 1–33.

Weirich, Paul. 1985. "Decision Instability." *Australasian Journal of Philosophy* 63 (4): 465–72.

Young, Peyton. 1974. "An Axiomatization of Borda's Rule." *Journal of Economic Theory* 9: 43–52.

Zimmerman, Michael J. 2008. *Living with Uncertainty: The Moral Significance of Ignorance*. Cambridge University Press.