ABSTRACT

Title of dissertation: RATIONALITY AND MORAL RISK: A MODERATE DEFENSE OF HEDGING

Christian Tarsney, Doctor of Philosophy, 2017

Dissertation directed by: Professor Dan Moller
Department of Philosophy

How should an agent decide what to do when she is uncertain not just about morally relevant empirical matters, like the consequences of some course of action, but about the basic principles of morality itself? This question has only recently been taken up in a systematic way by philosophers. Advocates of *moral hedging* claim that an agent should weigh the reasons put forward by each moral theory in which she has positive credence, considering both the likelihood that that theory is true and the strength of the reasons it posits. The view that it is sometimes rational to hedge for one's moral uncertainties, however, has recently come under attack both from those who believe that an agent should always be guided by the dictates of the single moral theory she deems most probable and from those who believe that an agent's moral beliefs are simply irrelevant to what she ought to do. Among the many objections to hedging that have been pressed in the recent literature is the worry that there is no non-arbitrary way of making the intertheoretic comparisons of moral value necessary to aggregate the value assignments of rival moral theories into a single ranking of an agent's options.

This dissertation has two principal objectives: First, I argue that, contra these recent objections, an agent's moral beliefs and uncertainties are relevant to what she rationally ought to do, and more particularly, that agents are at least sometimes rationally required to hedge for their moral uncertainties. My principal argument for these claims appeals to the *enkratic conception of rationality*, according to which the requirements of practical rationality derive from an agent's beliefs about the objective, desire-independent value or choiceworthiness of her options. Second, I outline a new general theory of rational choice under moral uncertainty. Central to this theory is the idea of *content-based aggregation*, that the principles according to which an agent should compare and aggregate rival moral theories are grounded in the content of those theories themselves, including not only their value assignments but also the metaethical and other non-surface-level propositions that underlie, justify, or explain those value assignments.

RATIONALITY AND MORAL RISK: A MODERATE DEFENSE OF HEDGING

by

Christian Tarsney

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Dan Moller, Chair/Advisor
Professor Samuel Kerstein
Professor Christopher Morris
Professor Eric Pacuit
Professor Karol Soltan

Acknowledgments

For helpful comments and/or discussion of various parts of this dissertation, I thank Ron Aboodi, Heather Adair, Krister Bykvist, Joseph Carlsmith, Owen Cotton-Barratt, Mathias Frisch, Andrew Fyfe, Andrew Garland, Hilary Greaves, Quinn Harr, Amelia Hicks, Samuel Kerstein, Daniel Kokotajlo, William MacAskill, Dan Moller, Christopher Morris, Toby Ord, Eric Pacuit, Caleb Pickard, Douglas Portmore, Chelsea Rosenthal, Anders Sandberg, Julius Schoenherr, Andrew Sepielli, Paul Shephard, Joshua Shepherd, Richard Teague, Sergio Tenenbaum, Taylor White, and many anonymous reviewers. Thanks also to audiences at the University of Maryland, the Uehiro Centre for Practical Ethics, the Oxford Applied Ethics Graduate Discussion Group, the 2014 Rocky Mountain Philosophy Conference at Colorado, the 2014 Mark L. Shapiro Graduate Philosophy Conference at Brown, the 2014 Northern Illinois Graduate Philosophy Conference, the 2015 Long Island Philosophical Society Conference, the 2015 Rocky Mountain Ethics Congress, the 2015 "Nature and Norms" conference at Johns Hopkins, the 2016 Central APA, the 2016 Virginia Philosophical Association Conference, the 2017 Pacific APA, the 2017 Columbia-NYU Graduate Philosophy Conference, and the 2017 British Society for Ethical Theory Conference. Special thanks to my adviser, Dan Moller, for tireless reading of drafts and many hours of fruitful discussion.

Personal thanks are also due to my mother, Juli, for unstinting editorial support; to my father, Jim, for passing along a regular stream of philosophical witticisms and irrefutable arguments for the existence of an absolute *up*; and to my

sister Catherine and girlfriend Erika for assiduously reminding me that, to most well-adjusted people, philosophy is not *that* interesting.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DNET	Democritean normative error theory
ECS	exclusionary conception of supererogation
EE	enkratic externalism
EP	enkratic principle
EPP	egoistic permissions principle
FDOT	Final Dominance over Theories
GDoT	Genuine Dominance over Theories
GDoT*	Strengthened Genuine Dominance over Theories
MEC	maximize expected choiceworthiness
MFO	My Favorite Option
MFT	My Favorite Theory
PEMT	Principle of Equity among Moral Theories
PES	principle of equal say
PIVA	problem of intertheoretic value aggregation
SCS	satisficing conception of supererogation
SD	stochastic dominance
SSR	Strong Subjective Rationalism
UCBN	universal content-based normalization
WSR	Weak Subjective Rationalism

Chapter 1: The Problem of Moral Uncertainty

Being a conscientious person means trying to do the right thing. But trying to do the right thing, even trying one's best, is no guarantee of success, for we often face substantial obstacles to acting well or rightly. One kind of obstacle is weakness of will, and challenges to the will like fear, temptation, or distraction. Another kind of obstacle is ignorance or uncertainty about the good or the right, about what we ought to do. Sometimes the source of uncertainty is empirical, e.g., uncertainty concerning the consequences of some available course of action. But sometimes we are also relevantly uncertain about the basic requirements of morality, e.g., about whether or to what extent some feature of the world like happiness, beauty, or knowledge has non-derivative value, or about whether the right thing to do is always the thing that maximizes the sum of value in the world.

This latter form of uncertainty, and how we can try our best to do the right thing in spite of it, is the topic of this dissertation. Following convention, I will use the term *moral uncertainty* to refer exclusively to this sort of "purely moral" uncertainty, uncertainty about basic moral principles as opposed to uncertainty about morally relevant empirical facts. As I hope will become clear, moral uncertainty even in this restricted sense is a pervasive feature of everyday life: Many basic questions about the requirements of morality remain unresolved despite millennia of philosophical effort. Moreover, many of the accounts of morality that philosophers have found most plausible make demands of which we routinely fall short, and that in many cases conflict with one another. Lying often seems likely to yield better overall consequences than truth-telling, putting Kantian and utilitarian theories in conflict. Commonsense morality often demands that we prioritize our friends or family members in a way that impartial, universalistic theories like both Kantianism and utilitarianism forbid. Egalitarian theories of justice demand that we support the socialist candidate for public office while historical entitlement theories demand that we support the libertarian. Therefore our uncertainty about morality is not just theoretical but practical: Any agent, no matter how conscientious, must often be unsure whether she is doing the right thing.

1.1 Moral Uncertainty

It may be tempting, in the face of such worries, to simply deny that we are relevantly uncertain about the basic requirements of morality. Especially when we focus on the most extreme demands of a given moral theory, or on cases where the theory conflicts with our own strongly held moral or political beliefs, we may be able to talk ourselves at least momentarily into the view that such a theory is *certainly* false. The importance of knowing how to make decisions in the face of moral uncertainty will depend very much, then, on how often we should really be uncertain, and how uncertain we should be, about the demands of morality with respect to the morally significant choices we face.

It is a trivial consequence of certain views about rational belief that we should *always* be at least somewhat uncertain about basic moral questions. In particular, if one accepts *probabilism*, according to which rational belief should come in degrees that obey the axioms of the probability calculus, and if one further accepts the requirement of *regularity*, according to which the probability assigned to any proposition that is not a logical truth or falsehood must be strictly greater than 0 and less than 1, then it seems we are rationally required to maintain at least some degree of uncertainty about any non-trivial moral proposition.

As I will explain in §1.7, I am sympathetic to these claims. But they alone are not enough to make the problem of moral uncertainty interesting, for our moral uncertainties about the choices we face as moral agents might nevertheless be so slight as to be practically irrelevant on any plausible account of rational choice under moral uncertainty. For instance, accepting regularity commits one to non-zero credence is the proposition that it is always morally obligatory to torture babies for fun. Nevertheless, one's degree of belief in this proposition may still be so *close* to zero that the reasons for action it generates, if any, are vanishingly weak, to the point that it is not worth attempting to explicitly account for them in practical reasoning.

But whether or not one accepts the general commitments of probabilism and regularity, there are more mundane, and more compelling, reasons for thinking that we ought to be *substantially* uncertain about a great many fundamental moral questions, and in a way that is practically relevant to a great many of our choices. First is the fact of widespread, intractable, and apparently reasonable disagreement among intelligent, reflective, and well-motivated moral agents. Among philosophers, centuries of debate between consequentialists, Kantians, virtue ethicists, et al have so far failed to yield consensus, and indeed the variety of moral theories that can claim active adherents in the philosophical literature continues to expand rather than contract. Likewise, traditional religious moral codes differ considerably in the requirements and prohibitions they enjoin (e.g. with respect to sexual morality, consumption of intoxicants, the relative weight of obligations to the poor, family members, ancestors, and religious institutions, etc), and these disagreements have proven no more tractable. Further, as any philosopher who has taught an undergraduate course on contemporary moral controversies can attest, even those without any antecedent commitment to either a philosophical theory of morality or a religious moral upbringing can differ greatly in their intuitive moral judgments concerning our obligations to non-human animals, the distant poor, or future generations, or concerning the permissibility of assisted suicide, killing civilians in war, abortion and infant euthanasia, and too many other topics to name. Thus, whatever population one takes to be the appropriate reference class of competent moral judges, it must be conceded that one's own moral convictions are far from universally shared within that class, and indeed that one's own comprehensive set of moral beliefs is almost certainly shared by a tiny fraction of moral judges at best.

In addition to these ongoing moral disagreements, the degree to which moral opinion has changed over time should challenge our confidence even in those moral judgments that are widely shared among present-day moral agents. Practices that have been generally accepted or encouraged in many times and places, like slavery, infant exposure, pederasty, and arranged marriage are now widely or universally viewed as morally repugnant, while requirements once seen as stringent, say with respect to personal or familial honor or sexual chastity, have lost their stringency or dropped out of our shared moral code altogether. There are good reasons to regard many of these belief changes as moral progress—we need not be too troubled by our disagreement with our ancestors regarding, say, the permissibility of declining a duel. But reflection on this history should lead us to conclude, by pessimistic induction, that many of our own moral beliefs will likely be seen as absurd and perverse by our own descendants, and likely in some respects that even the most morally reflective among us would be unable to guess.

Likewise, we ourselves may have experienced significant changes of moral belief in our lifetimes. Indeed, it seems likely to me that anyone who has thought long about questions of moral theory has experienced at least one such change in their initial convictions. Here again, though we may have good reason to regard our past opinions as wrongheaded, memories of misplaced confidence in those past opinions should leave us open to the possibility of further significant revisions in future.[1]

Further grounds for modesty in our moral beliefs comes from the finding that we are disposed toward overconfidence across a wide variety of doxastic domains.[2] In general, where such assessment is possible, we find that people's explicit degrees of belief display significant calibration errors in the direction of excessive confidence.

[1] For remarks on these and related considerations, see Moller (2011, pp. 432-4).

[2] For a useful overview of recent research on miscalibration and overconfidence, see Part III of Kahneman (2011).

For example, when faced with questions that have a well-ordered range of possible answers (e.g. dates or quantities) and asked to produce, say, 90% or 95% confidence intervals on this range (such that one is 90% or 95% confident that the correct answer falls in between the lower bound and upper bound), most of us will produce intervals that fail to cover the correct answer at far more than the expected rate of 10% or 5% (Soll and Klayman, 2004). Likewise, when asked to assign degrees of confidence to one's answers to common-knowledge questions, our stated confidence tends to substantially exceed our measured reliability (Alpert and Raiffa, 1982). Of course, in the moral context we have no means of checking the reliability of a given individual's moral judgments, but insofar as the finding of overconfidence can be inductively generalized to domains in which it cannot be directly confirmed, we have reason to suspect that our confidence in our moral judgments may be prone to this same bias.

A final reason for general uncertainty about questions of morality is the lack of well-established methods or foundations for moral theorizing. Proposed starting points for moral theory include conceptual truths about action and rational agency (e.g. Korsgaard (1996)), self-evident axioms about the good and the right (e.g. Sidgwick (1874), de Lazari-Radek and Singer (2014)), the considered case judgments of reflective moral agents (e.g. Rawls (1951), Daniels (1979)), the predicted choices of agents in idealized bargaining conditions (e.g. Gauthier (1986)), intentional design of moral principles to meet the needs of human societies, anthropological observation of the moral practices of existing societies, scientific study of the evolutionary function of moral norms, and many more. The most popular method of investigation in contemporary analytic moral philosophy, the method of reflective equilibrium based on heavy appeal to intuitive judgments about cases, has come under concerted attack and is regarded by many philosophers (e.g. Singer (2005), Greene (2008)) as deeply suspect. Additionally, every major theoretical approach to moral philosophy (whether at the level of normative ethics or metaethics) is subject to important and intuitively compelling objections, and the resolution of these objections often turns on delicate and methodologically fraught questions in other areas of philosophy like the metaphysics of consciousness or personal identity (Moller, 2011, pp. 428-432). Whatever position one takes on these debates, it can hardly be denied that our understanding of morality remains on a much less sound footing than, say, our knowledge of the natural sciences. If, then, we remain deeply and justifiably uncertain about a litany of important questions in physics, astronomy, and biology, we should certainly be at least equally uncertain about moral matters, even when some particular moral judgment is widely shared and stable upon reflection.

The only plausible way one might deny that we should be substantially uncertain about many important moral matters, it seems to me, is to deny that morality is a matter of belief in the first place—that is, to adopt a non-cognitivist account of moral judgment. Still, even non-cognitivists must account for the *appearance* of moral uncertainty, for the fact that something very much like uncertainty is unquestionably a part of our moral experience. Some philosophers (e.g. Smith (2002)) have argued that non-cognitivism must be rejected on the grounds that it *cannot* adequately account for the uncertainty-like features of moral judgment. Others (e.g. Sepielli (2012)) have proposed sophisticated versions of non-cognitivism meant to address these objections and explain how non-doxastic attitudes can display features closely analogous to uncertainty. I have nothing to add to these debates and will simply assume, as others in the literature have done, that any adequate version of non-cognitivism must allow for features of moral judgment structurally similar enough to what the cognitivist calls "moral uncertainty" that discussions of the phenomenon conducted in the language of cognitivism can be readily translated into the non-cognitivist framework.

1.2 Extant Approaches and Open Problems

In light of the pervasiveness of moral uncertainty (or its non-cognitivist analogue), and of compelling reasons for that uncertainty, it is more than a little remarkable that the question of how agents should rationally resolve moral dilemmas in the face of such uncertainty received almost no attention from analytic philosophers until quite recently.[3] The recent spate of philosophical interest in the problem dates to the publication of Ted Lockhart's *Moral Uncertainty and Its Consequences* (Lockhart, 2000), which developed an ambitious theory of expected value maximization in the face of moral uncertainty.[4] Unfortunately Lockhart's proposal, which turns on a

[3] Speculatively, it may be that the increased focus in the second half of the 20th century on moral dilemmas like the trolley problem, as well as questions of practical ethics like philanthropic obligations to the distant poor and the moral status of animals, led moral philosophers to pay more attention to the practical conflicts between prominent moral theories like Kantianism and utilitarianism and hence to the need for practical guidance in the face of uncertainty concerning these cases of practical conflict.

[4] The philosophical literature on choice under moral uncertainty before Lockhart comprises a very few papers, chiefly Greenwell (1977), Lockhart (1977), Pfeiffer (1985), Hudson (1989), Oddie (1994), and Gracely (1996). There is also, interestingly, a much earlier and fairly extensive literature from Catholic moral theology dealing with a particular aspect of the problem of moral uncertainty, namely, how confident one must be that an act is objectively permissible rather than forbidden before one is subjectively permitted to perform it. This debate mainly pitted rigoristic Jansenists (including Pascal (1657)), who demanded near-certainty before a morally risky act could

clever but counterintuitive account of how to make quantitative value comparisons across rival moral theories, has been subjected to powerful criticisms (especially by Sepielli (2006, 2013)) and has found no subsequent defenders in the literature.

In the years since Lockhart, other philosophers have suggested partial or general solutions to the problem of intertheoretic value comparisons (e.g. Ross (2006), Sepielli (2009)), while still others have offered arguments that the problem is unsolvable (e.g. Gustafsson and Torpman (2014), Nissan-Rozen (2015)). We will examine these debates at length in Chapters 5-6.

Others, meanwhile, have explored the practical implications of moral uncertainty with respect to particular moral dilemmas. Most of this literature has focused on cases of "moral risk" (actions that seem to be either forbidden or merely-permissible), rather than cases of "moral conflict" (actions that may be either forbidden or obligatory), with the most widely discussed cases being abortion and vegetarianism. Prominent in this literature are Ross (2006), Guerrero (2007), Moller (2011), and MacAskill (2013).

Finally, in the last few years, there has been concerted pushback against the very idea that what an agent ought to do is sensitive to her uncertainties about basic moral questions and hence that there is need for a theory of rational choice under moral uncertainty. The extreme position in this literature (exemplified by Weatherson (2014), Harman (2015), and Hedden (2016)) claims that what an agent

be contemplated, against the more laissez-faire Jesuits. (For a useful overview , see Sepielli (2010, pp. 48-53).) Kant was a latecomer to this debate, opining in his *Religion within the Limits of Reason Alone* that "it is a basic moral principle, which requires no proof, that one ought to hazard nothing that may be wrong" (Kant, 1793, p. 173; cited in Kerstein, 2002, p. 214). (Thanks to Dan Moller for bringing this passage to my attention.)

ought to do is simply insensitive to her moral beliefs, while a more moderate position (exemplified by Gustafsson and Torpman (2014) and Nissan-Rozen (2015)) merely denies that agents should engage in expected value-style weighing of the reasons posited by rival moral theories. We will consider several of the concerns raised in this literature, beginning in Chapter 3.

Despite this recent flurry of interest, most of the central problems associated with moral uncertainty still lack any satisfactory resolution, as we will see. In addition to the extreme lack of consensus regarding whether and in what ways the normative requirements incumbent on an agent depend on her moral beliefs (and, indeed, lack of clarity concerning the range of possible views on this question and which normative concepts it centrally concerns), no approach to the problem of intertheoretic value comparisons has achieved anything like widespread support even among those philosophers who agree that such an approach is needed. In addition, consideration of the problem of moral uncertainty draws attention to a deeper and more general problem concerning the nature of subjective normativity, the so-called "regress problem" which we will take up in Chapter 7, a problem that has been widely recognized but to which the extant literature contains only one or two hints at potential solutions.

Fortunately, recent developments in closely related philosophical literatures have given us new tools for approaching many of the problems related to moral uncertainty. Recent work on rationality, stemming from Broome (1999, 2001) and Kolodny (2005) *inter alia*, has furnished more sophisticated ways of thinking about both the nature and the content of rational requirements. Attention among metaethicists to the relationship and relative priority of objective ("fact-relative"), subjective ("belief-relative"), and prospective ("evidence-relative") normativity (Zimmerman (2008) and Parfit (2011), among many others) has sharpened these distinctions and brought clarity to long-standing debates between "objectivists" and "subjectivists" about moral obligation. A closely related debate on the relationship between moral ignorance and culpability, stemming largely from work by Michael Zimmerman (1997) and Gideon Rosen (2002, 2004) has helped to illuminate questions of the normative force of moral belief from a different direction. And work on value incommensurability/incomparability, driven forward in large part by the essays contained in Chang (1997a), has yielded new technical approaches to the apparent incomparability of disparate ethical considerations that is the central difficulty for theories of choice under moral uncertainty.

Despite its many difficulties, then, it is reasonable to hope that we will be able to make substantial progress on the problem of moral uncertainty in the coming years, at least to the point of offering more complete and more plausible statements of the various rival views than are presently available. In the pages that follow, I aim to make some contribution toward this goal.

1.3 Reasons, Oughts, and Rationality

It will be useful to spend some time setting out a repertoire of basic concepts to which I will regularly appeal in later chapters. I take the domain of practical reasoning to be characterizable in terms of four fundamental concepts:

agents, options, objective reasons/oughts, and subjective reasons/oughts. To the extent that other practical concepts like rationality/rational requirement, obligation/permission/prohibition, moral rightness/wrongness, praiseworthiness/blameworthiness, and so on are meaningful, it seems to me, they must be either interdefinable with or definable in terms of these four basic concepts. Short of examining the other candidates for fundamental practical concepts one at a time, I can defend this meager conceptual base only by an appeal to parsimony and by demonstrating in subsequent chapters how much can be said in the terms in offers.

Agents we can characterize simply as the sort of beings capable of choosing between options: the conceptual and empirical criteria for practical agency have little affect on the questions we will be considering. About options, a bit more needs to be said. Options are potential objects of choice for agents—that is, things that an agent *can choose*.[5] I will generally speak of "options" rather than "actions" as the objects of practical choice since an agent's options can include omissions, sequences of actions, and perhaps also general strategies/plans/dispositions to act that don't correspond to any specifically contemplated action. Nevertheless, actions and courses of action are paradigmatic examples of practical options.

In general I will assume that we are concerned with choice between *maximal* options, that is, options that are fully specified from an agent's perspective, having no more specific variants that the agent is capable of choosing between. For instance, *going to the store* is an option of which *walking to the store* and *driving to the store*

[5] One might, I suppose, take the notion of "choosing" to be fundamental and the notion of an "option" derivative, but this won't matter for our purposes.

are variants, so *going to the store* is non-maximal. If the agent anticipates, as she mulls over her options, that she can drive to the store via by way of either Oak St or Peach St, then "driving to the store" will not count as a maximal option for her either, since it has these more specific variants. It is very difficult to give any exact criterion for what an agent's options are at a given moment, and hence what her maximal options are, but we can safely leave this question unanswered.

Choosing an option is importantly different from *intending* to choose or to perform an option. By "choosing option O" I mean something like "attempting to perform option O": It is useful to talk of "choosing O" rather than "O-ing" since an agent may choose an option (e.g., paradigmatically, an action) and yet fail to perform it, and may perhaps perform an option (" ") accidentally or involuntarily. Forming an intention to perform an option O means making up one's mind that one will, at least a short time in the future, perform option O. Choosing O is, as I am using the expression, just the first step in the process of *actually performing* O. For example, an agent who manages to best Kavka's toxin puzzle (Kavka, 1983) *intends* to drink the toxin at midnight, but does not *choose* to drink the toxin unless and until she orders her hand to begin reaching for the vial. It is cases like this that highlight the need for practical principles to distinguish between intending an option and choosing/performing that option.[6]

[6] What I have said does not rule out rational requirements to form intentions. Rather, it simply *distinguishes* between a rational requirement to choose option O (which might follow from a belief that one ought to choose option O) and a rational requirement to *intend* to choose option O (which might follow from a belief that one *ought to intend* to choose option O). Very often these requirements will coincide—perhaps they will *always* coincide when the action in question is to be performed immediately. But they can also come apart, not just in anomalous cases like Kavka's, but also in the ordinary case where one believes that one is (now) rationally required to intend to perform some action in the future, but does does believe that one is (now) rationally required

Where merely schematic description is called for, I will denote agents by the terms A, B, C... or A_1, A_2, A_3... as the situation demands, and likewise to options by the terms O, P, Q... or O_1, O_2, O_3...

The two remaining elements in our repertoire of fundamental concepts, about which most needs to be said, are objective reasons/oughts and subjective reasons/oughts. I say that these are, together, just two fundamental concepts rather than four because I take reasons and oughts to be interdefinable. Roughly, an (objective/subjective) "ought" claim asserts that an agents has on-balance or all-things-considered (objective/subjective) reason to choose a particular option from some set of alternatives. Conversely, an (objective/subjective) reason to choose O is just a consideration that counts in favor of its being the case that the agent (objectively/subjectively) ought to choose O.[7] Whether one takes reasons or oughts as the more fundamental notion out of each pair is, I think, not particularly important. But for expository purposes, it is perhaps a bit easier to start with reasons.[8]

to choose/attempt to perform that action. For a more extended argument to similar effect see Reisner (2013), who suggests that all rational requirements must take the form of "matching attitude requirements" (requiring that an agent's attitudes match her beliefs about what attitudes she ought to have) and that therefore a rational requirement to intend can follow only from a belief about what one *ought to intend*, not from a belief about what one *ought to do*.

[7] I can only acknowledge in passing that these interdefinitions contradict the views of various philosophers (e.g. Foot (1972), Horty (2014)) according to whom true "ought" claims can fail to correspond to reasons for action. But this dispute may be merely verbal: it seems to me that I have in mind a different, more restrictive sense of "ought" than these philosophers.

[8] I don't want to say a great deal about the relationship between reasons and oughts, as I don't have a great deal to say and the details will generally be unimportant for our purposes. But I will take for granted that the relationship is *monotonic*: That is to say, both reasons and oughts express two-place relations between agents and options, and if given a certain set of reasons it is the case that agent A *ought to* choose option O, then the addition to A's set of reasons of a further reason for A to choose O cannot make it no longer the case that A ought to choose O, except insofar as the addition of that reason to choose O requires the withdrawal of other reasons to choose O, the addition of reasons not to choose O, or the addition of reasons for/withdrawal of reasons against choosing some alternative to O. (Reasons against choosing O can be thought of as reasons for choosing the non-maximal option $\neg O$, or as reasons for choosing each of the maximal alternatives to O.)

If the concepts of objective and subjective reasons are indeed basic, then it is of course no use attempting to give genuine *definitions* for them. The best I can do, in the style of Wittgenstein's characterization of "ethics" (Wittgenstein, 1965), is to attempt a few roughly synonymous expressions of each concept, and hope that the reader ends up with the intended concept in mind. Reasons are, then, "considerations that count in favor of actions." Objective reasons are considerations that count in favor of action *objectively*, in light of the way the world really is and independent of the agent's beliefs about the world (except in cases where these beliefs are relevant features of the world, e.g., when an agent has objective reason to take some course of action that would remedy her ignorance or false beliefs). *Very* roughly, objective reasons count in favor of action "from the standpoint of omniscience," and correspond in general (though not as a matter of definition or conceptual truth) to the considerations that would influence an agent's epistemically (and perhaps motivationally) idealized counterpart in wanting or advising the agent herself to adopt a certain course of action. In Parfit's well-known idiom, therefore, objective reasons are "fact-relative" rather than "belief-relative" (Parfit, 2011, p. 150).

As objective reasons count in favor of action "from the standpoint of omniscience," so subjective reasons count in favor of action from an agent's own standpoint, and in particular from her *doxastic* standpoint. In Parfit's idiom, they are "belief-relative" (and indeed, as it will be the major task of the coming chapters to argue, *moral*-belief-relative). In the classic case from Williams (1979), an agent who desires to drink from a glass that she believes to contain gin but that in fact contains gasoline has objective reason not to drink from the glass, in virtue of the fact that it contains gasoline, but has subjective reason to drink from the glass, in virtue of her belief that it contains gin. More particularly, an agent's subjective reasons depend either on her ("*de dicto*") beliefs about what objective reasons she has, or on her ("*de re*") beliefs about the properties that underwrite or subvene on her subjective reasons (of which distinction we will have much more to say in the coming chapters).

It is important to be clear that this objective/subjective distinction is *not* the distinction between agent-relative and non-agent-relative reasons. While objective reasons are (generally) independent of an agent's beliefs, they need not be independent of her desires or other aspects of her mental state. For instance, in the Williams case, the agent's objective reason not to drink from the glass does not depend on her belief but may well depend (as Williams would have us believe) on her desires, e.g., to avoid suffering and death. So talk of "objective reasons" does not by itself commit us to the existence of objective-in-the-sense-of-desire-independent value/disvalue, rightness/wrongness, etc. (This point will be further clarified in the next section when we discuss the Humean conception of rationality.) Moreover, even desire-independent objective reasons may still be agent-relative, e.g., each agent may have special reason to pursue her own objective welfare, where this consists of a list of objective goods whose contents are independent of her desires.

To reiterate the assumed relationship between reasons and oughts, an agent A objectively ought to choose option O iff she has on-balance, all-things-considered objective reason to choose O, and subjectively ought to choose O iff she has on-balance, all-things-considered subjective reason to choose O. I take subjective oughts to be synonymous with *rational requirements*: Agents are, always and only, rationally required to do what they have on-balance subjective reason to do. To be rational is to non-accidentally satisfy one's rational requirements: that is, to be sensitive to one's reasons in such a way that one generally does what one has on-balance subjective reason to do. I will frequently use "A subjectively ought to choose O" and "A is rationally required to choose O" interchangeably, as two ways of saying the same thing.[9]

1.4 Two Conceptions of Rationality

The concepts of reasons and oughts admit of two very different conceptions, each of which has been defended in various forms in the philosophical literature. According to what I will call the *Humean* conception of rationality: (i) An agent A has *objective reason* to choose an option O iff choosing O would in some way

[9] Kolodny (2005) has influentially argued that there are no general reasons to comply with rational requirements and that the apparent normativity of rational requirements stems from the fact that, whenever an agent is rationally required to choose option O, she is so required *because she believes herself* to have reason of some more mundane sort for choosing option O. I am generally in sympathy with Kolodny's account of rationality, and I take our disagreement with respect to the relationship between reasons and rational requirements to be merely verbal: Kolodny does not draw a distinction between objective and subjective reasons, reserving the term "reason" exclusively for what I have been calling "objective reasons." But everything I have said in the preceding section can be paraphrased in Kolodny's idiom if we simply take a "subjective reason" to choose option O as a positive degree of belief that one has some objective reason to choose O, on the assumption that such degrees of belief contribute monotonically toward generating a rational requirement to choose O. (Strictly, the account of rational requirement articulated in Kolodny (2005) implies that only *full* belief that one has objective reason to choose O can contribute toward generating rational requirements, but this feature of the view clearly requires amendment: even very partial belief that, say, my drink has been poisoned and that I therefore have objective reason not to drink it can generate a rational requirement against drinking.) On this way of putting things, then, one feature of Kolodny's view is that reasons are fundamental in the domain of objective normativity (that is, more fundamental than objective "oughts"/obligations/requirements), which rational requirements are fundamental in the domain of subjective normativity (that is, more fundamental than "subjective reasons"). But this is compatible with the framework I have adopted.

contribute to the satisfaction of her desires (or other motivational states); (ii) A *objectively ought* to choose O iff O satisfies her desires, on balance, better than any alternative option; and (iii) A has *subjective reason* to choose O iff she has positive degree of belief that choosing O would in some way contribute to the satisfaction of her desires (“ ”).

The notion of subjective oughts/rational requirements requires a bit more care. As I have said, I take it to be a conceptual truth that an agent subjectively ought to do what she has most subjective reason to do, so we may say that on the Humean conception of rationality, A subjectively ought to choose O iff her degrees of belief concerning the satisfaction or dissatisfaction of her various desires conditional on choosing each of the options available to her favor O above the available alternatives. But this leaves open the question how degrees of subjective-reason-giving weight are assigned to an agent's degrees of belief concerning the satisfaction or dissatisfaction of her desires conditional on choosing a particular option. The standard answer to this question is that an agent subjectively ought to maximize the *expected satisfaction* of her desires, i.e., that the weight of a subjective reason for/against choosing O is simply the product of the strength of the underlying desire times the agent's degree of belief that choosing O will result in the satisfaction/dissatisfaction of that desire. But, as we will see, other theories of rational requirement are possible within the general framework of the Humean conception of rationality.

In contrast, according to what I will call the *enkratic conception* of rationality: (i) An agent A has *objective reason* to choose an option O iff O is in some respect good, valuable, or choiceworthy; (ii) A *objectively ought* to choose O iff O is better, more valuable, or more choiceworthy than any alternative option; and (iii) A has *subjective reason* to choose O iff she has positive degree of belief that O is in some respect good, valuable, or choiceworthy.

As on the Humean conception, the enkratic conception of rationality admits some variation when it comes to subjective oughts/rational requirements. I have chosen to call this conception of rationality "enkratic" because, unlike the Humean conception, it validates the *enkratic principle*, which asserts (in the idiolect I have introduced) that if an agent fully believes that she objectively ought to choose an option O, then she is rationally required to choose option O. (Analogously, on the Humean conception, an agent who fully believes that choosing O will best satisfy her desires is rationally required to choose O.) But this is only a limiting case of rational requirement, that must be strengthened to yield a complete theory. Paralleling the standard Humean view on which agents are rationally required to maximize expected desire satisfaction, one version of the enkratic conception holds that agents are rationally required to *maximize expected choiceworthiness* (Wedgwood, 2013). But as with the Humean conception, other variations are possible, as we will see in Chapters 2 and 7.[10]

The Humean conception of objective reasons/oughts emerges as a limiting case of the enkratic conception, if one believes that the goodness, value, or choiceworthiness of an option consists exclusively in its capacity for desire satisfaction. But one

[10] Just as the enkratic conception of rational requirements will, in one way or another, generalize the enkratic principle, so it will take as the object of rational prohibition the classic philosophical notion of *akrasia*. To act akratically is to act contrary to one's beliefs about one's objective reasons for acting, meaning most centrally to do something that one is certain one objectively ought not do, or fail to do something that one is certain one objectively ought to do.

can also hold the Humean conception while acknowledging the existence of desire-independent goodness, value, or choiceworthiness, if one holds that the reason-giving force for an agent of these properties of her options is mediated by her desires.[11]

Likewise, the Humean conception of *subjective* reasons/oughts emerges as a limiting case of the enkratic conception, if one considers an *agent* who believes that the goodness, value, or choiceworthiness of her options consists exclusively of their capacity to satisfy her desires. For such an agent, her beliefs about the choiceworthiness of her options will always coincide with her beliefs about their capacity for desire satisfaction, and so the two conceptions of rationality will agree about the rational requirements to which she is subject.

Despite this potential for convergence, the Humean and enkratic conceptions of rationality are incompatible. Even if the Humean is in fact correct that all objective reasons for action consist in the action's contributing to the satisfaction of an agent's desires, an agent may certainly *believe* otherwise, and may believe that she has most objective reason to choose option O while believing that some incompatible option P would best satisfy her desires, if she herself is not a Humean.

[11] At the level of objective reasons/oughts, therefore, one who adopts the enkratic conception but does not believe in any form of goodness, value, or choiceworthiness that does not directly depend on an agent's desires must find their view collapse either to the Humean conception or to nihilism about objective reasons. For this reason, I will generally assume that the enkratic conception of rationality is accompanied by a belief in desire-independent goodness, value, or choiceworthiness and a belief that these properties constitute or furnish desire-independent reasons for action.

While this rules out the strong desire-dependence entailed by the Humean conception, it does not rule out lesser forms of agent-relativity. I might have objective reason to choose O because O is (or will result in a state of affairs that is) *good for me* (relative to me, from my perspective) rather than good *simpliciter*, and O might be good for me in virtue of comporting with my values, plans, projects, etc. Just how much agent-relativity the enkratic conception can allow before collapsing into a version of the Humean conception is a question I will leave to the side. The important point is that, on any interesting version of the enkratic conception, the mere fact that one of my desires or other motivational states would be satisfied by choosing O is not *ipso facto* a reason for me to choose O. For a defense of this claim and a representative presentation of the enkratic conception of rationality, see for instance Quinn (1993).

A would-be religious ascetic might sincerely believe that he has most objective reason to fast in the desert even while believing what would best satisfy his *desires* is a life of drunken revelry. On the enkratic conception of rationality, this agent's belief that he has objective reason to fast in the desert gives him subjective reason to fast in the desert, even if this belief is utterly unmotivating and even if there is no part of his motivational set that favors fasting in the desert (that is to say, even if he is in this instance fully akratic). Conversely, on the Humean conception, the agent's belief that drunken revelry would satisfy some of his desires provides some reason for him to choose drunken revelry, even if he believes that there is nothing good or choiceworthy about this option and therefore that he has no objective reason to choose it.

Further, suppose the agent recognizes in himself *some* desires that would be satisfied by fasting in the desert, but believes that his desires would overall by better satisfied by drunken revelry. And conversely, he recognizes that he has *some* objective reasons to choose drunken revelry, but believes that his objective reasons to fast in the desert are much stronger. In this case, all else being equal, the enkratic conception of rationality implies that this agent has *more* subjective reason to fast in the desert while the Humean conception implies that he has *more* subjective reason to choose drunken revelry. Thus, the two conceptions will disagree about what course of action this agent is rationally required to adopt.

It should be obvious that, even at the coarse-grained level at which I have so far described these two views, there is much that one might dispute internal to either view. That is, I do not claim that everyone who takes a broadly Humean view of rationality subscribes to every claim I have packed into the Humean conception, and likewise for the enkratic conception. What I do claim, admittedly without much argument, is that all philosophical theories of reasons and rationality represent elaborations of or variations on these two basic conceptions. This claim will be substantiated, if at all, mainly by the use to which I am able to put the Humean/enkratic distinction in the following pages. But to lend it some initial plausibility, let me make three brief observations.

First: It is often said that the conceptual core of rationality is *coherence*, a notion which in some way generalizes the more limited notion of (mere logical) *consistency*. Theories of rationality, then, will be defined by what combinations of mental states apart from inconsistent belief they identify as *incoherent*. It seems to me that there are just two alleged paradigms of incoherence in the practical domain:

1. I have desires that induce a certain ranking of states of affairs - I believe that choosing option O will result in a more highly desired state of affairs than if I choose incompatible option P - and yet I choose (or form an intention to choose/perform) option P.

2. I believe that option O is better, more choiceworthy, favored by stronger objective reasons than option P - and yet I choose (or form an intention to choose/perform) option P.

These two forms of alleged incoherence correspond to the Humean and enkratic conceptions of rationality respectively. The Humean identifies certain triples of $\langle$*motivational state*, *belief about what will best satisfy my motivational state*, *choice/intention*$\rangle$

as incoherent, while the enkratic theorist identifies certain pairs of ⟨*normative/evaluative belief*, *choice/intention*⟩ as incoherent. It is of course possible to hold that *both* of these represent forms of incoherence, but only on pain of concluding that someone like our imagined ascetic, who believes that one course of action would best satisfy his desires while believing that he has most objective reason to perform another course of action, is doomed to incoherence whatever course of action he chooses. The best way to argue for one conception of rationality over the other—and the method I will adopt when I argue for the enkratic conception in the next chapter—is to argue that in cases like these, one resolution of the practical dilemma looks more coherent than the other.

Second: The enkratic/Humean distinction emerges naturally in another way, from the idea of *direction of fit*. A Humean sees desires as exhibiting predominantly world-to-mind direction of fit. While the normative standards that govern desires will of course often call on me to revise my particular desires in light of new information about the world (e.g. the information that the substance in this glass is gasoline rather than gin), my basic, non-instrumental desires are subject to revision only in light of other desires, to meet standards of internal coherence and consistency. Once that set of basic desires has been made maximally coherent, my job as a practical agent is to shape the world so as to best fit those desires. According to the enkratic conception of rationality, on the other hand, desires and other motivational states are governed by norms that have mind-to-world direction of fit: that is, there are normative standards governing my basic, non-instrumental desires beyond the standards of internal coherence, standards that direct me to desire certain things in the world in virtue of properties that have nothing to do with the relationship between those things and my existing desires. "Value" or "choiceworthiness" are merely convenient shorthands for these properties, whatever they turn out to be, that direct me to form a basic, non-instrumental desire for something.

Third and finally: The distinction between Humean and enkratic views of rationality tracks and, I believe, refines the more standard division between *internalist* and *externalist* theories of practical reasons, a distinction famously introduced by Williams (1979). The reason internalist holds that all reasons for action are contingent, perhaps in very complex ways, on the contents of an agent's motivational set, while the reason externalist holds that some reasons for action are not contingent in this way. The Humean/enkratic distinction is more fundamental than the internal/external distinction, it seems to me, because a reason for action can *exist in virtue of* an agent's desires without being *contingent on* the particular content of those desires. Williams, for instance, defends reason internalism in a way that largely takes for granted the Humean conception of reasons (or at least, that ignores the possibility of the enkratic conception): He claims that for an external reason statement to be true, it would have to be the case that any agent who deliberated well enough would eventually come to be motivated to act in a certain way. And this is indeed (at least to a first approximation) what it would take for an external reason statement to be true, *if* the Humean conception of reasons is right. But one who held, for instance, that any agent who deliberated well enough would develop at least some motivation to avoid physical pain, despite counting as a reason externalist under Williams' schema, would be better classified with Humeans like

Williams than with those who hold that reasons for action need have nothing at all to do with an agent's actual or potential motivations.

To summarize: On the Humean conception, objective reasons for action exist in virtue of an agent's desires or other motivational states, and agents are rationally required to (for instance) maximize expected desire satisfaction. On the enkratic conception, objective reasons for action exist in virtue of desire-independent evaluative features of the world (goodness, value, choiceworthiness...), and agents are rationally required (for instance) to maximize the expected satisfaction of their objective reasons, desire-based or otherwise, and thus in particular to choose options that they are certain they objectively ought to choose, even when they are also certain that the option in question would not best satisfy their desires.

The divide between the Humean and enkratic conceptions of rationality is a theme to which we will repeatedly return. Whether these views indeed represent rival conceptions/theories of the same basic concepts, or instead represent fundamentally different *concepts* (e.g., each laying claim to the expression "objective reason" to pick out essentially unrelated concepts of "contributing to desire satisfaction" and "value/choiceworthiness," respectively) is an important question that we will take up in Chapter 6. But for the moment I want my vocabulary, at least, to remain neutral between the two views and so I will treat them as rival conceptions of one underlying concept. (The possibility of understanding the two views as rival elaborations of an underlying idea of rationality-as-coherence provides at least provisional justification for this choice.)

I have of course left a great deal unsaid about the nature of reasons and oughts. Some of the details will be filled in by subsequent chapters, others will not. I can only hope that the questions I leave unanswered—in particular, the great many questions concerning kinds of reasons, defeators to reasons, and the ways in which reasons can interact to generate on-balance or all-things-considered reasons/oughts/requirements—will not affect the substance of my arguments in ways I've failed to anticipate.

1.5 Moral Reasons

Fundamentally, the topic of the coming chapters will be the problem of *normative* uncertainty, which can be roughly characterized as uncertainty about one's objective reasons that is not a result of some underlying empirical uncertainty (uncertainty about the state of concretia). However, I will confine myself almost exclusively to questions about *moral* uncertainty: uncertainty about one's objective *moral* reasons that is not a result of etc etc. This is in part merely a matter of vocabulary: "moral uncertainty" is a bit less cumbersome than "normative uncertainty," a consideration that bears some weight when the chosen expression must occur dozens of times per chapter. It is also in part because the vast majority of the literature on normative uncertainty deals specifically with moral uncertainty, and because moral uncertainty provides more than enough difficult problems and interesting examples, so that there is no need to venture outside the moral domain.

Additionally, however, focusing on moral uncertainty is a useful simplification that allows us to avoid difficult questions about the relationship between moral and non-moral reasons (though I am hopeful that the theoretical framework I develop can be applied straightforwardly to normative uncertainties of a non-moral kind). For myself, I have no taste for the moral/non-moral distinction: To put it as crudely and polemically as possible, it seems to me that all objective reasons are moral reasons. But this view depends on substantive normative ethical commitments that it is well beyond the scope of this dissertation to defend. So I will make no use of this claim, or indeed of any claims about the relationship between moral and non-moral reasons (excepting some relatively modest claims required by the discussion of supererogation in Chapter 4).

If one does think that all reasons are moral reasons, or that moral reasons always override non-moral reasons, then a complete account of how agents ought to act under moral uncertainty can be given without any discussion of non-moral reasons (Lockhart, 2000, p. 16). To the extent that one does not share either of these assumptions, theories of choice under moral uncertainty must generally be qualified with "insofar as there are no relevant non-moral considerations."

1.6 Simplifying Assumptions

One simplifying assumption that will generally be in force as we proceed, then, is that non-moral reasons are inert in whatever cases we consider.

Another is that the practical agents we consider have no relevant empirical uncertainties. For example, if we consider an agent who has some degree of belief in a consequentialist moral theory, it will generally be assumed that she is not relevantly uncertain about the consequences of her various options. This is in general an innocuous simplification, though we will encounter a few cases in which potential interactions between moral and empirical uncertainties require examination.

Most importantly, we will generally assume that the decision-relevant beliefs of the agents we consider—both their empirical and normative beliefs—are *conclusively justified*, i.e. epistemically/theoretically rational and more rational than any alternative beliefs the agent might have on the subject matter in question. This is in order to sidestep the debate between subjectivism and evidentialism/prospectivism about rational requirements and the more recent debate between "narrow-scope" and "wide-scope" principles of rational requirement.[12] When an agent A has an unjustified belief that she objectively ought to choose option O, the subjectivist/narrow-scoper says that A is rationally required to choose O, while the evidentialist/wide-scoper says that A is not rationally required to choose O but is instead rationally required to revise her beliefs (and, presumably, to then choose whatever practical option is favored by her new, justified doxastic state).[13]

[12] I take these debates to be closely related. If one accepts a narrow-scope principle of practical rational requirement like "If A believes that she objectively ought to choose O, then she is rationally required to choose O," one is left with the subjectivist conclusion that what an agent is rationally required/subjectively ought to do depends simply on her beliefs and not on the *evidence* for or against those beliefs. If on the other hand one accepts only the wide-scope principle "It is rationally required that, if A believes that she objectively ought to choose O, she chooses O," then the evidentialist conclusion seems natural: If A's evidence does not support her belief that she objectively ought to choose O but rather supports the belief that she objectively ought to choose P, the best way of satisfying her rational requirements, it seems, is to abandon the belief that she objectively ought to choose O, adopt the belief that she objectively ought to choose P, and then choose (or form the intention to choose) P.

[13] I confess myself somewhat skeptical whether the debates between subjectivists vs. evidentialists on the one hand, and narrow- vs. wide-scopers on the other hand, are substantive. The evidentialist/wide-scoper wants principles that express the joint requirements of theoretical and practical rationality, while the subjectivist/narrow-scoper prefers to keep principles of practical rationality separate from principles of theoretical rationality. But both parties will agree that an agent with unjustified beliefs should, rationally, revise her beliefs, and that if she is ideally theoretically and practically rational, she will revise her beliefs and then act rationally in light of her

When the antecedent of a principle of rational requirement concerns an agent's beliefs about her objective reasons (which, on the enkratic view for which I will argue in the next chapter, is the form taken by all principles of practical rationality), the wide-scope formulation of that principle will collapse into the narrow-scope formulation if we assume that the agent has conclusive reason to hold the belief mentioned in the antecedent. For instance, consider the wide-scope principle: "It is rationally required that, if A believes she objectively ought to choose option O, then she chooses option O." This principle can be equivalently stated in disjunctive form: "It is rationally required that *either* A does not believe that she objectively ought to choose option O *or* A chooses option O." But if this principle holds, and it is *also* the case that A is rationally required to believe that she objectively ought to choose option O (because she has conclusive evidence for that belief), then it follows that she is rationally required to intend to choose option O, since forming this intention is the only way for her to satisfy all of the rational requirements that apply to her. Thus the wide-scope principle implies the following narrow-scope principle: "If A has conclusive reason to believe that she objectively ought to choose option O, then she is rationally required to choose option O." Thus, by limiting our attention to cases where the agent has conclusive justification for her relevant beliefs, we can generally sidestep debates about the scope of rational requirements. I will therefore typically adopt narrow-scope formulations for the sake of convenience, with the understanding that these can be seen as limiting cases of wide-scope principles if

new, justified beliefs. So the views do not differ in terms of their ultimate practical implications.

that is to one's taste.[14]

Each of these simplifying assumptions (no relevant non-moral reasons, certainty about all relevant empirical matters, and conclusively justified normative and empirical beliefs) will be dropped at various points, as needed, but it should be assumed that they are in force unless otherwise specified.

1.7 Substantive Assumptions

Along with these simplifications, I should acknowledge at the outset two important philosophical commitments of which I will make regular use, even though they will go largely undefended. The first, which has already been introduced, is the assumption that agents subjectively ought/are rationally required to do what they have all-things-considered subjective reason to do. I take this to be a conceptual truth, but it is not uncontroversial, and defending it lies beyond the scope of our discussion.

The other important philosophical assumption, to which I have so far only alluded, is a standard Bayesian picture of rational belief: namely that (i) belief comes in cardinal *degrees*, (ii) it is a requirement of theoretical rationality that an agent's degrees of belief satisfy the probability calculus, and (iii) agents should

[14] To the extent that there is a substantive choice between narrow-scopism and wide-scopism about rational requirement, I incline toward the narrow-scope view, for reasons much like those given in Kolodny (2005): Principles of rational requirement ought to specify the ways in which an agent can respond-by-reasoning to her situation, and while an agent can reason her way from the belief that she objectively ought to choose option O to an intention to choose O (or, I think, simply to *choosing* O), she *cannot* reason her way from her lack of an intention to choose O (or her failure to choose O) to giving up the belief that she objectively ought to choose O. If the most rational thing for her to do is to give up this belief, it is in light of a *different* rational requirement, a requirement of theoretical rationality that concerns how she ought to respond to her evidence or her beliefs about evidence.

update their beliefs by some form of Bayesian conditionalization (Talbott, 2015). The first two components of the Bayesian picture will be crucial to the view I eventually defend. The third is important only insofar as it lends support to the requirement of regularity, i.e., that a rational agent should not assign minimal or maximal degrees of belief (0 or 1) to any propositions apart from logical truths.[15]

I will refer to degrees of belief interchangeably as degrees of belief, credences, or subjective probabilities. I assume that an agent's degree of belief in a proposition corresponds to something like the odds she is willing to accept in betting on the truth of that proposition. In its naive form, this "betting interpretation" of subjective probabilities is subject to several powerful objections (for an overview, see §3.3 of Hájek (2012)), and the refinements needed to meet these objections require strong idealizing assumptions. Indeed, it must be conceded that the Bayesian picture in general is at best an enormous idealization that ordinary human agents can only hope to very roughly approximate in the best of circumstances. But I do think that it's the *right* ideal to approximate, and I know of no alternative that is more psychologically realistic while also representing a plausible normative ideal. So in attempting to develop an ideal theory of rational choice under moral uncertainty, I will help myself to a manifestly ideal(ized) theory of the nature of that uncertainty. (At the end of Chapter 7, I will briefly how non-ideal agents can attempt to

[15] Often, regularity is characterized more loosely, as assigning 0 and 1 only to *necessary* or *a priori* truths. But insofar as we should be uncertain about the basic principles of morality, and basic normative principles generally, even though the truths of these domains are presumably necessary rather than contingent and arguably *a priori* rather than *a posteriori*, it is useful for our purposes to make the more demanding regularity assumption.

For a typical defense of regularity as a requirement of rational belief, see for instance Hájek (2003, pp. 31-2).

approximate the requirements of ideal rationality.)

1.8 Moral Theories

Throughout the following chapters I will speak mainly of uncertainty concerning moral *theories*, rather than moral principles or particular moral dilemmas. By "theory," however, I mean simply a maximal consistent set of propositions in a given domain—hence, a moral theory is a maximal consistent set of moral propositions. Understood in this way, I think, the language of theories is innocuous, and does not exclude or bias the playing field against "anti-theoretical" views like common-sense morality or moral particularism.[16] An agent who is uncertain, for instance, whether it is permissible to tell a white lie to her friend in the particular choice situation divides her belief between at least two moral theories, one of which includes the proposition that it is permissible and another of which includes the proposition that it is not. If she is uncertain about many moral questions, then she divides her beliefs between many moral theories.[17]

[16] Thanks to Christopher Morris for pressing me on this point.

[17] Granted, there is a hefty dose of Bayesian idealization in the claim that, say, a particularist distributes her moral belief over maximal consistent sets of moral propositions. The particularist denies that morality is *axiomatizable*, i.e., that there is any finite (or at any rate, reasonably small) set of true moral propositions of which all the rest are logical consequences. Therefore she will presumably hold that the true moral theory (i.e., the maximal consistent set of *true* moral propositions) is not the sort of thing that any human agent could consciously entertain, and perhaps even that it is too complex to be represented in the unconscious resources of the human mind (e.g., in the form of non-occurrent beliefs). Nevertheless, throughout the dissertation (except briefly in Chapter 7) my focus will be on ideal rationality, setting aside the further question of how bounded, finite human agents are to approximate that ideal. Therefore I will assume that agents assign degrees of belief to all propositions that their conceptual resources enable them to entertain, including moral propositions. As long as there are only finitely many *atomic* moral propositions that such an agent can represent with her conceptual resources, she will assign degrees of belief to maximal consistent conjunctions of literals (atomic moral propositions or their negations), which amounts to an assignment of belief to moral theories. (It is worth reiterating here that I am helping myself freely to talk of "truth" on the assumption that those who deny the truth-aptness of moral

In fact, if we want a general theory of rational choice under *uncertainty* that treats normative, empirical, mathematical, and other forms of uncertainty in a continuous, unified manner, then that general theory will make reference not to *moral* theories but simply to *theories of the world*, i.e., maximal consistent sets of propositions without any domain restriction. Since we will generally be operating under the simplifying assumption that the agents we consider are *only* uncertain about moral matters, and experience no uncertainty (or at least, no relevant uncertainty) about any non-moral matters, I will mostly set this point aside and speak simply of moral theories. But in Chapter 6 and Appendix B we will have some reason to think of agents as distributing belief over complete theories of the world that include propositions of every kind, moral and otherwise.

Finally: In Chapter 2, I will defend the enkratic conception of rationality as more plausible than the Humean conception, and thereafter it will sometimes be convenient to speak of "value" or "choiceworthiness" in place of "objective reasons." In the idiolect of subsequent chapters, unless otherwise specified, the *value* or *choiceworthiness* of an option is just the degree to which it is supported by objective reasons, relative to other options.[18] I will also sometimes speak of the "subjective choiceworthiness" of an option as a convenient locution for the degree to which that

propositions will nonetheless supply some truth-surrogate that permits us to say things analogous to what we can say in the language of truth.)

[18] There is a narrow sense of terms like "value" or "choiceworthiness" on which these are, as a conceptual matter, desire-independent features of the world or properties of practical options, and it was to this sense that I appealed above in attempting to distinguish the enkratic from the Humean conception of rationality. But a reader who remains unconvinced by the anti-Humean arguments of the next chapter should be able to understand these terms, thereafter, as simply synonyms for "objective reason," leaving open the possibility that all objective reasons involve and depend on the satisfaction of desires.

option is supported by *subjective* reasons. The *value scale* of a moral (or other normative) theory is the structure provided by that theory for comparing degrees of objective reason: a set of possible values that the theory assigns to practical options, together with a more or less enriched ordering on that set. For instance, classical utilitarianism has a cardinal value scale, with practical options assigned real-numbered values over which are defined not only the relation $\geq$ but also the operations $+$ and $-$.[19] But one way of conceiving certain deontological theories, for instance, is as ordinally structured, ranking options as better or worse while denying that there are meaningful quantitative answers to the question "How *much* better/worse?"[20]

A theory's *value assignment*, then, is its mapping from practical options to points on its value scale. Where convenient, I will denote the value assigned by a theory T to a practical option O by the expression $V^T(O)$. It will also occasionally be convenient to speak of the value assigned to *payoffs* (increments of some value-bearing phenomenon, like a gain or loss of one hedon) using the same formalism

[19] Whether multiplication and division are defined depends on whether the utilitarian value scale is an *interval scale* (giving cardinal degrees of difference between values but having no privileged or non-arbitrary zero point and hence not giving *ratios* between absolute degrees of choiceworthiness) or a *ratio scale* (giving both cardinal degrees of difference between values and a non-arbitrary zero point, and hence permitting talk of ratios between values). For my part it seems most natural to conceive of consequentialist theories as having ratio structure, since if those theories are *aggregative* (determining the value of the states of affairs that might result from an agent's actions by summing the values instantiated at multiple, independent value locations, e.g. persons or spacetime regions), then there is a non-arbitrary zero point on the value scale, namely, an empty universe that does not contain any value locations at all. For a risk-neutral expected value maximizer, however, only the interval structure of a normative theory is relevant, so it will generally make no difference whether we conceive of these theories as interval or ratio.

[20] For detailed discussion of the distinction between ordinal, interval, and ratio structures in the context of moral theories and its relevance to rational choice under moral uncertainty, see MacAskill (2014), pp. 14-15, 52ff. We will return to these questions in greater detail in Chapter 6, where I argue that many other structures are possible as well—indeed, all possible enrichments of the very simplest structure a normative theory can possess, namely, a *binary* structure that simply classifies options as permissible or impermissible.

(e.g., V^T(one hedon)). Strictly speaking, such expressions should be understood to denote the *difference* in value between practical options that are equivalent in all morally relevant respects but for the payoff in question (e.g., a pair of options that produce all the same consequences, except that one produces an additional hedon that the other does not).[21]

In addition to an objective value assignment, a normative theory may also include one or more additional value assignments that are "subjectivized" to various degrees, i.e., relativized to various aspects of an agent's credal state. For instance, hedonistic utilitarianism gives an objective value assignment according to which the value of any practical option is equal to the net hedonic value of its actual (or counterfactual) consequences, but also a partially subjective value assignment according to which the value of any practical option is equal to the *expected* net hedonic value of its consequences.[22] As we will see, a large part of the problem of decision-making under moral uncertainty has to do with the *aggregation* of value assignments, i.e., the method by which an agent goes from a *probability distribution* over value assignments to a new value assignment that represents the overall degrees of subjective reason for choosing particular practical options in light of that probability distribution.

[21] This latter locution is intelligible, and will be used, only in the context of theories that recognize value differences between options more fine-grained than "better" or "worse"—e.g., in the context of cardinally structured theories but not in the context of merely-ordinal theories.

[22] While our primary interest will be in the *fully subjective* value assignment that compares an agent's subjective reasons for choosing her various options in light of the *entirety* of her credal state, I will argue in Chapter 6 for a *multi-stage aggregation procedure* on which we have reasons to be interested in partially subjective value assignments, e.g., the value assignment to an agent's options conditional on a particular moral theory in which she has positive credence, which takes account of her empirical uncertainties but not her moral uncertainties. We will also consider value assignments that are conditioned on *disjunctions* of moral theories, which play an important role in the multi-stage aggregation process.

Again it is worth emphasizing that talk of value scales and value assignments does not presuppose any tendentiously theoretical conception of ethics or rule out anti-theoretical views like particularism. All that is presupposed is that a maximal consistent set of moral propositions should, in one way or another, assess and compare an agent's (moral) reasons for or against particular practical options.

1.9 Summary of the Dissertation

Having at least roughly described the conceptual framework and philosophical background assumptions on which I will draw in the coming chapters, we are now ready to engage with the problem of moral uncertainty proper.

I begin in the next chapter by making the positive case for the relevance of moral uncertainty to rational decision-making. I will describe and defend two principles of rationality, one a strengthening of the other, but either of which is enough to establish the need for a theory of rational choice that is sensitive to an agent's moral beliefs (on the weaker principle) and/or degrees of belief (on the stronger principle). Chapter 3 then attempts to answer some challenges that have recently been forwarded against this conclusion by philosophers like Brian Weatherson and Elizabeth Harman who hold that what an agent ought to do, in every important sense, depends on the moral facts and the agent's empirical beliefs but not on her moral beliefs or degrees of belief.

Having addressed these challenges, we can take up the question of *what* principles of rational choice should guide an agent who is uncertain about the requirements of morality. We will begin in Chapter 4 by considering the capabilities and limitations of approaches based on *dominance reasoning*, which have recently received significant attention from philosophers. I will conclude that, contra worries in this recent literature, interesting normative consequences concerning moral nihilism, supererogationism, and certain controversial moral topics like meat consumption can be derived from relatively weak and uncontroversial dominance principles, but also that these dominance principles only have interesting scope for application if they can be construed as limiting cases of some more general (e.g. expectational) theory of choice under moral uncertainty.

Chapter 5 then embarks on the task of describing this more general theory. The chief obstacle to this task is the problem of finding a non-arbitrary basis for intertheoretic comparisons of moral value. I argue, however, that rather than looking for a single, maximally general principle to govern such comparisons (as Lockhart and others have attempted), we should take a bottom-up approach that starts from what I will call "comparability classes" of moral theories, within which there are clear and rationally perspicuous bases for intertheoretic comparison. Any general theory of moral uncertainty, I suggest, must be compatible with the clear and compelling results we find in these more limited cases of uncertainty between closely related and structurally similar accounts of morality.

Chapter 6 will then attempt to flesh out the general approached introduced in Chapter 5 into the outline of a general theory of rational choice under moral uncertainty. I will distinguish three obstacles to aggregation of rival moral theories under uncertainty, enumerate possible responses to each of these obstacles, and argue for my own preferred response to each. From this will emerge a novel picture of decision-making under moral uncertainty, though with many important details still left to be filled in.

Finally, in Chapter 7, I take up one of the most difficult problems for any approach to moral uncertainty. This is the so-called "regress problem," which stems from the observation that, just as an agent can be uncertain about first-order normative principles, she can likewise be uncertain about second-order principles of choice under uncertainty like those proposed in Chapters 5 and 6. Therefore, it seems, the principles of rational choice that demand sensitivity to moral uncertainty in the first place force a rational agent to deliberate not just on the basis of second-order principles that accommodate her first-order uncertainty, but also on the basis of third-order principles that accommodate her second-order uncertainty, fourth-order principles that accommodate her third-order uncertainty, and so on *ad infinitum*. In response to this challenge, I argue that at least some normative principle(s) must possess "external," belief-independent normative force, but that, contra first-order moral externalists like Brian Weatherson, this external status is best attributed to a fundamental principle of *rationality*: the enkratic principle, correctly formulated. I then distinguish two aspects of the regress problem, one having to do with ideal rationality and the other with non-ideal rationality, and argue that this "enkratic externalist" position enables solutions to both problems.

Chapter 2: Rationality and Moral Belief

In this chapter and the next, I aim to establish the *need* for a theory of rational decision-making under moral uncertainty. To do this, I will defend a weaker and then a stronger thesis: first, that an agent's full moral beliefs give her (subjective) reasons for action and second, that her partial moral beliefs (i.e., positive degrees of moral belief) give her (subjective) reasons for action. If the weak thesis is correct, then a complete theory of rational choice must take account of an agent's moral beliefs. If the strong thesis is correct, then a complete theory of rational choice must take account of her moral *uncertainties.* This establishes the project to be pursued in the remainder of the dissertation as one worth pursuing.

In this chapter I spell out the two theses more precisely, then describe what I take to be *prima facie* compelling arguments for each thesis. In the next chapter, I will consider arguments for the practical irrelevance of moral belief and/or moral uncertainty raised by Brian Weatherson, Nomy Arpaly and Timothy Schroeder, and Elizabeth Harman, and argue that none of these challenges is as compelling as the *prima facie* case for the two theses.

Throughout, my goal will be to establish that an agent's moral beliefs and degrees of belief are relevant to what she *subjectively ought to do* or *is rationally* *required to do*, in the sense of these terms introduced in the previous chapter. By contrast, I do *not* aim to show that an agent's moral beliefs and uncertainties are relevant to what it is *right* for her to do, or what she would be *praiseworthy* or *blameworthy* for doing, or to what it would be *virtuous* for her to do, etc. One or more of these notions may turn out to have some necessary connection to subjective oughts/rational requirements, such that what I say about the latter concepts will have implications concerning the conditions for rightness, blameworthiness, virtue, or the like. But I will make no claim to this effect.

By making this stipulation, I run some risk of talking past my philosophical adversaries, who as we will see often appear to be concerned in the first instance with characterological notions or with praise and blame rather than with subjective rationality. But my goal is simply to vindicate the project of looking for principles of decision-making that account for an agent's moral beliefs and uncertainties, and insofar as moral uncertainty is relevant to subjective oughts/rational requirements, this project is vindicated. Conversely, insofar as the philosophers to whom I respond are rejecting this project (and not merely declaring it irrelevant to some other project, like theories of blame and responsibility), we are in substantive disagreement. As will become clear in the next chapter, Weatherson and Harman at least *do* seem to reject the project of looking for moral-belief-relative norms, and hence reject the claim that moral beliefs and degrees of belief are relevant to rational requirement, so we can focus on the concept of rational requirement without missing the point of their arguments.

2.1 Moral Belief and Reasons for Action: Two Theses

These, then, are the two principles that I will defend in this chapter and the next:

Weak Subjective Rationalism (WSR) If an agent A *fully* believes that she is objectively morally required to choose option O, and is conclusively justified in so believing, then A thereby has some subjective reason to choose O.

Strong Subjective Rationalism (SSR) If an agent A *partially* believes that she is objectively morally required to choose option O (i.e., has positive degree of belief that she is so required), and is conclusively justified in so believing, then A thereby has some subjective reason to choose O, with the strength of this reason covarying monotonically with A's degree of belief.

Put intuitively, WSR asserts *the rational relevance of moral belief* and SSR further asserts *the rational relevance of moral uncertainty.* Assuming that full belief can be understood as any degree of belief that equals or exceeds some "Lockean threshold" L, WSR asserts that an agent whose credence that she is objectively required to choose option O is greater than or equal to L has thereby at least some subjective reason to choose O, meaning among other things that *ceteris paribus* (absent any other subjective reasons for or against choosing O or any of its alternatives) she is rationally required to choose O.[1] SSR asserts that an agent who assigns any

[1] This Lockean threshold view, it seems to me, is the best way of accommodating the notion of full belief within a Bayesian framework. Nevertheless, is carries certain counterintuitive implications, most importantly that rational belief does not *agglomerate*: a fully rational agent may

positive probability to the proposition that she is objectively required to choose O has some subjective reason to choose O, which increases in strength as that probability increases, meaning among other things that an agent who is uncertain whether she is objectively morally required to choose option O or objectively morally required to choose incompatible option P thereby has both some subjective reason to choose O and some subjective reason to choose P, reasons which must be accounted for in determining what she subjectively ought to do, all things considered.

Note that neither WSR nor SSR denies that there may be *other* sources of subjective reasons entirely unrelated to an agent's beliefs about her objective moral obligations and even entirely unrelated to an agent's beliefs generally. Thus, these principles only assert the rational *relevance* of full/partial moral belief, not that such belief is the *only* determinant of rational requirement.

Given this limitation, the weaker principle WSR may appear so trivial that no one could deny it. But in fact, Harman (2015) does pretty clearly deny WSR, and it's possible that Weatherson does as well.[2] More importantly, though, even WSR "gets us in the game" of looking for norms of rational choice that are sensitive to an agent's moral beliefs, contra the spirit of Weatherson's and Harman's critiques, both of which criticize the project of seeking such norms (rather than, for instance, claiming

believe P, believe Q, but fail to believe P & Q. Rejecting agglomeration is, to my mind, the most attractive solution to the lottery paradox: Assuming that one wishes to allow for full belief in anything other than the limiting case of absolute certainty, then given a fair lottery with a billion tickets, a rational agent ought to believe that Ticket 1 will not win, ought to believe that Ticket 2 will not win, ... but *ought not* believe that none of the tickets will win. Defending this view is beyond the scope of the present discussion, but nothing I have to say in defense of WSR will, as far as I can tell, turn on the exact nature of full belief.

[2] See the quoted statement of Harman's view that opens §3.5 in the next chapter. Weatherson's view is complex and difficult to situate relative to my theses, but I make the attempt in footnote 1 of Chapter 3.

that these norms should be understood *only* as norms of rational requirement and not of some other normative concept like blameworthiness).[3]

Strictly speaking, WSR must be conjoined with some minor auxiliary premises to establish the relevance of moral beliefs to rational requirements. If agents are rationally required to do what they have most subjective reason to do, then given WSR the reason to choose O generated by an agent's full belief that she is objectively morally required to choose O will sometimes produce a rational requirement to choose O—when no other reasons are present, or when the balance of other reasons is against choosing O, but close enough that the reason postulated by WSR tips the scales in its favor. This could fail to be the case only if reasons of some other kind are always present and always decisive: for instance, one might in principle allow that an agent's moral beliefs give her subjective reasons, but hold that the combination of her empirical beliefs with the true moral theory always gives her *stronger* subjective reasons, such that the reasons given by her moral beliefs can never make the difference in deciding what she is rationally required to do. While this view is consistent, I can see no motivation for defending it and know of no

[3] In a more recent article, Brian Hedden (2016) gives the clearest articulation of the anti-WSR position to date. On the initial statement of his view, "There is no normatively interesting sense of *ought* in which what you ought to do depends on your uncertainty about normative facts" (p. 3). It is clear from the arguments Hedden marshals, however, that he means to deny the relevance (to any "normatively interesting sense of *ought*") not just of moral uncertainty but of moral belief generally. Indeed, later in the paper he confirms that he is defending this stronger position, in contrasting his view with Weatherson's: "Weatherson is arguing that moral beleifs do not affect what you *morally* ought to do, but denying that moral beliefs affect what you morally ought to do is compatible with thinking that there is some other sense of *ought* (e.g., not the primary moral *ought*, but rather a super-subjective *ought*), in which what you ought to do depends on your moral beliefs. It is this latter claim that I am denying" (p. 23n). This distinction notwithstanding, it seems clear to me that Weatherson as well as Hedden and Harman would be unwilling to concede that the all-things-considered rational *ought* that I have been calling the "subjective ought" is sensitive to an agent's moral beliefs or uncertainties.

one who has tried. So while my defense of WSR and SSR will allow for subjective reasons of other kinds that are insensitive to an agent's moral beliefs, I will take for granted that these other reasons are not *always* rationally decisive, and hence that if WSR/SSR are true, then an agent's moral beliefs/degrees of belief will *sometimes* generate rational requirements.

As WSR implies that moral *belief* is relevant to rational requirements, so SSR implies that moral *uncertainty* is relevant to rational requirements. For instance, if SSR is true, then an agent's worry that it *might* be morally impermissible to eat meat will provide her with some subjective reason to avoid eating meat, and where no other reasons are present or the balance of those other reasons does not strongly favor eating meat, this reason may prove decisive, generating a rational requirement against eating meat.

WSR and SSR are, by design, somewhat independent of the gradations of viewpoint most commonly drawn in the literature on moral uncertainty. Among those who accept that an agent's moral beliefs are relevant to what she subjectively ought to do, the most modest position in the standard taxonomy is the view Lockhart (2000) calls "My Favorite Theory" (MFT): An agent subjectively ought to do whatever is prescribed by the moral theory to which she assigns the highest degree of belief.[4] The next most modest position is a view sometimes called "My Favorite Option" (MFO): An agent subjectively ought to take whatever action she judges most likely to be objectively permissible (where various moral theories that judge

[4] This view is defended under another name by Gracely (1996), and as My Favorite Theory by Gustafsson and Torpman (2014).

the action permissible might contribute to her overall degree of belief that it is so).[5] Finally, there are "hedging" views, according to which what an agent subjectively ought to do is sensitive not only to her degrees of belief about objective rightness and wrongness, but to the degrees of potential rightness and wrongness associated with her various options. Unlike My Favorite Theory and My Favorite Option, for instance, hedging views imply that if an agent believes to degree .8 that she is under a weak obligation to keep a promise to a friend by donating some money to a high school reunion fund, but believes to degree .2 that she is under a much stronger obligation to donate that money to GiveDirectly instead, the comparative strength of her potential obligation to donate to GiveDirectly can make it the rationally required option.

Both WSR and SSR are compatible with any of these three views (though SSR combines very uncomfortably with My Favorite Theory). Importantly, even the stronger SSR does not imply that one must adopt a hedging view, for it does not assert that the strength of the subjective reasons given by partial moral beliefs is in any way related to the degrees of rightness or wrongness that those beliefs assign to an agent's options. While the argument offered for SSR in the next section can be easily seen as an argument for hedging, objections to hedging views in particular (especially, arguments stemming from the problem of intertheoretic value comparisons) might lead one instead to a view like MFO, even if one accepts the argument for SSR. Though in chapters 5-6 I will argue that the problem of intertheoretic

[5] MFO is considered and rejected under this name by Gustafsson and Torpman (2014) and MacAskill (2014). Lockhart considers and rejects the same view, which he calls "PR2" (Lockhart, 2000, p. 26ff).

value comparisons can be at least sometimes overcome and therefore that hedging is at least sometimes the rational response to moral uncertainty, my concern for the moment is not to defend anything more than SSR.

The opponents of the moral uncertainty project, in denying both WSR and SSR, and adopt a position even more conservative than MFT. If what an agent ought to do is determined simply by the true moral theory plus her empirical beliefs/evidence, then the belief that one is or might be objectively required to choose O does not in itself generate any reason for action. For instance, a Kantian who believes (even with certainty) that she is morally required to tell a harsh truth rather than a harmless white lie nevertheless has *no subjective reason at all* to tell the truth or avoid the lie, if in fact utilitarianism is the right moral theory. For lack of established terminology, I will refer to this position as *irrelevantism*: the irrelevantist holds that moral beliefs and degrees of belief are irrelevant to what any agent subjectively ought to do.[6]

In examining the arguments on offer for the irrelevantist position, I will do my best to distinguish between those that challenge only SSR and those that go further and challenge WSR. But first, I will describe what I take to be the compelling *prima facie* basis for accepting the two theses and hence the rational relevance of moral uncertainty. The case for WSR and the relevance of moral belief turns on its connection to the enkratic principle and the enkratic conception of rationality. The further case for SSR and the relevance of moral *uncertainty*, as we will see in §2.3,

[6] Harman, in advocating this view, dubs it "actualism," but I avoid this terminology since it risks confusion with the somewhat-related metaethical debate between actualism and possibilism.

turns on a close analogy between moral and empirical uncertainty.

2.2 The *Prima Facie* Case for WSR: The Enkratic Principle

The enkratic principle, in its most plain-vanilla formulation, asserts the following:

EP1 If an agent *A* believes that she ought to (perform) *O*, then *A* is rationally required to intend to (perform) *O*.

In light of claims made and apparatus introduced in the preceding chapter, we can immediately substitute a slightly revised formulation:

EP2 If an agent *A* believes that she objectively ought to choose option *O*, and is conclusively justified in so believing, then *A* is rationally required to choose option *O*.

EP2 improves on EP1 in three ways: first, by clarifying that it is beliefs about *objective* rather than subjective reasons/oughts that generate rational requirements;[7]

[7] The relationship between the subjective oughts/rational requirements to which an agent is subject and the agent's *beliefs* about the subjective oughts/rational requirements to which she is subject poses difficult questions, which we will say something about in Chapter 7. For the moment, however, it is worth noting that endorsing principles that derive rational requirements from an agent's justified beliefs about *objective* reasons does not involve any commitment with respect to subjectivism vs. evidentialism about subjective oughts. It *does* rule out a certain sort of subjectivism on which an agent always ought to do whatever she *believes* that she subjectively ought to do (at least, it rules out this view on the assumption that an agent could, e.g. via a failure of rationality, believe that she objectively ought to choose option *O* and yet believe that she *subjectively* ought to choose some incompatible option *P* rather than *O*). But it says nothing about the rational requirements to which an agent is subject when her beliefs about what she objectively ought to do are unjustified, i.e., it leaves open both the subjectivist possibility that the agent subjectively ought to act in accordance with her unjustified belief and the evidentialist possibility that she subjectively ought to act in accordance with the evidence that supports a different belief.

second, by limiting our attention to *justified* beliefs and therefore setting aside the subjectivist vs. evidentialist and narrow-scope vs. wide-scope debates;[8] and third, by framing the consequent rational requirement in terms of choice rather than performance or intention.

At least one more revision is required, however, to produce a fully plausible formulation of the enkratic principle. As Wedgwood (2013) points out, a principle like EP2 seems to wrongly imply that if I believe (with, say, 99.9% confidence) that choosing option O will save one life and have no other morally significant consequences, but recognize a 0.1% risk that choosing option O will result in the destruction of a medium-sized city and the deaths of a million innocent people, and am certain that the only alternative option P would have no morally significant consequences at all, I am rationally required to choose option O (since I *believe*, just with something less than certainty, that O will in fact have better consequences than P and is therefore the option I have most objective reason to choose). If the notion of "belief" that figures in EP2 is taken to be full belief in the Lockean threshold sense, as in WSR, then this unwanted consequence follows straightforwardly.[9]

[8] A stronger condition that might set these debates even more conclusively to the side would require that the beliefs in question be *conclusively* justified, i.e., that the agent be epistemically *required* rather than merely *permitted* to hold those beliefs in light of her evidence. This distinction will be important, for instance, if one draws a more basic distinction between *justifying* reasons and *requiring* reasons in the epistemic domain (Gert, 2003). The issues raised by this distinction would be a distraction in the context of the present discussion, but in any case the various principles defended in this chapter suffice to establish the rational relevance of moral belief/uncertainty even with the more rigorous condition of conclusively justified beliefs, so long as one accepts that it is possible for an agent to have conclusive justification for believing a false moral theory.

[9] Accepting this as a counterexample to standard formulations of EP does not require one to accept (begging the question of the present chapter) that purely moral uncertainty can affect what an agent is rationally required to do, but only the more or less uncontroversial point that *morally relevant empirical uncertainty* can affect what an agent is rationally required to do. For anyone who doubts this latter claim, some defense of it will be given in the next section.

Wedgwood proposes replacing the standard formulation of EP with a principle asserting that agents are rationally required to form sets of intentions that *maximize expected choiceworthiness*, and he takes this revised principle to be "the fundamental principle of rationality" (Wedgwood, 2013, p. 505). (MacAskill (2014) also defends a somewhat restricted version of this view, which he dubs *MEC*.) Adapted to our idiolect, Wedgwood's version of EP can be stated as follows:

EP3/MEC Assuming that all of agent A's relevant beliefs are conclusively justified, then in any choice situation, A is rationally required to choose an option with maximal expected choiceworthiness (i.e., such that the probability-weighted sum of possible objective reasons to choose that option is equal to or greater than the equivalent sum for any alternative option).

Such a principle would certainly be suitable to my purposes, lending clear support not only to WSR but also to SSR and indeed to moral hedging. But we need not go nearly so far ourselves to accommodate Wedgwood's objection to standard formulations of EP.

We might also, for instance, simply weaken EP2 by confining its application to cases of moral *certainty* rather than mere belief:

EP4 If an agent A *believes with certainty* (i.e., probability 1) that she objectively ought to choose option O, and is conclusively justified in so believing, then A is rationally required to choose option O.

Of course, if we accept the assumption of regularity then this restricted principle will strike us as entirely uninteresting, since an agent is never justified in believing with probability 1 that she ought to choose option O. As a compromise between the ambitious EP3 and the overly-modest EP4, then, we might adopt the following *ceteris paribus* formulation of EP:

EP5 If an agent A fully believes (in the Lockean threshold sense) that she objectively ought to choose option O, and is conclusively justified in so believing, then A is rationally required to choose option O, *unless* her beliefs about her objective reasons conditional on its *not* being the case that she objectively ought to choose option O generate subjective reasons against choosing O that are strong enough to defeat her subjective reasons for choosing O.

EP4 and EP5 can be seen as limiting cases of EP3, or of some other similarly strong formulation of EP that posits, for instance, a more complex attitude toward risk in place of Wedgwood's risk-neutral, expectation-maximizing view.

Anyone who accepts *any* of these formulations of EP, though, should also accept WSR. Let's focus on EP5 as the most interesting case. Deriving WSR from EP5 requires just two auxiliary premises: first, that an agent who believes that she is objectively morally required to choose option O believes that she objectively ought to choose O, and second, that the *ceteris paribus* rational requirement to choose O described in the consequent of EP5 stems from subjective reasons to choose O which are *overridden*, rather than undercut or otherwise caused to disappear, in cases where the risks associated with option O prevent it from being rationally required.

I see no reason to doubt the latter assumption. With respect to the former, note that this does not require the controversial claim that moral reasons always override non-moral reasons. One might hold, with Portmore (2008), that non-moral reasons can sometimes prevent moral reasons from generating moral requirements (as when the prudential costs of some altruistic act make it morally permissible not to perform that act, even though the moral goods that would result from the act are such that it would be morally obligatory if there were no prudential cost). The claim is only that moral requirements are a species of requirement *simpliciter*, i.e. that to be objectively morally required to choose option O is just to be objectively required to choose option O, for moral reasons.

This should at least make plausible the general claim that, if we accept any plausible formulation of the enkratic principle, we should also accept WSR. Indeed, even without examining particular formulations of EP, it should be clear that WSR is hard to resist if one accepts the general spirit of the enkratic conception of rationality. On the enkratic conception, the paradigmatic form of incoherence and hence irrationality in the practical domain is a mismatch between one's beliefs about what one objectively ought to do and one's intentions or choices. Even allowing for Wedgwood's point that full belief in the Lockean threshold sense does not always generate rational requirements, an agent who believes in this sense that she ought to choose option O and yet fails to experience even *pro tanto* motivation to choose O clearly manifests incoherence in this sense (i.e., manifests akrasia). To say that a fully rational agent in this situation would feel at least some motivation to choose O is just to say that an agent in this situation *has some subjective reason* to choose option O, which is what WSR asserts.

But whether we should accept *any* formulation of the enkratic principle depends on whether we should accept the broader enkratic conception of rationality. I attempted to show in the previous chapter that there is a basic divide between theories of practical rationality that elaborate the enkratic, normative-belief-based conception of coherence in the practical domain and those that elaborate the rival Humean, desire-based conception. If this is so, then why should we accept the enkratic conception over its Humean competitor?

This is of course too large a debate to settle in a few pages. So I will only gesture at what I take to be the most compelling reason to prefer the enkratic conception. (I will then close the section by arguing that even one who prefers the Humean conception of rationality should accept WSR or something like it.)

The well-trodden path of argument against Humean theories of practical rationality involves agents with bizarre motivational states, like Parfit's "Future Tuesday Indifference" (Parfit, 1984, pp. 123ff) or Quinn's imagined basic disposition toward turning on radios (Quinn, 1993, pp. 236ff). But among other standard rejoinders, the Humean may simply maintain that in such alien cases it is difficult for us to appreciate the reason-giving force that such an agent's desires would have for her. Let's instead consider the case I described in the previous chapter as a test case for the Humean and enkratic conceptions: the would-be religious ascetic—let's now call him Simon—who believes that one course of action (fasting and repentance in the desert) would be most objectively good or choiceworthy while believing that another (endless debauched partying) would best satisfy his desires. The crucial question is, of these two choices, which would manifest a lack of internal coherence on Simon's part, given the motivational and belief states we have ascribed to him.

It seems intuitively clear to me that the choice to fast in the desert *coheres better* with Simon's overall internal state that the choice to revel in the city. One obvious way in which this is so is that Simon will feel that he can *explain* or *justify* to himself the choice to fast, despite the sacrifices it entails, in a way that he could not explain or justify to himself the choice to revel. Whichever choice he makes, he may feel some sense of ongoing internal struggle after the choice has been made, and even a species of regret. But if he chooses to revel, he will have reason to believe *that he made an all-things-considered mistake*, *that he was weak-willed*, *that he cannot justify or defend what he is doing*, whereas if he chooses to fast he will have no reason to think any of these things (assuming his beliefs about the objective choiceworthiness of fasting remain unchanged).

Of course, there is still the following puzzle: If Simon does in fact successfully choose to fast in the desert, doesn't this very choice show that fasting-in-the-desert was in fact what he had most desire to do, and therefore that, despite appearance, he was acting rationally on the Human conception as well as the enkratic conception of rationality? Three remarks on this puzzle: (i) A point of divergence among Humean theories of rationality is whether rationality concerns only the satisfaction of present desires/motivational states or of anticipated future desires/motivational states as well. If one adopts the latter view, then there is no puzzle, since an agent may easily act against the sum of his present and anticipated future desires. (ii) In the last chapter, I assumed that the Humean accepts a picture of *subjective* reasons and hence of *rationality* on which these things are determined not by what will best satisfy an agent's desires but by his *beliefs* about what will best satisfy those desires. In seems inevitable that the Humean must say something like this, since in the absence of perfect knowledge about what will satisfy our desires we can only be guided by our beliefs. If this is right, and if we keep the objective/subjective distinction clearly in view, then there is no puzzle: Objectively, of course, Simon might choose to fast even though reveling *will in fact* best satisfy his desires—suppose he simply underestimates the degree to which his desire to avoid physical suffering will be dissatisfied by choosing to fast. And subjectively it is equally clear that Simon might choose an option that he *believes* will not best satisfy his desires—he might, for instance, regard desire as inherently sinful and hold that in choosing to fast in the desert he is actuated by some entirely different source of motivation, like divine grace (or a Kantian sense of duty). (iii) The Humean may still, if they like, insist that their view does not entail drunken revelry to be the more rational choice for Simon. They might say that, in any event, he must believe when he chooses to fast that this is the choice *most consistent with his overall motivational state at the moment*, even if he believes that motivational state to include things other than *desires*, and that this is enough to make his choice rationally permissible. Or they might give their conception of subjective rationality a *de re* twist and say that, even if Simon does not believe at the moment of choice that fasting in the desert will best satisfy his desires, nevertheless the fact that he chooses it means that he believes it will best satisfy the *objectives* that he *in fact* most strongly desires, even though he would deny those objectives to be objects of desire. But either of these views has the obvious consequence of making the Humean theory trivial, implying that any or virtually any option that an agent in fact chooses must *ipso facto* have been rationally permissible for that agent.

To sum up, then: If the Humean view is to avoid triviality, it must claim that Simon is rationally required to revel in the city rather than fast in the desert. Therefore if one judges, as I do, that fasting is in fact the *more* rational option for Simon given his various beliefs and motivations, one must reject the Humean conception of rationality in favor of the enkratic conception (and therefore accept some version of the enkratic principle, which will entail WSR).[10]

But in any case, it seems to me that even Humeans about rationality should accept WSR or something very much like it, and hence should accept the relevance of moral belief to rational requirement. The tightness of the connection between

[10] Niko Kolodny (2005) offers a different line of argument for what I have been calling the enkratic conception of rationality. He argues at length that we do not have (objective) reasons to comply with rational requirements as such—that is, when in light of my beliefs and/or evidence I am rationally required to choose option O, although *in fact* there is no objective reason to choose O (e.g. because, contrary to my beliefs and evidence, O will in fact produce exclusively bad consequences), the mere fact that rationality requires it of me is not an objective reason for me to choose O. But, he further claims, rationality is *apparently* normative, and he proposes to explain this apparent normativity by what he calls the "Transparency Account": Rational requirements are apparently normative because, whenever one is subject to a rational requirement to choose option O, it is *in virtue of the fact that* one believes oneself to have reasons of some more ordinary kind (e.g., based on consequences) for choosing O that one is so required. Hence, from a first-personal standpoint, rational requirements always line up with (because they represent) the normativity of objective reasons. But, the argument continues, this Transparency Account only succeeds if there is an appropriate conceptual connection between rational requirements and one's beliefs about objective reasons. Kolodny proposes, to this end, that all rational requirements can be derived from what he calls the Core Requirements. In his preferred idiom of forming attitudes rather than choosing options, these requirements go as follows.

C+ If one believes that one has conclusive reason to have [attitude] A, then one is rationally required to have A.

C− If one believes that one lacks sufficient reason to have [attitude] A, then one is rationally required not to have A. (Kolodny, 2005, p. 557)

C+ and C−, of course, closely resemble standard articulations of the enkratic principle, and like the articulations I have offered above, C+ implies that Simon is rationally required to choose/intend to fast in the desert, while C− implies that he is rationally prohibiting from choosing/intending to revel in the city.

Humeanism and WSR depends on whether one accepts motivational internalism or externalism about moral judgments. On the internalist view, an agent who sincerely believes (or judges, on non-cognitive accounts) that she is morally required to choose option O must possess at least some *pro tanto* motivation toward choosing O. For the Humean, such motivation is at least *prima facie* reason-giving. Therefore, agents who believe that they are obligated to choose O will generally have some objective reason to choose O, and provided they recognize their motivation toward choosing O, will have some subjective reason as well, just as WSR asserts.

The motivational externalist, on the other hand, holds that it is possible in principle for an agent to judge that she is morally required to choose O and yet to experience no motivation of any kind toward choosing O. Nevertheless, even externalists accept that *as a matter of fact*, nearly all our moral judgments are accompanied by at least some degree of motivation, and therefore the Humean motivational externalist will still accept that in nearly all cases, agents who believe that they are morally required to choose an option O have at least some reason to choose O. Thus, while the proponent of the enkratic conception of rationality should (as I have argued) accept WSR outright, the Humean should accept WSR at least as a defeasible generalization holding true in most cases (defeasible in that an agent *may* fail to exhibit any motivation corresponding to her full moral belief that she is required to choose option O and/or in that this motivation *may* fail to generate subjective reasons if the agent does not recognize it in herself or if other, deeper features of her motivational state generate undercutting defeaters to the *prima facie* reason for choosing O).[11]

Summarizing the argument of this section: WSR asserts that an agent who *fully* believes, with conclusive justification, that she is objectively morally required to choose option O has thereby some subjective reason to choose O. To accept WSR is to accept (*contra* critics like Harman, Weatherson, and Hedden) that an agent's moral beliefs affect the rational requirements to which she is subject. There is a strong *prima facie* case for accepting WSR and therefore the rational relevance of moral belief: (i) WSR is plausibly entailed by the enkratic conception of rationality, which I have argued is more plausible than the rival Human conception; and (ii) even one who accepts the Humean conception of rationality should accept WSR at least as a defeasible generalization, and therefore should accept that moral belief is *generally* reason-giving and hence relevant to rational requirements.

2.3 The *Prima Facie* Case for SSR: Parallels between Moral and Empirical Uncertainty

The case for SSR might also be made in terms of rationality-as-coherence: If you believe that option O *might* be more objectively choiceworthy than any alternative but experience no corresponding *pro tanto* motivation to choose O, this is arguably also an instance of irrational incoherence. Likewise, if your degree of belief that O is objectively most choiceworthy increases but your motivation to choose

[11] An analogous argument could be made that the Humean should accept SSR, if one believes that *partial* as well as full moral belief is necessarily or generally motivational. These claims, however, are more controversial than the corresponding claims with respect to full moral belief.

O does not correspondingly increase, this might be seen as a form of diachronic irrationality.

But rather than pursuing this line of thought, I will rely instead on an argument commonly offered in support of moral hedging, that also supports the somewhat weaker SSR. This argument takes the form of a challenge: If, as nearly everyone agrees, *empirical* risks and uncertainties are relevant to what agents subjectively ought to do, then why not also *moral* risks and uncertainties? What is the relevant difference between uncertainty about the empirical state of the world and uncertainty about the basic principles of morality, such that the former but not the latter are relevant to the requirements of practical rationality?

In the next chapter, we will examine some arguments that implicitly or explicitly endeavor to meet this challenge. But first, let's try to cast the argument in the following more positive light: Whatever arguments can be given that rational requirements should be sensitive to an agent's empirical uncertainties can be closely paralleled by arguments that rational requirements should likewise be sensitive to her moral uncertainties.

The case for the rational relevance of empirical uncertainty is most commonly made by appeal to intuition. Two famous and structurally analogous thought experiments, Frank Jackson's drug case and Derek Parfit's mineshaft case, serve as highly effective pumps for this intuition. In both cases, an agent has three options, O, P, and Q. She knows that, of O and P, one is the best of her options and the other is catastrophic, but she has no way of knowing which is which and assigns equal likelihood to either possibility. Q, meanwhile, she knows for certain to be nearly as good as the best option (either O or P) and much better than the worst. In Jackson's case, O, P, and Q are drugs that a doctor may prescribe to her patient. Of O and P, one is a complete cure and the other a deadly poison, while Q is known to be a partial cure that leaves the patient with only a few minor symptoms (Jackson, 1991, pp. 462f). In Parfit's case, a mine is flooding and one hundred miners are trapped in one of two shafts. The agent may either (O) block the lefthand shaft, (P) block the righthand shaft, or (Q) block neither shaft. If she blocks the shaft in which the miners are trapped, the flood waters will not enter that shaft and the miners will be saved, but if she blocks the other shaft, all the water will be directed into the shaft containing the miners, all of whom will be killed. If she takes option Q, blocking neither shaft, the flooding will be moderate and only ten miners—at the bottom of the occupied shaft—will be killed (Parfit, 2011, p. 159). In cases of this sort, often collectively referred to as "Jackson cases," it seems obvious to most people that the agent is rationally required to choose option Q, as implied by expected utility theory.

The equivalent of irrelevantism with respect to empirical beliefs is *objectivism* about rational requirements: the view that an agent is always required to perform the action that is objectively best, e.g., the action that has the best consequences. But this is incompatible with the judgment that, in Jackson cases, agents are rationally required to choose option Q, since Q by stipulation does *not* have the best consequences (and in fact, is the one option that the agent knows for certain will not have the best consequences). Even if O, for instance, is the optimific action (in Jackson's original case, the drug that will completely cure the patient's condition), the agent's uncertainty about the consequences of O should lead her to choose Q instead.[12]

The moral of Jackson cases is that agents are not always rationally required to choose the action that is objectively best, or even an action that has maximal probability of being objectively best. Rather, rational requirements depend crucially on an agent's uncertainties and the balance of risk, at least with respect to the empirical facts. If anyone is inclined to resist this conclusion, the intuition may be driven home as follows: (1) Suppose you find yourself in the position of Jackson's doctor, and must prescribe your patient one of the drugs O, P, or Q. Which drug *would you* in fact choose to prescribe? (2) Assuming your answer is drug Q, do you view this choice as a mere idiosyncratic preference? An arational foible? Something akin to an involuntary twitch? Or was your choice guided by reasons? It is very hard to resist the conclusion that, in practice, we choose options like Q over options like O and P without any hesitation, and that in so doing we view these choices as rationally prescribed or required. This seems to commit us to the view that rational requirements are sensitive to degrees of belief. Call this two-step argument the Practice Test. In a moment, I will try to show that the same test can be applied, with nearly the same force, to cases of moral uncertainty in service of the analogous conclusion.

But first, let's note a more abstract argument that might be given in favor of

[12] Even philosophers who defend an objective understanding of *moral* requirements typically do not defend objectivism about *rational* requirements in the above sense. Rather, they attempt to explain how the practical conclusion that a rational agent should choose option Q is compatible with the claim that she *ought to* or *is morally required to* choose whichever of O or P is in fact optimific. For one example of this strategy, see Oddie and Menzies (1992).

choosing option Q in Jackson cases: namely, an appeal to Kant's famous principle that “‘ought’ implies ‘can’.” To one who asserts that an agent subjectively ought (= is rationally required) to choose whichever option of O and P in fact has the best consequences, we are inclined to point out that the agent in the Jackson case *cannot comply* with this prescription, because she cannot reliably identify the optimific option. Of course, there is a sense in which the agent *can* choose the optimific option: namely, she *can* choose O and she *can* choose P, and one or the other of those options is optimific. But the true force of “‘ought’ implies ‘can”’ is clearly stronger than this. It is not enough that an agent might manage to comply with a prescription by luck or accident. For that prescription to represent a genuine rational requirement, she must be able to comply with it *rationally*, that is to say, *reliably* and *on the basis of reasons.* Her ability to do what she ought to do must not be contingent on a lucky guess. Principles of rational requirement that are sensitive to an agent's degrees of belief in normative or evaluative propositions can satisfy this constraint; norms that are sensitive only to facts about the world to which the agent may lack epistemic access (e.g., which drug is in fact the perfect cure) cannot.

Now let's consider an analogue of Jackson's case, where empirical uncertainty is replaced with moral uncertainty, and see what the “‘ought’ implies ‘can’” principle tells us about this case. Suppose you have just been elected World Dictator for Life, and must choose a path for your loyal subjects to follow. Your moral beliefs are evenly divided between hedonistic and perfectionist forms of average consequentialism. You have a group of learned advisers who have crafted three possible social schemes for you to consider, and have run (highly reliable) projections of the average hedonic and perfectionist welfare levels that would result from each scheme.

- On the *Spartan* scheme, citizens will live in communal barracks, subsist on a high-protein bean paste, and develop their moral and intellectual virtues by 18-hour days of work, reading, and exercise. This plan carries an expected value of 100 perfectons per person, but zero net hedons (hedons minus dolors).

- On the *transhumanist* scheme, citizens will live inside machines that simulate pleasant experience while regulating their hormone levels to maintain a hedonic optimum. (The machines will be serviced in perpetuity by an army of highly dependable automata.) This plan carries an expected value of 100 net hedons per person, but zero perfectons.

- Finally, on the *eudaemonist* scheme, citizens will live as participants in a free society with a strong social safety net. They will be taught the value of work, education, and exercise, and also given access to the same prohedonic drugs that would be used in the experience machines. Because some in this society will eschew work etc, and others will eschew drugs, the eudaemonist scheme cannot generate hedonic or perfectionist optima. But it still carries an expected value of 95 perfectons and 95 net hedons per person.[13]

[13] For a somewhat more down-to-earth moral-uncertainty-based Jackson case, see MacAskill's "Susan and the Medicine II" (MacAskill, 2014, pp. 9-11).

Nota bene that the assignment of cardinal values does not imply commensurability between the hedonistic and perfectionist theories. We could, for instance, change the size of a perfecton so that the three schemes generate 1,000,000/0/950,000 perfectons per person respectively, without changing the substance of the thought experiment. But for the thought experiment as it stands, we should assume that the size of a perfecton and a hedon are such that five perfectons and five hedons both represent (what we would intuitively describe as) a minor but not trivial change in someone's welfare.

Given your distribution of moral belief, you are certain that either the Spartan or the transhumanist scheme is objectively best. Nevertheless, it seems intuitively rational to choose the eudaemonist scheme. And it seems likely that this is the scheme that nearly anybody *would* choose, if they had the imagined distribution of credence and accepted their advisers' projections. As with Jackson's original case, we can apply the Practice Test: If you *would* select the eudaemonist scheme, given your stipulated doxastic state, and if you view this choice as non-arbitrary and guided by reasons, then you are committed to the conclusion that degrees of moral belief are relevant to rational requirement (that is, your are committed to SSR). Though you lack a full belief that the eudaemonist scheme is better than, say, the transhumanist scheme, your partial belief (in the moral theory that implies this to be the case) is, under the circumstances, sufficient to generate a rational requirement, regardless of which moral theory ultimately turns out to be true.

Likewise, as with Jackson's case, we can run an "'ought' implies 'can'" argument: To say that you ought to choose whichever option is objectively best, as the irrelevantist would have us say, is to give a prescription with which you have no capacity to (non-accidentally) comply. Just as in Jackson's case you cannot discover which of drugs O and P is the cure and which the poison, so in our moral analogue you cannot discover which of the Spartan and transhumanist schemes would be morally ideal and which a moral disaster. So a rational requirement to choose the scheme that is in fact morally ideal would be a rational requirement with which reason alone does not enable you to comply.

"'Ought' implies 'can'" furnishes a negative argument against objectivist/irrelevantist views of rational requirement. But one might also try to offer a positive explanation for the rationality of risk aversion in the empirical domain, and here too, the most natural explanations seem to transfer seamlessly from empirical risks (e.g. uncertainty about the consequences of your actions) to moral risks (the risk that the true moral theory is one that implies your action to be substantially morally wrong). Consider, for instance, the following passage from Oddie and Menzies (1992), in which they attempt to explain how an objective-utilitarian theory of value can justify subjective expected-utility maximization as a "selection procedure" for actions.

> The regulative ideal for the moral agent is to maximize objective value. But we have already noted two features of the ideal. First, it is an ideal which admits of degrees of realization, and there is an obvious and natural ordering of options with respect to the realization of the ideal...Failing to achieve the ideal admits of degrees, and it is objectively preferable for the agent to perform an act with more objective value rather than less in those choices which do so fail. Second, the agent is not usually in the position of knowing which option fully realizes the ideal. If we put these two features together, then it is clear that in attempting to realize the ideal the agent is not justified in selecting the option which, by her lights, has the greatest probability of fully realizing the ideal, since such an option may also have a significant probability of missing out by a large margin. Rather, the agent is justified in selecting the option which overall has the best prospects for objective value enhancement.

(Oddie and Menzies, 1992, pp. 530-1)

Reasoning along these lines plausibly explains our ordinary, obvious responses to empirical risk: for instance, why I should not drink (or serve) coffee that I believe with subjective probability .1 might contain a deadly poison, and why the doctor in Jackson's case should not prescribe either Drug O or Drug P. And this reasoning requires little if any amendment to act as an argument for aversion to moral risk: I know that the correct regulative ideal for the moral agent is the true moral theory, or the features of the world to which that theory attributes moral significance. But I don't know what features those are (hedons? friendships? the good will?), insofar as I am uncertain what moral theory is correct. Rather, the target I know I must aim at is moral value as such, and in aiming at this ideal I will naturally be led to account for all those features of the world that *might* have moral significance. In other words, given that I recognize objective moral value as the goal towards which I should act, I can do no better than to minimize the degree to which I deviate from or fall short of this standard in expectation, and this requires that I be sensitive in my practical deliberations to moral uncertainty and moral risk.

In summary, three of the most natural arguments for the claim that rational requirements are sensitive to an agent's empirical uncertainties and degrees of belief also suggest that rational requirements are sensitive to an agent's *moral* uncertainties and degrees of belief. We can (i) construct intuition-pumping thought experiments analogous to Jackson's drug case and Parfit's mineshaft case (augmenting the intuition-pumping force of these thought experiments by means of the Practice

Test), (ii) appeal to the same negative "'ought' implies 'can'" argument to explain the intuitively falsity of the objectivist/irrelevantist response to these dilemmas, and (iii) offer the same sort of positive explanation for the rationality of attending to moral risk that we would naturally give with respect to empirical risk.[14]

There is, then, at least a *prima facie* analogy between empirical and moral uncertainty. We need norms of rational choice that are sensitive to an agent's morally relevant empirical uncertainties, and there is no obvious reason why these same norms should not extend to at least some cases of "pure," non-empirical moral uncertainty. Thus we have at least *prima facie* reason to accept SSR, which asserts that rational requirements are sensitive to an agent's moral degrees of belief. Of course, it is still only a *prima facie* case, and in the next chapter we will examine arguments that purport to defeat the analogy between these two kinds of uncertainty and to show a fundamental distinction between the deliberative roles of empirical and moral belief.

[14] It might seem that the arguments from Jackson cases and from "expected shortfall" principles go further than justifying SSR and act as arguments for moral hedging or even a full-blown expected value theory of rationality under moral uncertainty (à la "maximize expected choiceworthiness"). But one can accept the analogy between empirical and moral uncertainty without being driven to any conclusion of a consequentialist or expected-value-maximizing flavor. In Chapter 5, for instance, we will examine a case in which a non-expectational "threshold principle" for deciding what to do when one is empirically uncertain whether a given course of action would violates a deontological moral constraint seems to generalize naturally from empirical to moral uncertainty. Thus the analogy between the two types of uncertainty can support various extensions of SSR, not all of which will be forms of hedging.

Chapter 3: Challenges to the Rational Relevance of Moral Belief

There is *prima facie* reason to believe that an agent's moral beliefs and degrees of belief are relevant to the requirements of practical rationality. But these reasons are not indefeasible. In this chapter we will examine some of the arguments that have been recently offered for the *irrelevance* of an agent's moral belief states to what she subjectively ought to do or what rationality requires of her.

3.1 Normative Externalism

I begin by setting aside a debate that might seem—but, I claim, is not—crucial to the defense of WSR and SSR. Weatherson frames his arguments for the irrelevance of moral uncertainty as part of a broader defense of *normative externalism*, the view that "[w]hat one should do, or should believe, in a particular situation is independent of what one thinks one should do or believe, and (in some key respects) of what one's evidence suggests one should do or believe" (Weatherson, 2014, p. 141). The externalist, as I will understand the position, claims that some norms of (theoretical and/or practical) rationality are incumbent on an agent, have rational requiring force with respect to her, whether or not she believes those norms are incumbent on her or has any reason to believe that they are incumbent on her.

As we will see in Chapter 7, there is a compelling (and, I think, ultimately successful) argument for normative externalism based on the threat of infinite regress that the rival internalist position faces. The internalist claims that the norms that apply to an agent are the norms that she accepts (or ought to accept, on the basis of her evidence), and that these norms apply to her, roughly, *to the extent that* she accepts them. Thus, when an agent is uncertain what "first-order" norm objectively governs a particular choice, the various norms across which she distributes her belief each contribute some subjective reasons for action, and (it seems) she must then go searching for some "second-order" norm to tell her how to compare these reasons. But what this second-order norm should be is itself an uncertain matter, so she will no doubt end up distributing her belief over various, sometimes conflicting second-order norms. When these norms do conflict she must, by the internalist's lights, go searching for some *third*-order norm that tells her how to deal with her uncertainty about the second-order norms, and so on *ad infinitum*. Thus, it is claimed, the internalist is trapped in a regress that paralyzes rational choice.

But even if, in order to block this regress, we must concede that *some* normative principle has its force independent of an agent's beliefs or evidence, this does not force us to accept the irrelevantist's claim that *first-order moral norms* have this sort of belief-independent normative force. Rather, as I will argue in Chapter 7, the most plausible way of halting the regress described in the last paragraph is to accept a basic principle of *rationality*—namely, the enkratic principle—as a constitutive or definitional condition of rational agency and therefore as incumbent on agents regardless of their beliefs or evidence. And since, as I argued in the last chapter, the enkratic principle implies that what an agent subjectively ought to do depends *at least* on her full moral beliefs, and plausibly on her partial beliefs as well, this externalist conclusion is actively incompatible with irrelevantism. The irrelevantist, therefore, cannot prove her case merely by arguing for normative externalism in general, but must offer some separate argument to establish that *first-order moral norms* have belief-independent normative force. More particularly, if she accepts the rational relevance of *empirical* uncertainty, she must offer some argument for a dividing line between "basic" moral norms (that apply to an agent regardless of her empirical circumstances) and "derivative" moral norms (that apply to an agent in virtue of the basic moral norms plus her empirical circumstances), such that the former but not the latter have belief-independent authority. The regress argument alone cannot discharge this burden of proof, but the arguments we consider in the rest of this chapter attempt to do so.

3.2 Weatherson on Moral Fetishism

Weatherson (2014) states his states his position on moral uncertainty as follows.

> Being uncertain about the physical consequences of your actions should matter both to what you do, and how you are assessed. The red light runner is immoral, even if she never actually harms anyone, because she endangers morally significant humans. But the meat eater [who believes but is not certain that meat eating is morally permissible] cannot be

> condemned on the same grounds. If she is wrong that meat eating is morally acceptable, that would be one thing. But a mere probability that meat eating is immoral should not change one's actions, or one's evaluations of meat eaters. (142-3)

I take this to be a denial of SSR and the rational relevance of moral uncertainty.[1]

Weatherson offers two arguments for this view, which deserve separate treatment. The first claims that an agent who is sensitive to moral risks must exhibit a defective motivational state that Weatherson, following Michael Smith, calls "moral fetishism" (Smith, 1994). Though Smith introduces the idea of fetishism as part of a very different argument, aimed against motivational externalism about moral judgments, we will shortly see that the details of his argument bear a surprising relevance to Weatherson's. It is therefore worth a brief digression to set out Smith's argument. Smith argues that the motivational externalist—who holds that the sincere moral judgment, e.g., that option O is morally required need not be accompanied by any motivation to choose O—must explain why morally good agents nevertheless *are* motivated to act in accordance with their moral judgments. This means that the externalist must claim that a morally good agent is possessed of desires that lead

[1] Beyond this, I am somewhat uncertain how Weatherson's view should be classified. At points in Weatherson (2014), he seems to suggest a full-on denial even of WSR, for example when he says: "I think the stronger, prima facie implausible, view is true: rightness and wrongness as such shouldn't even be part of our motivation" (p. 160). But he also concedes that an agent who acts against her full moral beliefs can be convicted of "hypocrisy," and seems to think that agents have reason to avoid hypocrisy, which would amount to an acceptance of WSR (pp. 157-8). He also regards it as at least acceptable for agents to be motivated by moral risks understood in terms of "thick" moral properties, e.g., the risk that eating meat you are personally unwilling to slaughter exhibits a vice of cowardice. If any case of moral uncertainty can be recast in such "thick" terms, then Weatherson has for all intents and purposes accepted SSR and should then accept the need for norms that guide agents' choices under moral uncertainty—with the thought, perhaps, that casting those norms in terms of thick moral properties will have substantial and interesting effects on their content. If this is in fact his view, then so much the better.

her to act morally. According to Smith, these desires could be of only two kinds: (i) "*de re*" motivations to take actions with the sorts of non-moral properties that the agent judges to subvene on moral rightness or (ii) "*de dicto*" motivations to perform morally right actions as such, i.e., to perform *whichever actions turn out to be morally right*. An agent whose moral judgments are utilitarian, for example, exhibits *de re* moral motivation if she desires to perform actions that promote the general happiness *independent* of her judgment that these actions are morally right, while she exhibits *de dicto* moral motivation if she desires to perform actions that promote the general happiness precisely *because* she judges that those actions are morally right.

Smith's argument against externalism then takes the form of a dilemma: *De re* moral motivation cannot be the mark of a morally good agent because it fails to track *changes* in moral judgments. For example, if our morally good utilitarian finally gets around to reading the *Groundwork* and is persuaded to abandon her utilitarian convictions in favor of Kantianism, then as a morally good agent she should *become* motivated to tell the truth and keep her promises even when doing so will not promote the general happiness. But her *de re* motivations will not support this shift, since they only impel her to promote the general happiness and are not conditioned on her judgments about moral rightness. *De re* moral motivations do not follow an agent's moral judgments and hence, Smith claims, cannot be the defining feature of a morally good agent.

If the morally good agent, as the externalist conceives her, cannot be characterized by *de re* moral motivations, then she must instead be morally motivated *de dicto*, i.e., motivated to perform right actions as such, because they are right. But this, Smith claims, is not a feature of morally good agents at all, but rather a vice, which he calls "moral fetishism." Here Smith draws on Bernard Williams' famous "one thought too many" objection to Kantianism and utilitarianism (Williams, 1981). Williams describes a man who must leap into a river to save his drowning wife rather than another, equally drowning stranger. Theories like Kantianism and utilitarianism that make one's personal attachments subordinate and accountable to the universalistic and impartial demands of the moral law, Williams thinks, attribute "one thought too many" to a morally good agent in this circumstance: Rather than thinking "It's my wife, so I must save her," he instead thinks "It's my wife, and the maxim of saving one's wife instead of a stranger can be consistently universalized" or "It's my wife, and in this circumstance saving her would be at least no worse for the general happiness than saving the stranger." It seems to Williams that such a motivational scheme, that places the impersonal demands of morality ahead of personal commitments and values, is a character defect and not a virtue. Likewise, it seems to Smith and Weatherson that an agent who is motivated to save her spouse, or help her friends, or keep her promises, or donate to GiveDirectly merely on the abstract and impersonal grounds that these things are morally right or good has "made a fetish out of morality."

Weatherson adapts Smith's fetishism argument against the externalist's *de dicto* model of moral motivation into an argument for the permissibility of what he calls "moral recklessness." To be morally reckless is to ignore moral risks, e.g., to eat meat despite one's suspicions that eating meat might be morally wrong. Weatherson argues that an agent who is motivated to avoid moral recklessness must be guilty of moral fetishism—for instance, she avoids eating meat *because it might be wrong.* If an agent who avoids moral recklessness must exhibit moral fetishism, and moral fetishism is a vice, then Weatherson thinks it must follow that there is nothing wrong with moral recklessness. This, then, is an argument against SSR: A moral agent whose practical deliberations are sensitive to any positive degree of moral belief, for instance that a certain action might be morally required or prohibited, is thereby criticizable and so, presumably, cannot merely be acting as rationality requires.

There are a lot of things wrong with this argument. First, I confess that I find Williams', Smith's, and Weatherson's case intuitions paper thin and badly misleading. To note the obvious, an agent can be responsive to "*de dicto*" moral *reasons* of the sort postulated by WSR and SSR without needing to think, each time she acts for those reasons, "...and I'm doing this because it is (or might be) morally required of me." Just as an airline pilot performing preflight safety checks can be motivated by concern for the safety of her passengers without thinking, at each step, a sentence like "By checking the deicing fluid levels, I help to keep my passengers safe," so an agent who acts to promote the general happiness because she believes it to be her moral obligation need not stop to think to herself, as she rushes into Singer's pond to save the drowning child, "The Axiom of Universal Benevolence demands it!" The "one thought too many" intuition that underlies the fetishism objection, then, is misleading in part because it is based on a misrepresentation of what it means for agents to be reasons-responsive and therefore what the defender of *de dicto* moral motivation is committed to.

For my part, I don't think that an agent who *did* reason through her moral decisions in this hyper-explicit way would thereby *necessarily* be guilty of anything worse than a little cognitive inefficiency. But contingently, human nature being what it is, we would be rightly put off by anyone who always had to think, explicitly, "...and doing such things is morally right" to summon up any motivation for right-doing. Our moral psychology does not easily beget characters like Kant's sorrowing philanthropist. Rather, it seems almost impossible for a human being consistently to act well without possessing the kinds of automatic sympathetic responses to the needs of others that obviate the need for conscious reflection on the demands of the moral law.[2]

Notably, also, it is easy to describe forms of *de dicto* motivation that do not intuitively feel fetishistic or otherwise objectionable. Someone who, for example, returns an accidentally over-generous tax refund to IRS "because it was the right thing to do" seems to express laudable, not objectionable, motivations. (Indeed, Weatherson admits that he must bite a bullet here and claim that people who say such things, who feel themselves to be motivated by "conscience" or "a sense of right and wrong," are mistaken about their own motivations (Weatherson, 2014, p.

[2] I suspect, speculatively, that the underlying "one thought too many" intuition gets its force in part from a superficial resemblance to a different, more persuasive argument, namely the "clean hands" objection to deontology. When the deontologist says that she will not commit one murder to prevent five, the objection goes, she places a concern for her own moral purity above a concern for the rights or interests of others. She is more concerned with earning morality's stamp of approval than with the *substance* of morality, viz., those rights and interests. Likewise, when we consider the sorrowing philanthropist or the man who pauses to think of the moral law before jumping in the river to save his wife, we are apt to worry that their concern with doing the right thing has become dangerous unmoored from the underlying substance of morality. And, in human beings as we are, this may often be the case. But this is not a *necessary* pitfall of *de dicto* moral motivation: It is perfectly possible, in principle, for an agent to be motivated *de dicto* while keeping clear sight of the fact that doing the right thing requires attending to the interests of others, and not mere mechanical adherence to a set of precepts.

161).)

But if one does find oneself compelled by the "one thought too many" intuition and the claim that pure *de dicto* motivation is a vice, there is more to be said. Sepielli (2016) offers three further objections to Weatherson's fetishism argument. Each of Sepielli's objections strikes me as compelling, but I will confine my attention to the first, namely, that consciously avoiding moral recklessness does not require an agent to exhibit *de dicto* moral motivation.[3] An agent might, Sepielli suggests, avoid ordering steak at a restaurant because she is motivated *to treat the cow with proper respect*, and she is morally uncertain whether eating the cow's meat is consistent with this goal. Here the agent exhibits something in between *de re* and *de dicto* motivation, since her desires track a "thick" moral property, namely, *treating beings with proper respect*. This sort of motivation seems to trigger the "fetishism" intuition much less strongly than do purely *de dicto* motives, as Weatherson himself concedes (Weatherson, 2014, p. 159).

Alternatively, Sepielli claims, an agent who avoids recklessly eating meat might be motivated by purely *de re* concerns, like concern for the cow's suffering, despite her uncertainty whether the cow's suffering has moral significance. Here, I think, lies the most compelling response to Weatherson's fetishism argument, though it will take some effort yet to suss out. Note, to begin with, that in availing himself of the second horn of Smith's dilemma for externalists (the fetishism objection to

[3] Sepielli's other objections are, in brief, that (ii) *de dicto* motivation in cases of moral uncertainty is compatible with exhibiting desirable *de re* motivations in other sorts of cases, and so even if *de dicto* motives were needed to avoid moral recklessness, these motives would not be objectionable; and (iii) even if acting on norms that are sensitive to an agent's moral uncertainties did require that agent to exhibit objectionable motives, that fact would not constitute a refutation of the norms themselves.

de dicto moral motivation), Weatherson impales himself pretty badly on the first horn. Smith takes it for granted that a morally good agent must have motivations that track her sincerely held moral judgments, and this assumption furnishes his argument against the *de re* externalist model of moral motivation. But Weatherson's use of the fetishism argument commits him to denying this tracking claim: a morally good agent, he claims, will be motivated *de re*, not by the features of the world that she judges to have moral significance, but by the features of the world that *in fact* have moral significance—and these, of course, do not change with an agent's moral beliefs.

This is a substantial intuitive cost to Weatherson's view, and moreover it as an unnecessary cost. If one wishes to avoid attributing *de dicto* motivations to morally good agents, one would do better to adopt Smith's own internalist view, according to which sincere moral judgments are internally motivating, except in agents who are suffering from akrasia or other rational failings (Smith, 1994, pp. 61f). Such a view clearly vindicates WSR, and contradicts Weatherson's view that an agent's moral beliefs have no bearing on what she subjectively ought to do. A full belief that option O is morally required constitutes, presumably, a *judgment* that one is required to choose O, and thus on a Smith-ian motivational-internalist view must give an agent at least *pro tanto* motivation to choose O, on pain of practical irrationality.

Moreover, it seems plausible that this sort of motivational-internalist view should vindicate SSR as well. Our ordinary moral judgments are beliefs to which we assign, presumably, high probability but something less than certainty. For instance, my judgment that I am morally required to rescue drowning children from shallow ponds is about as strong as any moral judgment I hold, but it falls short of certainty since I assign some credence, for instance, to moral error theories according to which all statements of moral requirement are false. But if my (say) .9 credence that I am morally required to save drowning children from ponds must be motivating in me, on pain of practical irrationality, why should the same requirement not apply to my .6 credence that I am required to avoiding harming distant future generations, or my .4 credence that it is morally good, *ceteris paribus*, to bring additional happy people into existence, or my .2 credence that it is good to preserve natural beauty even when no one will benefit from the aesthetic experience of it? An internalist view like Smith's will end up endorsing SSR (i.e., that agents should be responsive to their partial moral beliefs), unless it attributes undue moral significance to some "Lockean threshold" for full moral belief/judgment.[4] And given that this internalist view escapes the fetishism objection and also capture the intuitive claim that a morally good agent's motivations should track her moral judgments, it is unclear why Weatherson should not prefer this to his own *de re*-externalist view.

But moreover—and here we pick up once again the main thread of Sepielli's objection—there are externalist models of moral motivation that avoid both horns of Smith's argument for internalism, and that explain how an agent might avoid moral

[4] And in fact, as Sepielli points out, Smith himself is fairly explicit in accepting this implication: "In his (2002), Smith tells us that for a rational agent, the motivatinoal force of a moral judgment is a function of two features—'importance' (i.e. the moral significance the judgment assigns to an action), and 'certitude' (i.e. the degree of confidence the judgment manifests). (pp. 309-310) Nor does he regard motivation arising from uncertain judgments as any more fetishistic than motivation arising from certain ones. (p. 213) Smith would agree, then, that an agent may be ultimately motivated by things she thinks might be morally significant, even if...she is not sure that they are." (Sepielli, 2016, pp. 2957-8)

recklessness on the basis of purely *de re* motives. Jamie Dreier (2000) suggests a model on which a morally good agent is characterized by a second-order motivation to develop *de re* first-order moral motivations: that is to say, she desires to care about the things that have moral significance for their own sake, and so for instance, if she comes to believe that animal suffering has moral significance, will attempt to cultivate in herself a *de re* concern for animal suffering. As Dreier points out, this second-order motive is not simply *de dicto* motivation or moral fetishism by another name, since the *de re* first-order desires at which the morally good agent aims are not *conditioned on* her belief that the phenomena in question (e.g. animal suffering) have moral significance, nor does this belief need to play any "maintenance role" once the motivation is established. Rather, she comes to care about animal welfare for its own sake, just as Weatherson would have her do.

Like Smith's internalism, Dreier's second-order model escapes concerns about moral fetishism, explains the tracking relationship between moral judgment and motivation, and in so doing clearly vindicates WSR: An agent who fully judges that it is wrong to eat meat should develop (at least some) motivation against eating meat. And again like Smith's internalism, it seems natural that Dreier's model should vindicate SSR as well, for if a good agent has a second-order desire to cultivate *de re* concern for the morally significant features of the world, then her judgment that animal suffering *might* have moral significance will presumably motivate her, at least a little bit, to develop a *de re* concern for animal suffering. This, then, explains how a non-fetishistic agent might avoid moral recklessness: Her judgment that eating meat might be wrong leads her to cultivate at least a limited *de* *re* concern for animal interests that then imbues her with at least some motivation to avoid meat consumption.

Indeed, it seems to me that Weatherson's view itself requires one to adopt something like Dreier's second-order moral motive: For if you judge, as Weatherson does, that a morally good agent is one who is motivated *de re* by the features of the world that have moral significance, and you judge that some feature F has or might have moral significance (e.g., animal welfare interests), it seems incoherent (i.e., irrational) to experience no consequent motivation to develop in yourself a *de re* concern for feature F. An agent who judges that, to be a good person, she must have these *de re* concerns would *at least* desire and aim for *de re* concerns for those features of the world that she fully judges to have moral significance, and again, it is hard to see why this motivation should not extend (in proportionally diminished degree) to partial judgments as well.

3.3 Weatherson on Prudential Uncertainty

Weatherson has a second argument for irrelevantism, to which I have already alluded. Rather than analogizing moral uncertainty to empirical uncertainty, he claims, we should instead draw an analogy with uncertainty about *prudential* values and norms, in which domain it seems counterintuitive that agents should be motivated by partial belief. To pump this intuition, Weatherson offers the following case.

> Bob has thought a bit about philosophical views on welfare. In particu-

> lar, he has spent a lot of time arguing with a colleague who has the G.E. Moore-inspired view that all that matters to welfare is the appreciation of beauty, and personal love. Bob is pretty sure this isn't right, but he isn't certain, since he has a lot of respect for his colleague and for Moore.
>
> Bob also doesn't care much for visual arts...Now 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. Still, Bob decides it isn't worth the trouble, and stays out. (Weatherson, 2014, pp. 148-9)

Bob, then, is insensitive to what we might call a "prudential risk," namely, the risk that by not stopping at the art gallery he fails to satisfy an important welfare interest. But, Weatherson thinks—assuming that the pseudo-Moorean view is wrong and that sterile "appreciation" of great artworks will not in itself make Bob's life go better—Bob has not acted irrationally or criticizably. And if Bob's case is relevantly similar to the case of the agent with a low but positive credence that meat consumption is morally wrong, then this agent would likewise not be acting irrationally or criticizably by continuing to consume meat.

I confess once again that I don't share Weatherson's intuitions. Assuming Bob has a justified conviction that he ought to pursue all the constituents of a good life, and is justified in assigning non-marginal credence to the view that the appreciation of beauty is an intrinsic constituent of the good life, then it does not seem utterly unreasonable that he should hedge his bets and make the occasional unrewarding visit to the gallery.[5]

Moreover, Weatherson's intuition can be explained away, revealing in the process an important methodological principle for debates over normative uncertainty. Note that Weatherson has chosen, as his example of a view about prudential interests in which one might have low but positive credence, a view that nearly everyone will find extremely uncompelling. It is quite implausible, to most of us, that staring joylessly at great works of art makes an indispensable positive contribution to the goodness of a life. This first-order intuition significantly biases us against the second-order conclusion that Bob should hedge his bets. Moreover, the first-order intuition makes it hard to genuinely internalize the required stipulation of the thought experiment, that Bob is *justified* in assigning non-trivial credence to the pseudo-Moorean view. Where a thought experiment requires us to accept implausible stipulations, our intuitions will often fail to reflect those stipulations, and that may be what is happening in the case of Bob. The methodological principle, then, is that thought experiments intended to pump intuitions for a particular view about how to respond (or not to respond) to normative uncertainty should not involve first-order views about which we have intuitions that mimic, or illicitly reinforce, the intended thrust of the second-order intuition pump.[6]

Consider, by contrast to Weatherson's case, a case of more reasonable pruden-

[5] Intuitions are confused here, I think, by the fact that we don't typically view prudential concerns as having quite the same sort of binding rational force as do moral concerns. It is unclear what reason Bob has to pursue all the constituents of a "good life" if, fully informed and having deliberated carefully, he still finds that some of those constituents leave him cold.

[6] I should now confess to having violated this principle myself in the Jackson case analogue from §2.3. The reader may discount the force of that thought experiment accordingly.

tial uncertainty. Psychologists who study happiness commonly draw a distinction between two aspects of subjective well-being, *affect balance* and *life satisfaction* (Diener et al., 1999). Affect balance reflects the average of moment-to-moment affective tone, including physical and emotional pleasures and pains. Life satisfaction reflects an agent's *assessment* of her life as a whole. While affect balance and life satisfaction are, unsurprisingly, positively correlated, the correlation is far from perfect and measures of happiness that focus on affect balance can reveal importantly different results than those that focus on life satisfaction. For instance, in one widely discussed study, Kahneman and Deaton (2010) find that being wealthy rather than merely middle class improves people's sense of life satisfaction without improving the balance of positive and negative emotion they experience. On a more immediate scale, many psychological studies find evidence for a "peak-end rule" by which a person's after-the-fact assessment of the goodness or badness of an experience is predicted, not by the net balance of positive and negative affect over the course of that experience, but primarily by her affective state during the most intense moments of that experience, and at the end of the experience. Thus, for instance, a painful experience that tapers off slowly can be remembered as less unpleasant, overall, than an experience of the same kind that was both shorter and less intense, but involved no tapering period (Kahneman et al., 1993).

Now, it strikes me as perfectly reasonable for someone to wonder whether they should aim more for experiences that maximize positive affect balance, or for experiences that they will retrospectively regard as satisfying. Further, it seems perfectly reasonable for an agent who believes (correctly, let us assume) that affect balance is the true essence of subjective wellbeing, but assigns some non-trivial credence to the possibility that self-assessed life satisfaction has intrinsic significance, should make some concessions to the latter possibility, by sometimes choosing experiences that involve slightly less positive affect but that she knows will seem more satisfying, or be remembered more positively, in retrospect. And if this is right, then Weatherson's case has led us astray with respect to the rationality of hedging for prudential uncertainty.

3.4 Arpaly and Schroeder on Inverse Akrasia

Though not explicitly presented as such, Nomy Arpaly and Timothy Schroeder's influential discussion of the case of Huckleberry Finn, and the more general phenomenon that they dub "inverse akrasia," presents an apparent argument for the irrelevance of moral belief and uncertainty to subjective normativity. In Twain's novel (Twain, 1884), Huck runs away from home with the escaped slave Jim. Believing that he has wronged Jim's owner by aiding his escape, Huck is wracked by guilt and nearly turns Jim in to slave hunters. But his friendship for Jim overcomes his moral convictions, and while he continues to believe that he is acting wrongly by helping Jim, he resolves to ignore the purported demands of morality in favor of helping his friend.[7]

Arpaly and Schroeder (1999) argue that Huck's choice to help Jim escape against his own moral convictions is fully praiseworthy, despite the fact that it is

[7] The case of Huck Finn originally entered the philosophical literature by way of Jonathan Bennett (1974).

akratic, and that this phenomenon of "inverse" or praiseworthy akrasia indicates that a morally good agent should not be predominantly motivated by her moral convictions but rather by more particular, less cognitive and more affective concerns, like sympathy and friendship.

Does this case refute WSR or SSR? First, note the gap between the claim that a morally good agent should be *exclusively* motivated by her moral beliefs and degrees of belief and the much weaker claims expressed by WSR and SSR, that a morally good agent should be at least *somewhat* motivated (because she has at least *some* reason) by her moral beliefs and degrees of belief. One can accept Arpaly and Schroeder's intuition about the Huck Finn case, and deny the former strong claim, without denying either WSR or SSR. Indeed, this seems intuitively the right response to the Huck Finn case: While we are glad that Huck ultimately decides to help Jim escape, it is also to his credit that he feels *bothered* by the tension between his sympathetic motivations and his conscience. If Huck were a simple amoralist who said, straightaway, "I know that it's wrong to help Jim escape, but so what, that's nothing to me!", he would seem a less worthy character than he does. So at minimum, we should say of the Huck Finn case that his *de re* concern for the welfare of a friend properly *overrides* a weaker, but not undesirable, *de dicto* concern for the dictates of conscience.

But actually, I think we can go a bit farther than this and undercut the force of the Huck Finn intuition, in several ways. First, recall the methodological claim from the previous section that intuition pumps for second-order views concerning how agents should rationally respond to their moral beliefs and uncertainties ought not take as their subject matter first-order moral views that arouse strong first-order intuitions. This stricture is now even more salient. To block the easy response that Huck's moral beliefs about slavery lack epistemic justification—and that what he ought rationally to do is *revise* his moral beliefs on the basis of evidence and rational reflection, then help Jim escape because *that* is the right thing to do—it must simply be stipulated that Huck's moral beliefs are rationally justified. But this is a hard stipulation to internalize, if ever there was one. Is it really plausible that Huck has adequate grounds for believing that slaves are property and that justice requires runaway slaves to be returned to their owners? And even if this stipulation feels plausible, are our intuitions about how Huck should respond to this allegedly justified moral conviction not still colored by our *own* conviction that Huck's belief is, in fact, badly mistaken?

To see that these forces are at work on our intuitions in the Huck Finn case, consider instead another case of an agent deciding whether to act on justified but false moral beliefs, where the belief in question does not arouse the same negative first-order intuitions. Samuel Kerstein (2002) offers a case structurally analogous to but substantively very different from the Huck Finn case, involving a one Colonel Mikavitch.

> A morally reflective person since she was a child, [Colonel Mikavitch] has embraced the Formula of Universal Law as the supreme principle of morality...Unfortunately and unforeseeably, a foreign power has attacked the colonel's country, bent on exterminating one of its ethnic minorities.

> With the enemy nearly on her doorstep and no hope of escape, she comes to the painful conviction that if she is captured, she will, under the weight of torture, reveal a secret known only to her: the location of several minority families. After careful consideration of the alternatives, she has decided that the only way to save the families is to kill herself. The colonel finds in herself no inclination to do so and, indeed, believes that suicide would require her last ounce of courage. Although she thinks she has a moral duty to save the families, she wonders whether it is morally permissible for her to take her own life...After careful thought she judges that this maxim passes the Categorical Imperative test...With the regretful thought that she must heed the call to save innocent lives, she takes poison. (Kerstein, 2002, pp. 121-2)

Suppose, however, that Colonel Mikavitch has committed some subtle error in reasoning when testing the universalizability of her maxim, and that in fact the Formula of Universal Law, the true supreme principle of morality, does *not* permit suicide even in these extreme circumstances. Kerstein asks whether, under the circumstances, the colonel's decision might still have moral worth, and concludes that it does. For our purposes, we should ask a slightly different question: Did the colonel act *rationally*, or *as she subjectively ought to have acted*? Given her prior mistake in reasoning when she derived the obligation to commit suicide from the Formula of Universal Law, it seems clear that the colonel's subsequent decision to commit suicide was fully rational, and indeed, it is hard to feel much hint of the intuition from the Huck Finn case that it would have been better or more praiseworthy for the colonel to disregard her moral convictions and avoid suicide. Where the first-order terrain is relatively neutral, any intuition that "inverse akrasia" is deeply laudable, let alone rational, seems to dissolve.

There is another way of testing whether the Huck Finn intuition is veridical. Consider a variant of the case, which we will call Empirical Huck. Empirical Huck's beliefs are quite a bit different than those of Twain's Huck. First, Empirical Huck is a classical hedonistic utilitarian—and, let us stipulate, this moral belief is both justified and true. But second, in this version of the antebellum South, *everyone* is a convinced utilitarian, and the moral justification for slavery rests not on mistaken moral doctrines of racial superiority or confused notions of property, but rather on the *empirical* belief that the institution of slavery maximizes the general happiness (say, by increasing economic growth rates). Huck is in no position to evaluate these empirical claims, but justifiably defers to his teachers and other elders who assure him that slavery is optimific. Nevertheless, in this world as in our own, slavery is not optimific.

From here, the story of Empirical Huck unfolds like the original. Huck runs away with Jim, finds himself pricked by the conscientious conviction that he should turn Jim in for the sake of the general happiness, but can't bring himself to do it and, moved by friendship and sympathy, helps Jim escape instead. Now, I think our intuitive reaction to Empirical Huck is not substantially different from our reaction to the original case: It is good, right, praiseworthy, etc, that Empirical Huck helps Jim escape. Presumably, though, no one is likely to mistake the Empirical Huck case for knock-down proof that what an agent subjectively ought to do does not depend on her empirical beliefs but only on the empirical facts. Empirical Huck does not, for instance, refute expected utility theory. Rather, it indicates either that our intuitions in both versions of the Huck Finn case are subject to interference from powerful *first-order* intuitions, or else it simply indicates that the judgments of goodness, rightness, praiseworthiness and so forth that we readily bestow on both Huck and Empirical Huck should not be confused with subjective ought judgments or judgments of practical rationality (a point to which we will return in the next section).[8]

3.5 Harman on Moral Ignorance and Blame

One of the most explicit advocates of the irrelevantist position is Elizabeth Harman. In "The Irrelevance of Moral Uncertainty," she puts her view as follows.

[8] One possible explanation for our intuitions about both the original Huck Finn case and Empirical Huck is the psychological phenomenon known as the "curse of knowledge": Better-informed agents often find it difficult to put themselves in the shoes of worse-informed agents and disregard their own supererior information in predicting what those worse-informed agents will do. In an experimental setting, for instance, participants in a simulated market who must predict the valuation of a firm by other participants whom they know to lack certain information about the firm are nevertheless biased in their predictions by their own possession of that information (Camerer et al., 1989). Plausibly, this bias in our predictive judgments begets a bias in our normative judgments: Because *we* know that slavery is wrong, we expect others to act as if they know this as well. Even given the stipulation that Huck lacks our knowledge of the wrongness of slavery, then, the idea of his turning Jim in seems not like expected behavior in light of his epistemic state, but rather like an unexpected norm violation that could only be attributed to hidden malevolence or some other defect of character. Or, on a slightly different hypothesis, perhaps the underlying cause of the "curse of knowledge" phenomenon is a general inability to fully internalize information about other agents' epistemic/doxastic states insofar as they differ from our own, in which case our intuitions about the case simply suffer from our failure to internalize the stipulation that Huck really believes (let alone that his evidence *justifies* him in believing) that he is morally obligated to turn in fugitive slaves. This is also suggested by the phenomenon of hindsight bias, our apparent inability to accurate internalize known limitations on our own past evidence, or that of others, in assessing what ought to have been known or expected given those limitations (Roese and Vohs, 2012). This is speculation, of course, but it should be matched against the arguably paltry evidence that our intuitive judgments about cases like Huck's are veridical to begin with.

> A person's moral beliefs and moral credences are usually irrelevant to how she (subjectively) should act. How a person (subjectively) should act usually depends solely on her non-moral beliefs and credences; her moral beliefs and credences are relevant only insofar as they provide warrant for beliefs and credences about what her non-moral situation may be. (Harman, 2015, p. 58)

Harman's argument for this position is straightforward: If an agent's moral belief state were relevant to how she subjectively ought to act, then agents who act on (justified) false moral beliefs would be exculpated from blame for their wrongful actions. But false moral beliefs don't exculpate. So moral belief states are irrelevant to subjective normativity.

Both premises of Harman's argument are wrong, I think. First, it is not clear why there must be the tight connection between subjective oughts/rationality and blameworthiness that Harman supposes. Harman's defense of this premise is fairly thin: She asserts without further argument that "A person is blameworthy for some behavior only if she should not have behaved in that way," in a subjective sense of "should" that appears to align with our notion of subjective oughts/rational requirements (Harman, 2015, p. 75). But the norms that govern praise and blame might be quite different from the norms of subjective rationality. There is nothing obviously paradoxical in the idea that an agent caught in the grip of a false moral view, for which she has adequate evidence and on which she acts enkratically and as she rationally ought to act, might nevertheless be a legitimate target of blame, and a victim in that respect of bad moral luck. Indeed, certain normative views like utilitarianism will imply that this is quite often the case: If norms of praise and blame are ultimately answerable to the principle of utility, then it may often turn out that subjectively rational agents can incur legitimate blame. That Harman's argument relies on normatively substantive assumptions about what makes an agent deserving of blame, and hence is inconsistent with at least some first-order moral views, is a point to which we will return in the final section.

Absent the link to subjective rationality, Harman can have her claims that false moral belief does not exculpate, without this constituting a refutation of WSR or SSR. As I said in §2.1, my interest is only in the connection between an agent's moral belief states and the basic action-guiding normative system that we have been speaking of in terms of subjective oughts/rational requirements. (Likewise, then, even if Weatherson is correct that *de dicto* motivation is a defect of character or Apaly and Schroeder are right that Huck is more virtuous and praiseworthy for ignoring his moral beliefs, these claims need not be seen as inconsistent with the theses I am concerned to defend, nor need they undercut the project of searching for principles of *rationality* that govern choice under moral uncertainty.)

Still, I think Harman's second premise—that false moral beliefs are never exculpatory—is probably wrong as well, and the support she offers for it insufficient. She argues for this premise by means of a few thought experiments, of which the following is a representative example.

> Max works for a Mafia "family" and believes he has a moral obligation

> of loyalty to the family that requires him to kill innocents when it is necessary to protect the financial interests of the family. This is his genuine moral conviction, of which he is deeply convinced. If Max failed to "take care of his own" he would think of himself as disloyal and he would be ashamed. (Harman, 2015, p. 65)

Harman claim that, if Max commits murder on behalf of his crime family, he is a "paradigm case" of an agent blameworthy for a wrongful action.

Two remarks: First, note once again the misleading force of first-order intuitions in this case. We should all feel, forcefully and without hesitation, that the *particular* false moral belief that Max has accepted is badly and obviously mistaken, and this prejudices us against thinking that his belief might be exculpatory. Likewise, it is difficult to appropriately internalize the stipulation (which Harman recognizes must be made, to counter responses from the proponent of theses like WSR and SSR) that Max's moral belief is fully justified, that he has made no mistakes of theoretical rationality in reaching this belief that he ought to have corrected before acting. Thus, our intuitions about the Max case are likely to be misleading.

But further, a single case in which false moral beliefs fail to exculpate does not Harman's case make for the much stronger premise, essential to her argument, that false moral beliefs are *never* exculpatory. If one shares Harman's judgment about Max, then, consider by contrast Kerstein's case of Colonel Mikavitch from §3.4. It seems perfectly intuitive that the colonel's justified false belief that the Categorical Imperative requires her to commit suicide at least goes some way toward *mitigating* her blameworthiness for violating the moral law. One can easily imagine other such cases where an agent acts on moral views that are false but neither unreasonable or horribly pernicious, out of a sincere sense of conscientious obligation, and in such cases it generally seems intuitive that the agent's moral belief state should affect her liability for blame.[9]

3.6 Concluding Thoughts

I will close with three general observations about the irrelevantist position and the arguments for it that we have just surveyed.

[9] At this point I can't resist quoting a few lines from Gilbert and Sullivan's *The Pirates of Penzance*, in which the plot revolves around just such a case of sincere though mistaken moral belief. As a young boy, the protagonist Frederic is mistakenly indentured to a pirate by his nurse (who mishears the word "pilot"), with his apprenticeship to last until his twenty-first birthday. Frederic "abhors [the] infamous calling" of piracy, but nevertheless does his best as a member of the pirate band because, as he says, "[it] was my duty under my indentures, and I am the slave of duty."

As the story opens, Frederic believes his apprenticeship has expired and that he can return to civilized society to atone for his piratical crimes. In the following exchange between Frederic and the Pirate King, the question of blame and moral belief is raised, and the Pirate King eloquently expresses a view of the subject that is sharply at odds with Harman's.

Frederic. "Individually, I love you all with affection unspeakable, but, collectively, I look upon you with a disgust that amounts to absolute detestation. Oh! pity me, my beloved friends, for such is my sense of duty that, once out of my indentures, I shall feel myself bound to devote myself heart and soul to your extermination!"

All. "Poor lad — poor lad!" (All weep.)

Pirate King. "Well, Frederic, if you conscientiously feel that it is your duty to destroy us, we cannot blame you for acting on that conviction. Always act in accordance with the dictates of your conscience, my boy, and chance the consequences."

Before he can put his plan into effect, however, Frederic learns to his shock that he was born on Leap Day, and that as a result he will not reach his twenty-first birthday and be free of his indentures until age 84. Horrified but obedient to his sense of duty, Frederic rejoins his pirate companions, rather than leading the local police force to arrest them. Informed of this development, the police are nonplussed:

Sergeant. "This is perplexing."

Police. "We cannot understand it at all."

Sergeant. "Still, as he is actuated by a sense of duty?"

Police. "That makes a difference, of course. At the same time we repeat, we cannot understand it at all."

It seems to me that the police, like the Pirate King, have the right intuition: Frederic's conscientious belief that he is duty-bound to rejoin his pirate comrades *does* make a difference to our moral assessment of his actions, of course.

First: The irrelevantist view involves a curious setting-aside of the first-person perspective on rational decision-making. How is someone who accepts a view like Weatheron's or Harman's to actually go about making moral choices in the face of moral uncertainty (for I assume it cannot be denied that we should all be at least a little bit uncertain about the big questions of morality)? She knows that she *ought* to do whatever the true moral theory implies she ought to do, but she does not know—at least, does not know with certainty—what the true moral theory is. So how will she proceed, when her view of subjective rationality gives her instructions with which she does not know how to comply?

I can see only two alternatives: Either the irrelevantist position collapses into something like My Favorite Theory, in that the agent who accepts irrelevantism acts most rationally by taking her best guess and acting on the recommendations of the theory she judges *most likely* to be correct — or it collapses into near-complete arationalism about moral decision-making, since in the absence of any norms the agent knows how to apply, she has no rational means of choosing among possible actions. The latter option seems wildly implausible: There must be *some* way for an agent to reason, non-arbitrarily, from her beliefs about morality to conclusions about how she should respond to moral dilemmas. The former option is more defensible, but to adopt it is to abandon irrelevantism and concede at least WSR. The advocate of My Favorite Theory is "in the business" of looking for norms of rationality that take account of an agent's moral beliefs, just as much as the advocate of a hedging view. If there is some other way, besides these two, of understanding how an irrelevantist decides what to do in the face of moral quandaries, I don't know what it could be.

Second: Even if one finds oneself moved by intuition pumps like Williams' "one thought too many," Weatherson's Bob, Arpaly and Schroeder's Huck Finn, or Harman's Max, the intuition behind the enkratic principle (or some suitably revised variant thereof) should seem like a far better bet. Apart from the specific reasons I have given for taking the former group of intuitions with a grain of salt, there are good reasons to be skeptical of reliance on case intuition in normative ethics and metaethics generally (see, for instance, Singer (2005)). On the other hand, the intuitions behind the enkratic principle (which arguably also lend their support to stronger principles, like MEC, that vindicate SSR as well as WSR) are of a very different and seemingly more reliable kind. If rationality is, conceptually, a matter of internal coherence, then it is seems like a conceptual truth (or even a tautology) that rationality requires an agent's motivations to align in some fashion with her beliefs about the good or about objective choiceworthiness. So if it comes down to a weighing of competing intuitions, the intuitions behind the *prima facie* case at least for WSR, and possibly for SSR, are strong enough to overwhelm their putative defeators.

Third and finally: It strikes me that there's something suspect, in the context of asking how agents ought to act under moral uncertainty, about appealing to substantive, first-order moral claims about things like virtue, praiseworthiness, or blameworthiness. This move is suspect just as it would be suspect, in developing a theory of how an agent should act when she is uncertain about the consequences of her actions, to appeal to claims about those consequences—the very sort of claim about which she is uncertain.[10]

The particular claims about virtue and blame on which Weatherson/Arpaly & Schroder/Harman depend are not trivial, but significantly constrain the space of possible moral theories. On a thoroughgoing consequentialist moral theory, for instance, none of their claims about virtue and vice (*de re* vs. *de dicto* motivation), praiseworthiness (inverse akrasia), or blameworthiness (Harman's Max) are even candidates for general truths about morality. Whether an agent is blameworthy for an act, to a strict consequentialist, must reduce to the question whether it is optimific to blame her for that act.[11] And whether *de dicto*, "fetishistic" motivations are a virtue or a vice, likewise, reduces to the question whether they lead an agent to take better actions, compared to alternate motives by which she might otherwise be moved. On these normative views, such assessment claims become fairly uninteresting from a metaethical standpoint, and even if contingently true, cannot be called on to do the work that the advocates of irrelevantism require.

Of course, not everyone (perhaps not even most consequentialists) will accept an equivalence between "blameworthy" and "ought to be blamed," and hence one

[10] As MacAskill points out in a similar context, "it would clearly be incorrect to argue against [maximizing expected choiceworthiness under normative uncertainty] 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." (MacAskill, 2014, p. 40)

[11] Sidgwick writes: "From a Utilitarian point of view, as has been before said, we must mean by calling a quality 'deserving of praise,' that it is expedient to praise it, with a view to its future production: accordingly, in distributing our praise of human qualities, on utilitarian principles, we have to consider primarily not the usefulness of the quality, but the usefulness of the praise..." (Sidgwick, 1874, p. 428). This passage is cited by Norcross, who adds that "The utilitarian will, of course, say the same about censure as Sidgwick says about praise: we should assess whether it is good to punish or blame someone by assessing the utility of doing so. Punishing and blaming are actions just like promisekeeping and killing and, like those actions, their value is determined by their consequences, their power to produce utility" (Norcross, 2006, p. 225).

might reject the claim that praiseworthiness and blameworthiness are essentially first-order normative notions.[12] Still, the existence of candidate first-order views that imply this sort of equivalence means that the arguments for irrelevantism we have surveyed are not neutral with respect to first-order normative theories.

It could turn out, I suppose, that the second-order question, what agents ought to do under moral uncertainty, *depends* on what first-order moral theory is true. But this in itself would not constitute a victory for irrelevantism. It would imply only that irrelevantism is true *if* some first-order theory is true that supports it (e.g., the theories of virtue and blameworthiness the the advocates for irrelevantism presuppose). Another kind of first-order theory, like consequentialism, might support My Favorite Theory, My Favorite Option, or a hedging view.

If the sorts of arguments for irrelevantism that we have been engaged with are to be accepted as methodologically admissible, despite their reliance on first-order normative claims that are potential objects of moral uncertainty, then we must accept the dispiriting conclusion that we cannot know what norms of rationality govern choice under moral uncertainty until we have resolved the very uncertainties that make this question seem so pressing. On the other hand, if this sort of "upward dependence" relationship between first-order moral theories and second-order norms of rationality seems improbable, then there are methodological grounds to reject the extant arguments for irrelevantism out of hand.

[12] An alternative analysis would equate "blameworthy" with "deserves to be blamed." But like blameworthiness, the notion of desert is either simply absent from classical utilitarianism or must be reduced to the principle of utility in such a way that either the fundamental principles of desert or the normative significance of those principles becomes a contingent, empirical matter.

Chapter 4: Dominance-Based Approaches

Having defended the claim that principles of rational decision-making must be sensitive to an agent's moral beliefs and degrees of belief, we can now ask *what* principles govern a rational agent's choices under moral uncertainty. Among philosophers who have sought to give a systematic treatment of the moral uncertainty problem (e.g. Lockhart (2000), Sepielli (2009)) the goal has generally been to give an *expected value theory* that assigns quantitative weights to the considerations put forward by rival moral perspectives and then discounts those weights according to the degree of belief the agent assigns to each perspective.

As we will see in Chapters 5-6, however, such expected value approaches encounter a very formidable obstacle: the apparent incomparability of moral considerations put forward by competing moral theories, especially theories like Kantianism and utilitarianism that differ radically in their basic structure. I will eventually argue that this problem can be at least partially overcome. But before we engage with the difficulties this task presents, it is worth asking what progress can be made by more modest means.

In the moral uncertainty literature, there is a widely discussed class of cases that have the following form: (i) An agent is faced with exactly two options, O and P.

(ii) It looks pretty clear that O is not morally required, and hence that P is morally permissible. But (iii) there is a substantial possibility that O is morally prohibited, and hence that P is morally required. Typically O is a morally risky *action* while P is a morally safe *omission*, at least by our intuitive sense of the act/omission distinction. Two widely discussed cases that fit the above schema are abortion and meat consumption: while it's pretty clearly morally permissible to *avoid* having an abortion or eating meat, there is at least a substantial possibility that having an abortion or eating meat are morally impermissible. As Moller (2011) puts it, these cases present an "asymmetry" of moral risk. As such they seem like low-hanging fruit for any theory of choice under moral uncertainty: the first interesting conclusion that such theories are likely to yield is that we have reason to avoid actions that carry such asymmetric or uncompensated moral risk.

A natural way of expressing this idea is by appeal to *dominance reasoning*. In the context of reasoning under empirical uncertainty, a *weak* dominance principle states that, if you are uncertain about the state of the world, but believe that *conditional on any possible state of the world* option O is better than option P, then you are rationally required to prefer O to P. What we want, however, is an analogue of the stronger dominance principle stating that, if you are uncertain about the state of the world, but believe that (i) conditional on any possible state of the world option O has *at least the same* value as option P and (ii) conditional on *some* possible state of the world option O is better than option P, then you are rationally required to prefer O to P.[1]

[1] Somewhat confusingly, the weaker dominance principle is typically referred to as "Strong

MacAskill (2013) offers the most careful analysis to date of dominance-based approaches to moral uncertainty. His initial statement of a strong dominance principle ranging over moral theories goes as follows:

> *Dominance over Theories (DoT)* — If some theories in which you have credence give you subjective reason to choose x over y, and no theories in which you have credence give you subjective reason to choose y over x, then, rationally, you should choose x over y. (MacAskill, 2013, p. 511)

As MacAskill argues, and as we will see, this principle requires at least one important revision. But it will provide a useful jumping-off point for our discussion.

It's not clear, sadly, that a principle like DoT can get much traction on the abortion and vegetarianism cases. For while these cases present a stark asymmetry of *moral* risk, there may be substantial *non-moral* reasons that, for particular agents, favor having an abortion or eating meat. Hence, in these cases, agents will be likely to have credence in theories according to which they have overall subjective reason to choose the morally risky option (abortion or meat), and so the antecedent of DoT will not be satisfied.

If one is inclined to think that moral reasons always override non-moral reasons, one might propose a modification of DoT that replaces "subjective reason" with "subjective *moral* reason." Such a principle would imply that asymmetric

Dominance" since it prescribes only the elimination of *strongly* or *strictly* dominated options, while the stronger principle is typically referred to as "Weak Dominance" since it *also* prescribes the elimination of *weakly* dominated options. It seems more natural to me, however, to call the weak principle "weak" and the strong principle "strong," so this is what I'll do, where the distinction needs to be made. In general, though, the following discussion will focus exclusively on the stronger principle and its analogues for moral uncertainty.

moral risks should *always* be avoided, even when there are strong non-moral reasons in favor of running these risks. But such a principle is controversial, to say the least—and, perhaps more importantly for our purposes, it is a principle about which agents should be substantially uncertain, i.e., reasonable agents will have positive credence in normative theories that reject this principle.

Instead of the abortion and vegetarianism cases, then, let's focus on a different pair of cases that seem like even lower-hanging fruit for a dominance-based approach. The first, which has received a certain amount of attention in the recent literature, concerns "deflationary" moral views like skepticism or nihilism, and the possibility that dominance reasoning can render such views practically inert. The second, which has not yet received much attention, is supererogation. In both these cases (I will argue), agents may be all-but-certain that some practical option is at least permissible and potentially obligatory, and that another, incompatible option is at best permissible and potentially forbidden. It is in cases like these that dominance reasoning holds the most promise for making progress on the problem of moral uncertainty.

The objective of this chapter, then, is to assess the potential for dominance-based approaches to moral uncertainty through an examination of these two cases. As we will see, the cases are anything but straightforward, and I will ultimately conclude that dominance reasoning alone is insufficient even in these apparently most promising contexts. But there will be an optimistic conclusion as well: Insofar as dominance reasoning can be conceived not as a freestanding principle but rather as a limiting case on which various expected value theories and other "hedging" approaches to moral uncertainty converge, such reasoning *can* be coherently leveraged to bolster the strength of moral reasons in the face of all but the most extreme nihilist or supererogationist challenges—implying that very often at least, moral considerations are still sufficient to generate rational requirements even when agents have reason to be skeptical of the objective requiring force of those considerations. So, while as a theoretical matter we cannot circumvent the more difficult task of developing a stronger theory of rational choice under moral uncertainty, we can as a practical matter have a great deal of confidence that whatever the correct comprehensive theory of moral uncertainty turns out to be, it will yield certain non-obvious conclusions: for instance, that the possibility that all actions might be morally indifferent or that many morally good actions might be merely supererogatory does not negate the reason-giving force of moral considerations or, in most cases, the rational requirement to act morally.

4.1 Moral Nihilism

Let's begin with moral nihilism. Ross (2006) argues that we may almost always disregard what he calls "absolutely deflationary" ethical theories according to which it's never the case that one course of action is more choiceworthy than another, whatever our credence in such theories. He distinguishes, within this class, between "uniform" theories according to which any pair of options is equally choiceworthy and "nihilistic" theories according to which "the notions of good and bad and of right and wrong are illusions and...objectively speaking, no option or state of affairs is better than any other, nor are any two options or states of affairs equally good" (Ross, 2006, p. 748).

Nihilism presents the more difficult and interesting case for Ross's dominance argument, so it is here that I will focus. Ross imagines a case in which he is nearly certain that moral nihilism is true and finds himself faced with a trolley dilemma. In such circumstances, he claims, he may simply disregard his credence in nihilism for purposes of practical deliberation. He supports this claim by the following dominance argument.

> [S]uppose...that I have a degree of credence of .01 in T_L [a non-nihilistic moral theory that prescribes turning the trolley to the left], but...I have a degree of credence of .99 in a nihilistic theory T_N. And again suppose that I must decide between sending the trolley to the right and sending it to the left. In this case we could reason as follows. 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 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.

(Ross, 2006, p. 748)

This line of reasoning seems straightforward, but it faces at least two significant objections.

4.1.1 The Cyclicity Problem

The first objection, raised by MacAskill (2013), suggests that the dominance principles on which Ross's anti-skeptical argument depends generate a number of intolerable results. In particular, attributing to Ross the principle DoT given above, MacAskill shows that for an agent who has positive credence in one or more theories that posit value *incomparability*, DoT has the unwelcome consequence of generating preference cycles. The demonstration of this result depends on a pair of complicated thought experiments. Stated schematically, and leaving aside the details MacAskill offers to illustrate this schema, the simpler of the two thought experiments goes like this: Agent A has three options O, P, and Q, and divides her moral beliefs among three theories, T_1, T_2, and T_3. T_1 implies that O is better than P and that Q is incomparable to both. T_2 implies that P is better than Q and that O is incomparable to both. And T_3 implies that Q is better than O and that P is incomparable to both. Hence, by DoT, $O \succ P \succ Q \succ O$ (MacAskill, 2013, pp. 513-514).

To avoid cyclicity, MacAskill claims, we must limit ourselves to dominance principles whose conditions of application exclude incomparability. In particular, MacAskill replaces DoT with what he calls

Genuine Dominance over Theories (GDoT) If some theories in which you have credence give you subjective reason to choose x over y, and all other theories in which you have credence give you equal subjective reason to choose x as to choose y, then, rationally, you should choose x over y. (MacAskill, 2013, p. 518)

This weaker principle is bad news for Ross's argument since nihilistic theories, as Ross has characterized them, do not claim that we have equal subjective reason to choose either option from any pair, but deny that notions like value, choiceworthiness, and subjective reasons are meaningful and hence that any comparisons in terms of these notions, even comparisons of equality, can meaningfully be made. Thus MacAskill concludes that dominance reasoning cannot justify rejecting nihilism.

It seems to me, however, that there is a relatively straightforward escape from the cyclicity problem, and that Ross has simply made a very inconvenient mistake by treating equality as a *positive* value relation, such that nihilistic theories deny that any two things are equally good. Consider by contrast John Broome's proposed definition of equality with respect to a property F: "'x is equally as F as y' means that [i] x is not Fer than y, and [ii] y is not Fer than x, and [iii] anything that is Fer than y is also Fer than x, and [iv] y is Fer than anything x is Fer than" (Broome, 1997, p. 72). If nihilism is true, then all four clauses in Broome's definition are trivially satisfied for any x and y and any evaluative property F (e.g. "good," "right," "choiceworthy," "supported by objective/subjective reasons"): If nothing is better than anything else, then x is not better than y, y is not better than x, and since neither x nor y is better than anything, it is vacuously true that for anything either x or y is better than, the other is better as well. Furthermore, by virtue of these last two clauses, Broome's definition distinguishes (as Broome intends it to) between equality and other relations like parity and incomparability in the context of non-nihilistic theories.

Thus, we can strengthen MacAskill's GDoT by simply replacing the positive relation of equality, which nihilistic theories decline to attribute to any pair of options, with Broome's negative relation. (Let's call the latter "equality*," for ease of reference.)

Strengthened Genuine Dominance over Theories (GDoT*) If some theories in which you have credence give you subjective reason to choose x over y, and all other theories in which you have credence give you equal* subjective reason to choose x as to choose y, then, rationally, you should choose x over y.

GDoT* allows Ross's anti-nihilistic argument to go through, while remaining immune to the preference cycles brought on by the incautious principle DoT: In MacAskill's case described above, T_1 does not imply that there is equal* subjective reason to choose O as to choose Q, since there is greater subjective reason to choose O than to choose P but not greater subjective reason to choose Q than to choose P, nor is there equal* subjective reason to choose P as to choose Q, by precisely the same reasoning (and likewise, *mutatis mutandis*, for T_2 and T_3).[2]

[2] For purposes of this argument, it makes no difference whether we frame GDoT* in terms of support by subjective reasons or any other normative property like rightness or choiceworthiness.

Thus, the dominance argument for rejecting nihilism can escape the threat of cyclicity. However, as we will see in §4.3.1, MacAskill's problem will reappear as a challenge to the dominance argument for rejecting *supererogationism*, and will somewhat constrain the scope of that argument.

4.1.2 Nihilistic Undermining

There is another difficulty that strikes me as more worrisome for Ross's argument and which has yet to receive attention in the literature. This is the worry that principles like DoT, GDoT*, and any other dominance principle to which Ross's argument might appeal seem to be *undermined* by the very same nihilistic hypotheses that the argument is meant to justify rejecting—or at least, by some of those hypotheses.

Consider a form of nihilism I will call *Democritean normative error theory* (DNET). According to this view, only fundamental entities exist (flatworldism[3]) and all fundamental entities are physical (physicalism). "Reasons," "obligations," "goods," "values," etc, are not fundamental physical entities, so no reasons, obligations, goods, values, etc exist. For any normative/evaluative proposition to be true, at least one such entity must exist to serve as its truthmaker. Therefore, DNET holds, no normative/evaluative propositions are true.[4]

Is such an extreme view plausible? Though I don't assign it the greater part

[3] I borrow this term from Karen Bennett (2011).

[4] The name I have given this view is not meant to imply an attribution to Democritus. But the spirit of the view is neatly captured by his famous adage: "By convention sweet and by convention bitter, by convention hot, by convention cold, by convention color; but in reality atoms and void" (Taylor, 1999). The DNETist just adds to this list: "by convention right, by convention wrong, by convention rational, by convention irrational...but still, just atoms and void."

of my own credence, it seems to me that DNET is a more reasonable contender for truth than many popular normative views (I leave it to the reader to fill in examples), and my impression is that a significant number of philosophers accept something like it (though, by self-selection, few of these are normative ethicists). It combines two widely held views in physicalism and flatworldism, and draws a plausible implication from the conjunction of these views.[5] I will assume, then, that DNET is plausible enough, as nihilistic theories go, to merit some investigation.[6]

Now, Ross's argument suggests that no matter how high my confidence in a nihilistic theory like DNET, I should nevertheless accept a second-order dominance principle like GDoT*.[7] The problem is this: Since GDoT* expresses an affirmative

[5] Of course, even if one accepts physicalism and flatworldism, one might still try to make normative propositions true by letting atoms and void alone serve as their truthmakers. This could be done by paraphrasing away existential claims concerning reasons and the like ("atoms arranged reason-wise," perhaps), by eschewing such talk entirely, or by defending it as ineliminable but ontologically non-committing. Since my goal is not to defend DNET, I won't attempt to assess these strategies individually. I do, however, find them at least mildly implausible. Suffice it to say that normativity in the most interesting sense (features of the world that *count in favor of* actions and make our actions *matter*) must be something that carves the world at a fairly fundamental joint, and analyses of normative concepts that make them *logically* or *conceptually* (rather than metaphysically) supervenient on atoms and void appear to carve no such joint.

[6] Ross attempts to avoid the undermining worry by fiat, restricting the scope of his argument to "a kind of nihilism that denies that there are any objective reasons for action but that concedes that if one held that there were such reasons, then it would be subjectively rational to act in accordance with them" (Ross, 2006, p. 749n). In fact, Ross's stipulation must go a bit further than this, since his dominance argument is meant to apply not only to agents who *hold* (i.e., fully believe) that there are objective reasons for action but to any agent who *has positive degree of belief* that there are such reasons. Thus, the version of nihilism to which Ross restricts his attention must deny that there are any objective reasons for action but also *affirm that there are subjective reasons for action*, and that any agent with non-zero credence in the existence of objective reasons has subjective reasons. There is a significant *prima facie* tension, however, between denying the existence of all objective reasons and asserting the existence of subjective reasons, and Ross gives no argument that plausible nihilistic theories should be understood in this way. So it seems that this self-imposed restriction would limit the scope of Ross's argument to a relatively implausible subset of nihilistic theories.

[7] A variety of other dominance-like principles could be offered to facilitate Ross's argument. MacAskill considers a second principle called "Modified Dominance over Theories" (MDoT) that achieves the same result as DoT (but shows that this principle still generates diachronic preferences cycles over certain sequences of pairwise choices) (MacAskill, 2013, pp. 515-6). Weatherson attributes a different principle to Ross which he calls "ProbWrong," then suggests a schematic version of this principle called "GeneralPrinciple" that could be filled in by various evaluative

normative proposition (a "should" claim), DNET $\vdash \neg$GDoT*. So if I accept DNET with subjective probability p, I must assign GDoT a degree of belief no greater than $(1 - p)$. If $p > .5$, then Cr(GDoT*) $< .5$. And, assuming that my beliefs ought to be consistent, if I *believe* DNET, I ought to disbelieve GDoT*.[8] But how can it be rational, let alone rationally required, to act on a principle that I regard as more likely false than true, or that I believe to be false?

It seems to me that the solution to this undermining worry, if there is one, must involve taking the relevant principle of rational requirement—a dominance principle like GDoT* or some stronger principle from which it can be derived—to have "external," belief-independent force in Weatherson's sense, such that the agent is subject to its requirements even if she justifiably disbelieves it. Whether we should accept normative externalism is a question I will not attempt to resolve in this chapter. (In Chapter 7, however, I will argue that the enkratic principle, rightly formulated, has external rational requiring force, and propose several candidate formulations each of which plausibly implies GDoT*.) But we will shortly return to the challenge of undermining, which arises in a closely analogous form in the context of the dominance argument for rejecting supererogationism. There I will argue that the versions of supererogationism that threaten to undermine the relevant

predicates to facilitate the same dominance arguments (Weatherson, 2014, p. 146). But the "undermining" objection described in this section applies equally and straightforwardly to all these principles.

[8] If GDoT* is read as a material conditional, this argument becomes slightly more complicated—GDoT* then is not itself an affirmative normative proposition, but implies such a proposition when conjoined with empirical propositions about my credence over theories that I presumably accept (and that I must accept, for GDoT* to apply to me and for Ross's argument to gain traction). But this does not affect the substance of the undermining problem: GDoT* is still inconsistent with the conjunction of empirical facts about my credal state (which make its antecedent true) and DNET (which makes its consequent false).

dominance principles are less plausible than the corresponding versions of nihilism, and hence that the issue of normative externalism is at least less urgent in the supererogation context.

4.2 Supererogation

Let's turn now to our second test case for dominance reasoning, the case of supererogation. Suppose agent A must choose between options O and P. Given these options, she is uncertain whether O is morally required or merely supererogatory and correspondingly, she unsure whether P is forbidden or permissible. In this circumstance, dominance reasoning seems to suggest that A is rationally required to choose O, despite having some (perhaps quite high) degree of belief that P is objectively permissible.

Paralleling (or rather, parroting) Ross, the argument can be put like this: Call the moral theory according to which O is objectively morally required T_R and the theory according to which it is merely supererogatory T_S. According to T_R, A is morally required to choose O and forbidden to choose P, so A's credence in T_R certainly seems to give her *pro tanto* subjective reason to choose O. The only way this could fail to be the most rational option would be if A's credence in T_S gave her sufficiently strong reason to choose P instead of O. But although T_S implies that it is *permissible* to choose P, it certainly does not imply that it would be *better* to choose P than to choose O. At most, it implies that the two options are equally choiceworthy (equally "eligible," one might say). Hence the *pro tanto* subjective reason to choose O is left unopposed, and so this is the rational option.

4.2.1 "All-Things-Considered" Supererogation

This argument depends crucially on the claim that, according to plausible supererogationist theories, the strength of one's all-things-considered reasons for choosing a supererogatory option O_S is (often or always) equal to or greater than the strength of one's all-things-considered reasons for choosing any of its merely-permissible alternatives. Let's call option O_S *morally* supererogatory iff for some alternative option O_P, O_S is morally better than O_P, but O_P is morally permissible; and let's call O_S *all-things-considered* supererogatory iff for some alternative option O_P, O_S is better than O_P all-things-considered (that is, supported by stronger all-things-considered reasons) but O_P is still all-things-considered permissible (that is, permissible as a matter of objective rationality). In these terms, then, the claim on which the dominance argument for rejecting supererogationism depends is that plausible supererogationist theories should regard most or all supererogatory options not just as morally supererogatory, but as supererogatory all-things-considered.[9]

This claim is *prima facie* plausible when we reflect on paradigm cases of supererogation, like donating a moderately substantial portion of one's income to charity. Even if we regard such actions as merely supererogatory, we do not re-

[9] This restatement sets aside the possibility of exact equality in the strength of reasons for performing O_S and O_P. Equality of reasons is, in general, extremely fragile, being disturbed by any increase or decrease in the strength of reasons favoring either option. Since it is fragile, equality is also extremely rare. Thus, while the dominance argument for rejecting supererogationism is compatible with the possibility that there is equal all-things-considered reason to perform O_S as to perform O_P, it will far more often be the case that there is at least slightly greater all-things-considered reason to perform O_S.

gard an agent who performs them as acting irrationality, contrary to the all-things-considered balance of reasons. We can of course describe cases where we might be inclined to say that, although an agent has acted well from the moral point of view, she has acted on comparatively weak moral reasons and against stronger prudential reasons, and therefore has acted irrationally overall: think of Ned and Maude Flanders sending a once-disheveled stranger off in Ned's best suit, with the promise to sleep on card tables if he ever wants the use of their master bedroom.[10] But such cases go well beyond ordinary acts of supererogatory altruism, and our attitude toward them is quite different from our attitude toward, say, someone who donates five percent of her weekly paycheck to GiveDirectly. Unlike the absurd self-abnegation of the Flanders family, common sense regards these more reasonable forms of altruism as admirable and praiseworthy, not just from a moral point of view, but *simpliciter*.[11] Though many of us judge that even moderate altruism (e.g. at the 5% level) is supererogatory rather than obligatory, it is hard to imagine that any plausible theory of all-things-considered practical rationality (so long as it incorporates non-egoistic moral considerations at all) will judge that we have less overall reason for making such sacrifices, when they will greatly improve the lives of the very badly off, than we have for spending the money on ourselves.

One might think, however, that this must be a mischaracterization of our

[10] *The Simpsons*, S3 E24, "Brother, Can You Spare Two Dimes?"

[11] Likewise, it is the Flanders-like extremes of altruism that give intuitive plausibility to Susan Wolf's skepticism about "moral saints" (Wolf, 1982). As Wolf argues, a life of moral perfection may be genuinely better from the moral point of view than a more ordinary life, but either worse or at least incomparable all-things-considered because of the nonmoral goods that must be sacrificed in the pursuit of moral perfection. But one can grant this conclusion without denying that *some* voluntary altruism in excess of the bare moral minimum may be good and praiseworthy from an all-things-considered (and not merely a moral) point of view.

commonsense judgments, since the all-things-considered understanding makes the notion of supererogation self-evidently and intolerably paradoxical. If rationality always requires us to choose the option that is supported by the strongest all-things-considered reasons, then the definition of all-things-considered supererogatory options as more strongly supported by reasons and yet rationally optional is a simple contradiction in terms, and so cannot be what is posited by plausible supererogationist moral theories.

There are a great many proposals in the literature that aim to resolve this "paradox of supererogation" and render the notion of the supererogatory coherent with a general theory of rationality. Fortunately, most of these proposals are consistent with if not actively supportive of the commonsense judgment that ordinary supererogatory acts better and more admirable than their merely-permissible alternatives, not just from the moral point of view, but *simpliciter*. Portmore (2008), for instance, considers five possible ways of understanding supererogation as part of a general theory of rational options, several of which allow us to make sense of the idea of all-things-considered supererogation. Specifically, we may understand all-things-considered supererogation in terms of a *satisficing conception of rationality* (that treats the merely-permissible alternative to a supererogatory act as rationally permissible because it meets some satisficing threshold, even though there is stronger overall reason to perform the supererogatory act); in terms of a distinction between the *justifying strength* and *requiring strength* of reasons (according to which the reasons that make the supererogatory option better supported by reasons overall nevertheless lack the requiring strength of produce a rational (or moral) requirement)[12]; or in terms of *imperfect reasons* (which, in supporting the supererogatory act, make it at least as rational all-things-considered as any alternative, but do not generate a rational requirement insofar as they can be satisfied by the agent performing other morally good acts in other circumstances).[13] In a similar vein, Raz (1975) proposes to understand supererogation in terms of what he calls "exclusionary permissions," which empower an agent to set aside or exclude the reasons that, on balance, favor a particular course of action.[14]

Finally, Urmson's classic article in defense of supererogation (Urmson, 1958) proposes that supererogatory acts be understood as differing from acts of moral obligation simply in that the difficulty of complying with the more demanding principles of supererogatory moral behavior means that demanding such behavior of one another, treating supererogatory acts as obligatory under our shared moral scheme, is likely to undermine the normative force of moral requirements in general to an extent that outweighs any potential social gains from compliance. On this sort of broadly pragmatic understanding of supererogation there is, likewise, no difficulty in the idea that many if not all supererogatory acts are more strongly supported by

[12] Horgan and Timmons, among others, have defended this understanding of supererogation: "What the case of Olivia suggests, then, is the idea that not all good moral reasons for an agent to perform some action, even reasons that are plausibly considered 'best,' are such as to require that she perform that action, even prima facie. Some moral considerations clearly do have a requiring force, but (we submit) others need not." (Horgan and Timmons, 2010, p. 50)

[13] The other two possibilities Portmore considers, drawing on the ideas of parity and incomparability/rough comparability between reasons respectively, are incompatible with the argument I have given, as we will see below. But, as I will argue, they are also incompatible with our commonsense judgments of the merits of supererogatory acts.

[14] "A person may have an exclusionary permission to perform an act even though there are conclusive reasons for him not to perform it, provided that he is entitled not to act for those reasons, to exclude them from his considerations. In other words: To say that a person is permitted to perform an act is to say that he may perform it, i.e., that he does nothing wrong in performing it. He is permitted to perform the act because there are no conclusive reasons against doing so or because he may exclude such reasons from his considerations." (Raz, 1975, p. 163)

reasons as their merely-permissible alternatives.

To come at the point from the opposite direction, there are just three possible understandings of supererogation with which the argument in the preceding section is incompatible. For ordinary cases of a putatively supererogatory option O_S and a merely permissible alternative O_P, in claiming that the all-things-considered reasons to perform O_S are equal to or stronger than all-things-considered reasons to perform O_P, I have thereby denied that (a) there is greater reason to perform O_P than to perform O_S, that (b) the reasons favoring the two options are incomparable, or that (c) the reasons favoring the two options are on a par (a fourth relation of evaluative comparison, supplementing the traditional trichotomy of better/worse/equal, proposed and defended by Ruth Chang, e.g. in Chang (1997b)). (That I must deny (b) and (c) as well as (a) will be seen in the following section.)

The first of these possibilities, that supererogatory acts are generally opposed by the balance of all-things-considered reasons, is strongly contradicted by our commonsense evaluative judgments, and has not found advocates in the philosophical literature.[15] The second possibility, that of incomparability, is likewise highly implausible, since an enormous number of our everyday moral judgments presuppose comparability between the sort of altruistic reasons that commonly favor supererogatory acts and the sort of prudential reasons that oppose them. For instance, the judgments on the one hand that one has stronger reason to save the drowning child in Singer's pond that to preserve one's new pair of shoes (Singer, 1972), and on the

[15] Douglas Portmore has come nearer than others to defending this view, but disclaims it in Portmore (2008, p. 382n).

other hand that Ned Flanders acts unreasonably in sacrificing his every personal interest for the sake of whatever needy stranger lands on his doorstep, both indicate that moral and nonmoral reasons are far from incomparable.

The most plausible understanding of supererogation that my argument must exclude, then, is one that draws on the notion of parity. On such an understanding, the strength of all-things-considered reasons for choosing some supererogatory option O_S is neither greater than, nor less than, nor equal to the strength of all-things-considered reasons for choosing some alternative O_P, but a sufficient strengthening or weakening of the reasons on either side (for instance, an increase in the good that might be done for others by, or in the personal costs of, performing O_S) would generate a definite inequality, making one option more rational overall.

It seems to me, however, that this parity-based account of supererogation is still inconsistent with our ordinary attitudes toward supererogatory acts. Parity implies a kind of symmetry between options: Neither option is better than the other, and either option could come to be better by the addition of reasons in its favor, or the subtraction of reasons for its alternative. But our normal attitudes toward the agent who gives more of her income to charity than morality strictly requires do not reflect such a symmetry: our admiration for her willingness to forgo creature comforts of the sake of others in greater need is not mirrored by a deprecating judgment that she is somewhat soft-headed or imprudent, nor by an equal admiration for the steely-eyed prudence of her colleague who only ever gives the morally obligatory minimum. Again, we view the agent who acts supererogatorily as having acted well and admirably, not simply from one point of view among many, but *simpliciter*.

It seems to me, then, that the most plausible understandings of supererogation will not treat supererogatory options as on a par with, incomparable with, or less overall choiceworthy than their merely permissible alternatives. Rather, they will endorse all-things-considered supererogation, and regard supererogatory options as more overall choiceworthy than their alternatives, though not all-things-considered required or obligatory. This gives the dominance argument for rejecting supererogationism at least *prima facie* plausibility.

4.3 Rejecting Supererogationism: Cyclicity and Undermining

Revealingly, however, the anti-supererogationist application of dominance reasoning faces a pair of objections that closely parallel the objections to Ross's anti-nihilistic argument discussed in §4.1.

4.3.1 Cyclicity and Supererogation

MacAskill's cyclicity worry, remember, arises insofar as nihilistic theories are construed as treating any two options as *incomparable* in terms of objective choiceworthiness. An analogous worry arises for the anti-supererogationist argument insofar as supererogation is construed as a matter of either incomparability or parity between supererogatory and merely permissible options. Suppose, for instance, I have some credence in a theory according to which the option of donating $500 to GiveDirectly is neither more, nor less, nor equally as choiceworthy as the option of buying a new TV (i.e., the two options are either incomparable or on a par) and some credence in a theory according to which I am obligated to donate to GiveDirectly. Then the principles GDoT and GDoT* will be inapplicable and will not generate a dominance argument for donating to GiveDirectly. The principle DoT will generate such an argument, but at the cost of preference cycles, as we saw in §4.1.

How much does a restriction to acyclic principles like GDoT* limit the anti-supererogationist argument? One the one hand, we have just seen reasons why the problematic theories of supererogation as incomparability or parity are implausible. On the other hand, to whatever extent these versions of supererogationism do seem plausible to a particular agent, there is no easy way of saving the dominance argument against supererogation as there is with nihilism. A normative theory that regards the options of buying a new TV and donating to GiveDirectly as all-things-considered incomparable does not treat these options as equally choiceworthy even in the "negative" sense of equality described in §4.1.1, the sense in which nihilistic theories *do* necessarily regard any two options as equally choiceworthy. Thus, if an agent does have positive credence that these options are overall incomparable or on a par, dominance reasoning alone cannot generate a rational requirement to perform the morally safe option of giving to charity, on pain of inviting cyclicity.

4.3.2 Undermining and Supererogation

But just as the greatest worry for Ross's anti-nihilistic argument is that certain forms of nihilism will undermine the very dominance principles on which it depends, so the greatest worry about the anti-supererogationist argument is that certain forms of supererogationism will undermine those same principles—albeit more subtly than radical skeptical theories like DNET.

One way this might happen is if the same arguments that support the existence of first-order supererogation—in particular, arguments from demanding-ness—can be simply re-run as objections to dominance principles like GDoT*: that is, GDoT* should be rejected, just like maximizing utilitarianism, because the demands it places on moral agents are excessive. But this line of argument is unconvincing for at least two reasons. First, even if one thinks it plausible that principles of morality must conform to a requirement of undemandingness, it is much harder to see why principles of *rationality* like GDoT* should conform to such a requirement. Morality, perhaps, is a human practice that must be designed to accommodate human weakness and imperfection. But the canons of rationality derive from more abstract standards of consistency or coherence that are not obviously sensitive to such pragmatic concerns. Second and more importantly, appeals to arguments for first-order supererogationism are question-begging, or something very much like it, when our goal is to formulate principles for choice under conditions of first-order normative uncertainty, where one is *inter alia* uncertain about the soundness of those very arguments. That is to say, the conditional claim that a "second-order" theory of rational choice under moral uncertainty should be rejected if it is highly demanding is much less plausible than the conditional claim that a "first-order" moral theory should be rejected if it is highly demanding, because the former conditional would generate illicit "upward dependence" of principles for choice under moral uncertainty on the very first-order moral questions we are uncertain about (cf. §3.6).

But there is another, more difficult way in which the undermining worry can arise. To see it, we must introduce a further distinction among conceptions of supererogation, at a somewhat higher level of generality than the distinctions discussed in the last section. Suppose we agree, as I have argued, that supererogation in most plausibly conceived in such a way that there is in some sense greater all-things-considered objective reason to choose a supererogatory option O than a merely permissible alternative P. It seems to me that there are then two very general ways of making sense of supererogation, such that P can be understood as all-things-considered rationally permissible at least by the lights of the supererogationist theory (that is, for an agent whose credence in the supererogationist theory is 1).

First, suppose one accepts the following two intuitively appealing principles: (i) an agent is always rationally required to do what she has most subjective reason to do and (ii) an agent's subjective reasons derive from her beliefs about her objective reasons—more specifically, that belief in an objective reason of a given strength gives rise to a subjective reason of proportionate strength. Then the most natural way of understanding supererogation is to suppose that, although there is more objective reason *available* to support the supererogatory option O, some of the reasons that favor O are of a sort that the agent has the power to set aside or eliminate: "justifying" or "enticing" reasons, imperfect reasons, or reasons that come along with exclusionary permissions. That is, the agent is permitted to perform option P, according to the supererogatory theory, because she has the power to alter the initial balance of objective reasons by setting aside certain optional reasons, thereby altering the balance of subjective reasons as well, in favor of P.[16] Since this approach is exemplified by the idea of exclusionary permissions, let's call it the "exclusionary conception of supererogation" (ECS).

The alternative, it seems to me, is to deny (i) and hold that supererogation arises because, in a certain class of cases, agents are rationally permitted to act against the balance of subjective reasons. That is, one is rationally permitted to choose P even though one has more subjective reason to choose O. Since this approach is exemplified by understandings of supererogation in terms of a satisficing conception of subjective rationality, let's call it the "satisficing conception of supererogation" (SCS).

Just as DNET posed an undermining threat to the dominance argument for rejecting nihilism, so SCS poses an undermining threat to the dominance argument for rejecting supererogationism.[17] To make this threat concrete, suppose that I have .6 credence in a version of SCS that implies the following principle.

Egoistic Permissions Principle (EPP) An agent is always rationally permitted to choose the option that is prudentially best in expectation, out of those options in a given choice situation that violate no moral constraint (e.g. against

[16] Alternatively, one might deny the objective-subjective linkage principle (ii), and hold instead that agents have the power, not to *eliminate* certain objective reasons, but rather simply to prevent them from generating subjective reasons. As far as I can see, everything to be said below about the view described in the main text will apply to this variant as well.

The idea that we can sometimes alter our reasons by an act of will is plausibilified by other sorts of cases where this seems to happen: for instance, when an agent forms or abandons a plan or life project, she may be thought of as creating or eliminating reasons for herself. Supererogatory moral reasons, on this conception, might be thought of as reasons "out there" in the environment that an agent can, but need not, adopt as her own.

[17] Thanks to Shelly Kagan for bringing this problem to my attention.

> killing the innocent, breaking a promise, etc), even when there is greater all-things-considered subjective reason for her to choose some other option.

Just as in §4.1.2 we saw that DNET $\vdash \neg$GDoT*, so here EPP $\vdash \neg$GDoT*. For consider a situation in which I can decide to spend some money harmlessly on myself or donate that money to charity, and in which I divide my beliefs between a moral theory on which donating the money is obligatory and one on which donating the money is supererogatory. In this situation, GDoT* implies that I am rationally required to donate the money, while EPP implies that I am not. Just as with DNET, then, it would seem that if I accept EPP with subjective probability p, I must assign GDoT* a degree of belief less than or equal to $(1 - p)$; that if $p > .5$, then Cr(GDoT*) $< .5$, and that if I *believe* EPP, I must disbelieve GDoT*. So again we may ask: How can I be subject to the rational requirements of a principle, GDoT*, that I (justifiably, we may stipulate) regard as more likely false than true?

This undermining worry can be mitigated in several ways. First, of course, there remains the option of taking a principle like GDoT*, or some stronger principle from which it derives, to have belief-independent rational requiring force. Thus we might say that, even if I justifiably believe that I am not rationally required to perform the potentially-obligatory option O because I have greater than .5 credence in a principle like EPP, I can nevertheless be so required.

But moreover, there is reason to think that the versions of SCS that undermine dominance principles like GDoT* are uniquely implausible, indeed more implausible than undermining forms of nihilism like DNET. It strikes me as a basic conceptual truth that I am rationally required to do what I have most subjective reason to do. To think otherwise is to multiply basic normative concepts beyond necessity, allowing the concept of rational requirement to float free of the concept of subjective reasons. And as I suggested above, the sort of demandingness arguments that might support a satisficing conception of rationality in the context of potentially-supererogatory acts seems much more plausible in the domain of first-order normative ethics than in the domain of rationality.[18] It is much more plausible, then, to posit the existence of certain objective reasons that I have the power to set aside or exclude from consideration such that they do not generate subjective reasons in the first place than to hold that I can simply ignore the balance of subjective reasons I do have.

Nevertheless, if we deny that a principle like GDoT* has belief-independent normative force, then SCS does pose an undermining threat to the dominance argument for rejecting supererogationism, for an agent who assigns the majority of her credence to this version of supererogationism. But what about ECS, the rival conception of supererogation according to which an agent can *alter* the initial balance of reasons so that she no longer has all-things-considered subjective reason to choose the supererogatory option? This seems to me the most plausible way of understanding supererogation, but if so, it raises an interesting complication for the dominance argument against supererogationism. It seems plausible to suppose, on this conception, that a fully rational agent exercises her power to alter the balance

[18] Another sort of argument for a satisficing conception of rationality invokes choice situations in which there are infinitely many options, none of which is better than all the rest (Landesman, 1995). But it is hard to see how a satisficing view motivated by cases of this kind could support a theory of supererogation.

of reasons in favor of the merely permissible option iff she actually *chooses* the merely permissible option: that is to say, it seems at least irrational if not simply inconceivable for an agent who has the power to alter the balance of reasons in favor of a merely-permissible option to choose that option but fail to alter the balance of reasons in its favor, or to so alter the balance of reasons (setting certain moral reasons aside) but then choose the supererogatory option anyway. Indeed, by what means could an agent alter the balance of reasons to favor the merely-permissible option, other than by *making a rational choice of that option*?

If this is so, then ECS treats cases of supererogation as exhibiting a kind of extreme act-state dependence: An agent A who is certain of ECS, and certain as she chooses between options O and P that O is supererogatory and P is merely-permissible, will be certain that whichever option she chooses will be the one favored by the balance of reasons. If she is exactly .6 confident that she will choose P, then she is exactly .6 confident that the balance of reasons will, in the end, turn out to favor P.[19] So, if A is not certain that she will choose O, then she is not certain that the balance of reasons at the time she makes her choice will not favor P, and so it looks as if dominance reasoning cannot serve to eliminate P from consideration.

But there is something intuitively wrong with this line of reasoning: For although A is not certain that her reasons to choose O will turn out to be equal to or greater than her reasons to choose P, she is certain that *if she chooses* O, then there will be equal or greater reason to choose O—indeed, she is specifically certain that

[19] I assume that deliberation does not "crowd out prediction," i.e. that an agent can have credences concerning what her own choice will be even as she deliberates about that choice. For defense of this view, see for instance Joyce (2002).

there will be greater reason. If she is certain that her supererogationist theory is right, then she has the same certainty about P, i.e. that conditional on her choosing P, P is the option favored by the balance of reasons. But if she has some degree of belief in a more rigorous moral theory according to which O is obligatory, i.e. on which she lacks the power to alter the balance of reasons in favor of P, then she can no longer have this certainty that if she chooses P, P will be the option most supported by her reasons. This creates an asymmetry in favor of O that seems to make O the most rational option.

And indeed, a slight reframing of the dominance principle GDoT* that accounts for the problem of act-state dependence allows us to accommodate this intuition. The fully precise way of expressing the idea of dominance reasoning over moral theories, it seems to me, is as follows.

Final Dominance over Theories (FDoT) If (i) every theory in which an agent A has positive credence implies that, conditional on her choosing option O, she has at least as much reason to choose option O as to choose option P, (ii) one or more theories in which she has positive credence imply that, conditional on A's choosing option O, she has greater reason to choose option O than option P, and (iii) one or more theories imply that, conditional on A's choosing option P, she has greater reason to choose option O than option P, then A is rationally prohibited from choosing P.

If a principle like FDoT that allows dominance reasoning to account for act-state dependence is correct, then dominance reasoning will succeed against supererogationist theories of the ECS variety. This is, of course, a heavily conditioned conclusion. Nonetheless, FDoT is intuitively plausible (or, at least, formalizes an intuitively plausible form of reasoning) and there is reason to think that ECS is more plausible than rival versions of supererogation—both those that deny that the balance of reasons, conditional on A's choosing O, is favorable to O and those according to which it is rationally permissible for an agent to disregard the balance of subjective reasons. Thus, even this limited conclusion may have substantial practical scope.[20]

4.4 Practical Upshots

To the extent that the dominance argument for rejecting supererogationism is effective, it sheds new light on some contested practical questions concerning the ethics of philanthropy. Suppose you are a typical member of the middle class in the developed world, and that on careful examination of the available evidence you judge that donating to well-chosen charities can save a life in expectation at a cost on the order of a few thousand dollars.[21] The upshot of a dominance principle like FDoT seems to be that even if you judge such charitable giving to be most likely morally supererogatory, you are still rationally required to donate a substantial portion of

[20] Note that to avoid the undermining problem posed by SCS it is only required *either* that a dominance principle like GDoT* or FDoT have external normative force *or* that the agent give the SCS version of supererogationism less than half her credence. The versions of supererogationism on which non-moral reasons are stronger than or incomparable with moral reasons pose a more substantial threat to the dominance strategy, insofar as *any* positive credence in these theories will block the application of GDoT* or FDoT. We will address this issue in the concluding section of the chapter.

[21] Figures in this range are well-supported by recent research. See for example GiveWell's estimate of the cost-effectiveness of anti-malarial bed net distributions (GiveWell, 2013).

your income to charity. More specifically, as per FDoT, you are rationally required to give at least to the point at which you judge that, by giving more, you *might* no longer be acting for the best, all things considered. In contrast to Peter Singer's famous injunction to "give to the point of marginal utility" (Singer, 1972), we might say that you are required to give (at least) "to the point of all-things-considered uncertainty." If you are not yet at this point, then you do not risk acting wrongly (contrary to the overall balance of reasons) by giving another dollar to charity. But there is a risk of acting wrongly if you keep the dollar, insofar as you judge that you might be objectively obligated to donate it. You may believe, with great confidence, that your donation is objectively supererogatory, but there is no (overall) harm done by performing a supererogatory action on the grounds that it might be obligatory, while there is (overall) harm done by refraining from an objectively obligatory action on the grounds that it might be supererogatory. The latter is a risk to be avoided, the former is not.

Giving to the point of all-things-considered uncertainty will generally be quite a bit less demanding than giving to the point of marginal utility. But just where is the point of overall uncertainty, when it comes to philanthropy? Here intuitions will no doubt diverge. As I have already suggested, that point can be reached more quickly when competing moral considerations are in play: moving your family into a hovel in order to squeeze more deworming money out of your Wall Street paycheck might well be an overall mistake. Of course, this risk of wrongdoing against one's family might be outweighed by the risks of not giving to the point of marginal utility, but it might not—dominance reasoning alone cannot speak to this question. But it should be readily conceded that there is a point, short of the point of marginal utility, at which potentially countervailing obligations or non-moral reasons come into play and the dominance argument can go no further. Still, common sense suggests that few First Worlders are presently in any danger of stepping past this point. Is there a significant risk that a typical American would violate a moral obligation, or otherwise act against the overall balance of reasons, by deciding that she and her family should live on $50,000 rather than $60,000 per year, if by this sacrifice they could save two or three lives every year in expectation? Even here, perhaps, intuitions may diverge. But I imagine no one will think that an agent in this position who donated, say, $1000 a year to charity would be acting for the worse all things considered. And so by the above reasoning, this much at least is rationally required.

More clearly, perhaps, the dominance argument holds implications for *where* one's charitable efforts should be directed. Suppose that a billionaire philanthropist every year makes a $1,000,000 donation to the local art museum. One year, an admirable friend suggests to her that this money would be better spent supporting some life-saving health initiative in the developing world. Here, I don't think reasonable intuitions can diverge. It seems simply obvious that, between the options of donating a million dollars to the art museum and donating a million dollars to the health initiative, there is more all-things-considered reason to choose the latter, and the former option can be rationally permissible only if the latter is supererogatory. Put another way, it is highly implausible that our imagined agent would violate a moral obligation or otherwise act for the worse by choosing to support the health initiative instead of the museum. In so doing, she might deprive the local museum-going public of the chance to feign enjoyment of one more Jackson Pollock canvas, but it is quite hard to believe that her obligations in this respect outweigh her obligations to the global poor.

Here, then, is a straightforward practical conclusion: When an agent is confident that it is all-things-considered permissible for her to devote certain resources (time, money, etc) to philanthropic endeavors, and has no reason to think herself under any other obligation to support some specific philanthropic endeavor (e.g. by having *promised* the money to the art museum), the possible obligation to support *efficient* endeavors—those that do the most good per dollar spent—is enough to generate a rational requirement that she choose these efficient endeavors over their less efficient alternatives.

Rejecting supererogationism may have practical consequences in other domains as well. For instance, an agent might be fully convinced that non-human animals have enough moral standing that vegetarianism is at least as choiceworthy, all things considered, as meat consumption, and yet be unsure whether vegetarianism is morally required or supererogatory. Likewise, a citizen of a democratic society might be certain that she has at least as much overall reason to vote as to stay home on election day, and yet be unsure whether voting is morally required or supererogatory. For such agents, the argument I have been defending implies that they are rationally required to choose the morally safe option (vegetarianism or voting, respectively).

A common theme among the consequences of rejecting supererogationism—

favoring altruism and philanthropy in general, efficient philanthropy in particular, and increased regard for potential moral subjects like non-human animals—is their broadly utilitarian character. The effect of rejecting supererogationism is largely pro-utilitarian, I suspect, because utilitarian theories are the only plausible moral theories that are *totalizing*, in the sense of rendering every or nearly every choice a matter of moral obligation. Thus, wherever a potentially obligatory action fails to maximize utility, it is potentially forbidden, i.e., there is a possible contrary objective obligation to counterbalance it (even if this possible obligation is outweighed in the last analysis)—so for any agent who has positive credence in classical utilitarianism, dominance reasoning cannot imply a rational requirement to choose any non-utility-maximizing course of action. By contrast, we often face choices where the only potential moral obligation one way or the other is that implied by a maximizing utilitarian theory, and in these cases dominance reasoning suggests that we are often rationally required to meet the utilitarian demand.

4.5 Can Dominance Reasoning Stand on Its Own?

There is a final, very general problem for dominance approaches that we have so far put off, and that serves to motivate the more ambitious project of the next two chapters. The problem is this: Any form of dominance reasoning requires a kind of certainty, namely, certainty that some dominant option is *at least as good as* a dominated option. But on the Bayesian understanding of probabilistic belief we have so far assumed, which accepts regularity, such certainties are disallowed except for some very narrow subclass of *a priori* truths (e.g. logical or analytic). No matter how confident we may be of the judgment that it is at least as good to give to charity as to spend money on oneself, or that it is at least as good not to torture babies for fun as to torture babies for fun, it is not generally believed that these are logical, analytic, or indeed even *a priori* truths.[22]

Of course, one could defend dominance reasoning by disputing the skeptical Bayesian attitude toward certainty. Many, no doubt, will be inclined to say it *is* certain that it's at least as good not to torture babies for fun as to torture babies for fun. But it seems to me that the Bayesian view is correct in this regard, though it is well beyond the scope of my current project to defend it. To put the point as intuitively as possible, we should remain open to the possibility that our ordinary beliefs are radically wrong, even diametrically opposed to the truth, in the moral domain as in the empirical domain. Even if, as some philosophers have held, it is not conceptually possible that causing pain to innocent people is good for its own sake or that knowingly punishing the innocent is just, the difficulties and uncertainties of conceptual analysis mean that these things remain both *epistemically* and *logically*

[22] This draws attention to another point we have passed over, though I believe justifiably. The mundane reason why it cannot be true *a priori* that, say, giving to GiveDirectly is better than buying yourself a new TV is that the moral comparison of these options depends crucially on their consequences, which are of course not an *a priori* matter. We have been assuming thus far that, in the sort of dominance reasoning we are interested in, these empirical uncertainties are suppressed. That is, the dominance principles we have considered say roughly that if each *moral* theory to which you give positive credence implies that your empirical-belief-relative reasons either favor option O over option P or are neutral between the two, then you should prefer option O. It seems permissible to reason in this way, rather than attempt the fruitless task of applying dominance principles over all maximal conjunctions of moral and non-moral possibilities, because particular theories will have their own internal methods for handling empirical uncertainty which are often much stronger than dominance reasoning. Thus, from the standpoint of an agent deliberating about what to do given her *moral* uncertainties, she can simply take it as input, for instance, that utilitarianism assigns greater *expected* utility to option O than to option P, suppressing any further consideration of the empirical uncertainties that go into the utilitarian assessment. (I will defend this sort of "multi-stage" reasoning under moral uncertainty at greater length in §6.2.)

possible. Therefore I think we should accept, regretfully, that it is never strictly certain that one option is at least as good as another.

Does this mean that dominance reasoning is useless (indeed, not just with respect to moral uncertainty, but in its more common applications to cases of empirical uncertainty as well)? No, but it does mean that the practical force of dominance arguments is derivative rather than fundamental. That is to say, in cases where we should act on dominance reasoning, it is because it *approximates* some other form of reasoning, e.g. expected value reasoning, in the limiting case where any possibility according to which option P is better than option O is seen as highly improbable. The more unlikely it is that spending money on luxuries for oneself is better overall than donating the money to charity, the more unlikely it is that the correct comprehensive theory of decision-making under uncertainty will regard the former option as rationally required or permissible. Dominance reasoning is useful, therefore, in that it identifies *robust* practical conclusions, conclusions in which we can have a great deal of confidence even while we remain deeply uncertain about the fundamental principles of rational choice and about what they will imply for more difficult cases. More specifically, for our purposes, we can say this: So long as *some* hedging view of moral uncertainty turns out to be correct (or more broadly, some view consistent with SSR), it will almost certainly imply that we are rationally required to do things like donate a substantial portion of our incomes to highly effective charities, and that having high credence in at least the ordinary forms of moral skepticism and nihilism do not make it rationally permissible to act like an amoralist.[23]

[23] In conversation, Samuel Kerstein has suggested that many agents have substantial credence

In the next chapter we will begin to develop such a comprehensive theory of rational choice under moral uncertainty, that attempts to specify precisely the contributions that an agent's moral beliefs and credences make toward determining what she subjectively ought to do. In so doing, we will allow ourselves to say something about the more difficult cases where plausible moral views conflict, prescribing incompatible courses of action as morally required.

in views like ethical egoism according to which any philanthropic self-sacrifice is worse overall that self-interested alternatives, and that for this reason such agents will never be close enough to certainty that an option O is either obligatory or supererogatory for dominance reasoning to serve as a plausible stand-in for more general principles of rational choice. I accept this possibility in principle, though it seems to me that many or most agents have very little credence in any form of ethical egoism and that such a credal state is at least rationally justifiable. But it is also worth noting that an agent's credence in such views should not problematize the conclusions drawn in the preceding section about *where* she should direct any philanthropic giving she chooses to undertake. For an agent choosing between similarly altruistic options (i.e., options that involve similar degrees of self-sacrifice) like donating \$1 million to the art museum or donating \$1 million to global health charities, egoistic theories will agree that the latter option is at least as choiceworthy as the former.

Still, whether and how often we are rationally required to choose potentially supererogatory options will depend on the more general theory of rational choice under moral uncertainty. Even if dominance reasoning is often not an acceptably close approximation of this theory, limiting the scope of the argument I have made in this chapter to a small number of cases, it may still turn out that we are rationally required to choose a great many potentially-supererogatory options in the last analysis, since the potential all-things-considered downside of such actions (e.g., if ethical egoism is true) is outweighed by the potential upside (if a non-egoistic theory is true).

Chapter 5: Intertheoretic Value Aggregation

As we have seen, drawing interesting normative conclusions about choice under moral uncertainty requires that we move beyond dominance principles and offer a general theory that accounts for cases of *conflict* between epistemically possible moral principles. In this chapter, we will begin the task of describing such a general theory. In the process, we will be able to address two more objections to moral hedging that we have thus far put off: the *problem of intertheoretic value aggregation* (PIVA) and the *grounding problem.* I will argue that both these objections can be overcome if we can locate the basis for a particular aggregation of the value scales of rival theories, not in some universal, freestanding second-order principle like Lockhart's PEMT, but rather in the shared structure and content of limited classes of epistemically possible moral theories, what I will call *comparability classes.*[1]

5.1 Two More Challenges to Hedging

We have already encountered the problem of intertheoretic value aggregation in earlier chapters, but let's now take the time to describe it in a bit more detail, and to make clear its intuitive force. In the process, we will see that it in fact represents

[1] Much of this chapter, particularly §§5.1, 5.3, and 5.4, is based on Tarsney (forthcoming a).

two problems, one epistemic and the other metaphysical, each of which must be addressed by any plausible account of intertheoretic aggregation.[2]

Suppose, then, that an agent who divides her beliefs evenly between Kantianism and utilitarianism (Cr(Kantianism) = .5, Cr(utilitarianism) = .5) is faced with a moral dilemma about which these theories disagree: Should she push a large man off a footbridge so that he falls in the path of an out-of-control trolley, stopping it before it can hit and kill some number $n > 1$ of innocent people? Or should she do nothing, letting the large man live and the n others die (Thomson, 1976)?

If one accepts the core principle of moral hedging, that an agent's subjective reasons for action under moral uncertainty are sensitive both to her degrees of belief in rival moral theories and to the strength of the objective reasons those theories posit, then it seems one will judge that what the agent in this case subjectively ought to do depends on the value of n. Perhaps if killing the large man will save only two others from the trolley (or only one other whose life happens to carry slightly more expected utility, because he is younger, or healthier, or of a more cheerful disposition, or valued by a slightly larger circle of friends and family), the epistemically possible Kantian prohibition on killing innocents as a means to an end

[2] What I will call the "problem of intertheoretic value aggregation," the literature to date has typically called the "problem of intertheoretic value *comparisons*." I choose the non-standard terminology because, as we will see in the next chapter (specifically §6.4), even when we are able to *compare* the value scales of two or more moral theories, a further question remains about how to *aggregate* those value scales into a single scale for purposes of rational choice. Thus, in my terminology, the problem of intertheoretic value comparisons is a part, albeit a very large part, of the broader problem of intertheoretic value aggregation.

The next chapter will also introduce another division between three kinds of barriers to intertheoretic value aggregation (two of which have to do with intertheoretic comparisons while the third has to do with the further problem of aggregation). This division is orthogonal to the epistemic/metaphysical distinction introduced in the current section, i.e., each of the three challenges to be described in the next chapter will have both an epistemic and a metaphysical dimension.

ought to win out over the equally possible but relatively weak utilitarian reasons for killing. On the other hand, if the trolley is about to crash headlong into a reactor of the local nuclear power plant, triggering a catastrophic meltdown that will cause 10,000 nearby residents to die agonizing deaths from radiation sickness, then perhaps the now-much-strengthened utilitarian reasons should take precedence.

So far so plausible. But what about the intermediate cases between these extremes? A commitment to hedging seems to entail that there is some value of n that is the tipping point, or the point of indifference, between Kantian and utilitarian considerations in this case as we have described it. Perhaps if pushing the large man will result in 27.38 or more net lives saved in expectation, then our agent subjectively ought to choose the utilitarian option, but if the expected number of lives saved is any less than this, she ought to choose the Kantian option.[3] This value of n, whatever it might be, represents the ratio between the wrongness of killing an innocent person as a means to an end, according to Kantianism, and the rightness of saving a (net) life, according to utilitarianism.

There is something unpalatable about the idea that such a ratio exists. In fact, there are at least two things unpalatable about it: (1) What could possibly *make it the case that* any particular value of n is the true tipping point between Kantian

[3] Of course, we have to fill in details before such a conclusion could be entailed, in particular the precise quantities of utility and disutility that would follow from each death caused or averted.

As may already be clear, the general principle of moral hedging as I have described it does not fully commit us to the existence of such a sharp cutoff; e.g., we might believe in hedging while also believing that Kantian and utilitarian reasons are only roughly comparable and hence that there is no definite point of indifference between them in the sort of case we have described. Nevertheless, it seems that even such a view will commit us to sharp boundaries of *some* sort, e.g., between the values of n for which the Kantian or utilitarian reasons are determinately stronger and the intermediate interval of incommensurability (Broome, 1997, pp. 85-86).

and utilitarian considerations, in the above case as we have described it? What facts about the (empirical or moral) universe could serve to *ground* the existence of such a ratio? (2) Assuming that there is such a ratio, how could we possibly *discover* it? How is a rational agent to figure out, for instance, whether the right tipping point when she is faced with such a choice is 3, or 5, or 27.38, or 12π, or ...? And if there *is* no way of figuring it out other than to take a plausible-sounding guess, then what becomes of the point that we need norms of rational requirement responsive to an agent's moral uncertainties precisely because we as agents need action-guiding norms whose prescriptions are *epistemically accessible to us*? If a commitment to hedging forces us to simply take a shot in the dark with respect to the exchange rate between Kantian and utilitarian values, what has been gained over the simpler option of taking a shot in the dark at which is the true moral theory and being guided by whichever theory we regard as more probable?

It is useful, I think, to consider these as separate problems. I will call the first, metaphysical worry the *grounding problem*, reserving *problem of intertheoretic value aggregation* for the second, epistemological problem. The former problem has been suggested by several authors in the literature, but to the best of my knowledge has not been explicitly distinguished from the epistemological problem.[4] But it is,

[4] I take this worry about grounding to be, for example, at least one of the lines of argument suggested in the following passage from Hudson (1989): "Even mere axiological uncertainty within an unquestioned subjective consequentialist framework is unhedgeable. Suppose the agent assigns probability 0.6 to the view that pleasure-minus-pain is the only intrinsic good, and 0.4 to the view that the good is self-realization. And suppose she must choose between an act that produces ten hedons and two reals and one that produces nine hedons and thirty reals. ('Reals' are the units in which self-realization is measured.) Which act should she do? The two axiological theories lead to different answers. Since the hedonic theory is more probable, perhaps she should accept its answer. But the self-realization theory seems to find more of a difference between the two actions, and perhaps this should outweigh its slightly lesser probability. But wait—is a difference of twenty-eight reals really greater than a difference of one hedon? What is the common measure between

in my view, at least as urgent as the epistemological problem, and it cannot be assumed that a solution to the latter will also be a solution to the former. Ross (2006), for example, suggests that we can find out how to compare the value scales of rival theories by reasoning backwards from our considered judgments about the subjective rationality of particular choices given particular credal states in imagined cases like the footbridge dilemma described above. But even if one regards our judgments about such cases as *prima facie* trustworthy, such an approach leaves it mysterious *how* or *in virtue of what* the underlying value comparisons that we suss out of those case judgments could be correct. Of course, we always have the option of asserting it as a brute fact that, say, the correct ratio between the Kantian wrongness of killing and the utilitarian rightness of saving lives is 27.38 (or, more plausibly, that there is some such precise ratio that is somewhere in such-and-such range). But it would be strange if facts like these were simply brute, part of the fabric of the universe on a par with fundamental physical constants. Why should the

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

Likewise, Gracely writes: "Consider again the extra-marital sex example. It certaintly seems possible to compare the fact that one system decalres the act to be a terrible wrong, while the other merely regards it as tolerable, or, at most, slightly more right than wrong. I don't agree. 'Terrible wrong' is only meaningful within the overall framework of that system. There is no abstract scale of 'wrongness' outside of the rank provided *within* a theory.

"Moreover, if indeed there is only one true moral theory, the noncomparabilty argument is further strengthened. In that case, a metatheory that attempted to design a weighting scheme applicable across systems would necessarily be comparing many false, or partly false, systems to each other and/or to the one that is true. Why would one expect there to be a valid method of comparing magnitudes other than that provided by the one true theory?" (Gracely, 1996, p. 331)

Finally, Nissan-Rozen (2015) raises a more specific version of the grounding objection to hedging, premised on the claim that a theory's "moral value function" is simply a representation of its ranking of lotteries (i.e., choices made under various states of empirical uncertainty), and that since these representations are unique only up to affine transformation (meaning that the choice of scale and zero point are arbitrary), there is no non-arbitrary or meaningful basis to compare the value function of one theory to that of another. (I address this argument in Appendix B, §B.2.)

universe have been so obliging as to provide us with the presumably thousands of such constants needed to compare the plethora of possible moral theories to which a reasonable agent might subscribe, especially given that all but one of those theories are in fact false?

Having described PIVA and the grounding problem somewhat impressionistically, let's make our terminology a bit more precise.

- An *intertheoretic value aggregation* is a mapping from probability distributions over value assignments to value assignments. The output of this mapping is meant to represent, in one way or another, a probability-weighted aggregate of theories in which an agent has positive credence.

- The *grounding problem* is the problem of what could make a particular intertheoretic value aggregation objectively correct, i.e., correct in a way not relative to any agent's beliefs/evidence.

- The *problem of intertheoretic value aggregation* is the problem of how to *discover* the objectively correct intertheoretic value aggregation and, relatedly, how an agent should decide what intertheoretic value aggregation to adopt, given her beliefs/evidence.[5]

Both the grounding problem and the problem of intertheoretic value aggregation have been put forward repeatedly in the literature as powerful objections to

[5] I group these problems together since I assume that an agent's decision about how to aggregate the value assignments of the various moral theories in which she has positive credence will be guided by a concern for truth, i.e., she will want as far as possible to adopt the *correct* intertheoretic value aggregation. In Chapter 7 we will consider how an agent ought to proceed when she cannot be sure what aggregation procedure is correct.

moral hedging, with particular emphasis being placed on one or both of these problems by Hudson (1989), Gracely (1996), Gustafsson and Torpman (2014), Nissan-Rozen (2015), and Hedden (2016).

The aim of this chapter is to begin addressing both of these problems. The strategy is as follows: I start by exploring two simple case of intertheoretic moral uncertainty that appear to offer rationally perspicuous, non-arbitrary methods for intertheoretic value aggregation. In one of these cases, the natural aggregation procedure involves both quantitative intertheoretic value comparisons and moral hedging. In the other case, it involves neither. Reflection on these cases suggests the more general idea that there are clusters of moral theories, which I will call *comparability classes*, that share common principles strong enough to establish clear internal procedures for value aggregation and choice under moral uncertainty, which sometimes will and sometimes will not take the form of hedging procedures like expected value maximization.

If this is so, I argue, then neither the grounding problem nor PIVA can be taken as *general* objections to moral hedging as such. That is, even if it ultimately turns out that *some* moral uncertainties (like that between classical utilitarianism and absolutist deontology) are too deep to permit the sort of comparison and aggregation necessary for hedging, or indeed any form of aggregation at all, this fact cannot plausibly be taken as a reason to forgo hedging with respect to those uncertainties that *do* permit non-arbitrary intertheoretic comparisons. There is no reason, in other words, to think that "incomparability anywhere is a threat to comparability everywhere." Instead, I will argue, if there is a line to be drawn between uncertainties for which a hedging procedure is appropriate and those for which it is not, that line should be drawn not between moral and empirical uncertainty but rather between those uncertainties that establish barriers to the necessary intertheoretic comparisons and those that do not. In the last section of the chapter, I will contrast this "content-based" approach to intertheoretic value aggregation with the "top-down" approaches that have dominated the literature to date, and offer some general arguments in favor of a content-based approach. Chapter 6 will then develop this approach in considerably more detail.

5.2 Moral and Empirical Uncertainty, Revisited

A theme of the examples of intertheoretic comparability that we will explore in the next section is that the decisions principles offered by a given class of moral theories for dealing with morally relevant *empirical* uncertainty can often be extended, in a natural and rationally perspicuous way, to structurally similar cases of *moral* uncertainty. We have already discussed some of the parallels between empirical and moral uncertainty in Chapter 2. But to prepare the ground for the arguments of the next section, it will be useful to bring out a slightly different point, namely the *continuity* between empirical and moral uncertainty—that is, the difficulty of drawing a clear and non-arbitrary line between the two, of the sort that is needed if we are to judge, with the opponents of hedging, that the principles of rationality treat the two sorts of uncertainty in fundamentally different ways (demanding, e.g., that a utilitarian hedge for her uncertainties about the state of the empirical world, but simply take a shot in the dark with respect to her uncertainties about the *moral* world when she divides her beliefs between several versions of utilitarianism).

The best way to illustrate this continuity is to simply enumerate a few of the cases in which the boundary between empirical and moral questions, and hence between empirical and moral uncertainty, is blurry, and which therefore make it implausible that entirely different sorts of principles should govern the two domains. Here, then, are several such liminal cases.

- Certain cases of uncertainty about moral considerability (or moral status more generally) turn on *metaphysical* uncertainties that resist easy classification as empirical or moral.

  - In the abortion debate, uncertainty about when in the course of development the fetus/infant comes to count as a *person* is neither straightforwardly empirical nor straightforwardly moral. Likewise for uncertainty in Catholic moral theology about the time of ensoulment, the moment between conception and birth at which God endows the fetus with a human soul (though this is perhaps closer to empirical than to moral uncertainty—after all, the soul is *concrete* even though it is not *physical*). Nevertheless, it seems strange to regard these uncertainties as fundamentally different from more clearly empirical uncertainties about the moral status of the developing fetus (e.g., uncertainty about where in the gestation process complex mental activity, self-awareness, or the capacity to experience pain first emerge), or from more clearly moral uncertainties (e.g., uncertainty, given a certainty that the fetus is a person, whether it is permissible to cause the death of such a person when doing so will result in more total happiness and less total suffering).

- Many ethical questions (e.g. about the weight of promises or of individual moral desert) turn in part of questions about the metaphysics of personal identity (Parfit, 1971, 1973). As Moller (2011) points out, one influential argument for the impermissibility of abortion (the "future like ours" argument given in Marquis (1989)) seems much more likely to succeed if one accepts an animalist rather than a psychological theory of personal identity. Uncertainty about the permissibility of abortion, or about whether one is still morally required to adhere to some long-ago childhood promise, that traces back to uncertainty about the criteria for personal identity is hard to describe as "empirical" uncertainty (after all, what empirical observation could resolve the debates between animalists and psychological theorists?), but neither is it uncertainty about basic *moral* principles or moral theory.

- Finally, on the theme of metaphysical uncertainty, a consequentialist (or anyone who recognizes the ethical relevance of consequences) who is also a physicalist about mental states may be uncertain about the moral status of various kinds of entities because she is uncertain what sorts of physical states subvene on morally significant mental states like pleasure/pain, preference, etc. Does a certain artificially intelligent computer program have *preferences*? Does a Martian whose pain-like functional states are realized by D-valves rather than C-fibers feel *pain*? Are various collectivities like corporations, nations, bee colonies, or mobs of human beings with walkie talkies acting out philosophical thought experiments capable of *forming desires*, *pursuing projects*, *willing ends/maxims*, etc? All of these cases likewise blur the line between empirical and moral uncertainty.

- I have already alluded to the possibility of uncertainty about "thick" moral properties, e.g., uncertainty whether some course of action would be *cruel* or *cowardly*. Weatherson, as we saw, is ready to admit that hedging for such thick moral uncertainty may sometimes be appropriate. If so, this cannot be because thick moral uncertainty is just empirical uncertainty is disguise: I might know everything there is to know about my beliefs and motivations in taking a certain action and still wonder whether I acted cruelly. If we admit this sort of uncertainty as hedge-able, then it will be even more difficult to draw a line around the "purely moral" uncertainties we regard as un-hedge-able, since the thickness of moral properties varies more-or-less continuously from *cruel* to *rights-violating* to *non-optimific* (each of which combines evaluative with empirical content).

- Certain views imply that first-order moral principles depend either epistemically or metaphysically on empirical or quasi-empirical facts. For instance, suppose one is convinced that divine command theory is the right account of the source of moral principles, but uncertain what God has commanded, e.g., because one is uncertain whether the Bible or the Quran contains the accurate reporting of God's commands. Here again it is hard to say on which side of a proposed empirical/moral divide one's uncertainty should fall.

- Some ethical views, most obviously rule consequentialism, imply that first-order moral principles depend on empirical counterfactuals: to a rule consequentialist, the true set of moral principles is that which *would* yield the most utility if generally practiced. If I am certain that rule consequentialism is true, but uncertain whether a rule that permits authorities to "punish" the innocent for the sake of deterrence is part of the optimific system of rules, is my uncertainty empirical or moral? This seems to depend on whether one understands rule consequentialism as itself constituting a first-order moral theory (that says "follow the set of principles that would, if generally practiced, maximize utility"), in which case one's uncertainty is empirical, or as constituting a metaethical theory of what *determines* or how we are to *identify* the true first-order moral theory (the content of which is, *de re*, the set of optimific-if-generally-practiced moral principles). But should anything of practical significance turn on such a question of taxonomy?

- Most generally, naturalistic metaethical views that treat normative ethical theorizing as continuous with natural science will see first-order moral principles as at least epistemically if not metaphysically dependent on features of the empirical world. For instance, on Railton's (1986) view, moral value attaches (roughly) to social conditions that are stable with respect to certain kinds of

> feedback mechanisms (like the protest of those who object to their treatment under existing social conditions). What sort(s) of social conditions exhibit this stability, given the relevant background facts about human psychology, is an empirical question. For instance, is a social arrangement in which parents can pass down large advantages to their offspring through inheritance, education, etc, more stable or less stable than one in which the state intervenes extensively to prevent such intergenerational perpetuation of advantage? Someone who accepts a Railtonian metaethic and is therefore uncertain about the first-order normative principles that govern such problems of distributive justice, though on essentially empirical grounds, seems to occupy another sort of liminal space between empirical and moral uncertainty.

In all these cases, it is implausible that important practical questions should turn on whether we classify these uncertainties as "empirical" (and hence, very likely, subject to some form of hedging if not to outright expectational reasoning) or as "moral" (and hence immune from such reasoning). Conversely, it seems plausible that the background principles one accepts for decision-making under empirical uncertainty should have no great difficulty guiding one's actions under these forms of uncertainty, if one allows them to do so. Attention to these intermediate cases highlights the fact that the moral principles we accept will themselves often tell us how to proceed under at least certain limited forms of moral uncertainty.[6]

[6] *Contra* Hudson, *inter alia*, who writes: "Any moral theory, in telling the agent what to do, will ignore the agent's possible commitment to other moral theories. And while it should be considered a defect in a theory that it issues instructions that the agent does not know how to follow, it is not similarly a defect if it issues instructions that the agent decides not to follow because she does not believe in the theory." (Hudson, 1989, p. 224)

For instance, suppose you are a hedonistic consequentialist and a physicalist about mental states, and that because you divide your belief among several versions of physicalism, you are uncertain whether some class C of beings (fish, insects, AIs, Martians, bee hives, corporations...) experience pain, even though you have no relevant uncertainties about the chemical composition or functional organization of those beings or any other relevant "narrowly empirical" uncertainty (uncertainty about how exactly the atoms are moving around in the void). You have, say, a credence of .6 in a type-identity theory according to which beings of class C do not experience pain, and a credence of .4 in a functionalist theory according to which they do. Though your uncertainty is not narrowly empirical, nevertheless it seems natural to treat it just as you would treat empirical uncertainty (i.e., just as you would treat a similar uncertainty about the moral considerability that *was* narrowly empirical, e.g., uncertainty whether rabbits experience pain that reduces entirely to uncertainty whether rabbits have C-fibers): You should weight the possible suffering of beings of type C by a factor of .4, reflecting the probability that they experience pain, when calculating the expected utility of your actions.

5.3 Case Studies in Comparability

In this section, we will consider two cases in which the same sort of thing can be said of clearly moral uncertainties as of the liminal uncertainties considered above: namely, that decision procedures for empirical uncertainty can and should be straightforwardly extended to deal with these non-empirical uncertainties.

5.3.1 First Case Study: Pluralistic Consequentialism

Consider, then, an agent Alice whose moral beliefs are as follows: (i) She is certain that consequentialism is true. (ii) She is certain that pleasure and pain have non-derivative value and disvalue, respectively. (iii) She is uncertain whether aesthetic goods (beauty etc) have non-derivative value as well. (iv) She is certain that nothing besides hedonic and aesthetic goods has non-derivative value. (v) She is certain that, when faced with purely empirical uncertainties, she ought to do what maximizes expected value. She therefore divides her beliefs between two moral theories: a monistic consequentialist theory T_1 that values only hedonic goods and a pluralistic theory T_2 that values both hedonic and aesthetic goods. The latter theory, we will stipulate, treats hedonic goods and aesthetic goods as comparable—that is, it establishes some "exchange rate" between these two categories of value-bearer. Since we will have reason to refer to Alice's case several times in this chapter and the next, let's give it a name, and call it the Easy Case.

It seems to me that, in the Easy Case, Alice should experience no great difficulty in hedging for her moral uncertainties. We may stipulate an arbitrary unit of hedonic value, the hedon, and may then define a unit of aesthetic value, the aestheton, as the quantity of aesthetic value that *according to* T_2 is equal in value to a single hedon. Now suppose Alice assigns T_1 probability p and T_2 probability $(1-p)$. In this state, I propose, she can quite naturally calculate the expected value of any practical option O as follows: $\text{EV}(O) = \mathit{hedons}(O) + (1-p)(\mathit{aesthetons}(O))$.

Prima facie, it seems quite plausible that an agent in Alice's doxastic situation rationally ought to choose the option that maximizes this formula. The two theories in which Alice has positive credence agree on the value of hedonic goods—take it as a stipulation, to which we will return shortly, that T_1 and T_2 differ not at all in the grounds they offer for, or anything else they have to say about, the intrinsic value of a hedon. As a result T_2 contains within itself, in the exchange rate it posits between hedonic and aesthetic goods, a procedure for comparing its own evaluation of practical options to those offered by T_1, which are exclusively hedonic.[7,8]

The suggestion from this simple case can be easily generalized. Let an agent distribute her beliefs over any number of pluralistic theories, along with her monistic theory, so long as all those theories agree on the value of hedonic goods. Then we can simply create a unit of value for each theory equal to the quantity of value that it assigns to a single hedon (an empirically measurable unit of pleasure/pain, whose value, we have by stipulation every reason to believe, remains constant across theories), thereby normalizing the value scales of the various competing theories. Then we can simply compute the expected value of an option O across theories $T_1, T_2, ..., T_n$ by something very much like the standard formula for expected value:

[7] For now, I leave these claims at the level of plausible intuition. Appendix B gives an extended explication and defense of the sort of reasoning employed in the last two paragraphs, making use of certain ideas to be introduced in Chapter 6.

[8] The thought that moral certainties, e.g. of the value of hedonic goods, can be used to construct a scale by which uncertain moral values and disvalues can be quantified for purposes of moral hedging is suggested by Sepielli (2009), who refers to this scale as a "background ranking" and shows that background rankings can be used to construct reasonable hedging procedures for limited forms of moral uncertainty, e.g. concerning the permissibility of abortion. Sepielli, however, attempts to infer a shared background ranking between a pair of theories simply from similarities in their value assignments (specifically, identical ratios between the differences in choiceworthiness assigned to some triple of options), which can generate multiple, inconsistent normalizations of the same pair of theories (MacAskill, 2014, pp. 142-5). Ross (2006) has suggested that some theories can be compared in the way that the hedonism/pluralism case illustrates, based on deeper agreement with respect to certain forms of value or kinds of reasons, applying this observation to a case where two theories disagree about the correct theory of rights but are in perfect agreement concerning all other moral considerations (Ross, 2006, pp. 764-5).

$$\mathrm{EV}(O) = \sum_{i=1}^{n} (\mathrm{EV}(O)|T_i)(\mathrm{Cr}(T_i))$$

This formula will be applicable even to an agent who distributes her belief over infinitely many competing theories, corresponding for instance to all possible real-valued exchange rates between hedonic and other forms of value over some bounded interval. In fact, we can generalize even further, dropping the assumption that there is a category of value-bearer common to every theory in which the agent has positive credence. Suppose, for instance, that an agent regards hedonic experience, beauty, and knowledge all as potential non-derivative value-bearers, and that her credences regarding the value of each of these goods are mutually independent, so that she has some credence in theories according to which all, none, or any other combination of the three are non-derivative goods. Comparisons between these goods can nevertheless be made by appealing to the exchange rates given by the various pluralistic theories that value more than one of the goods in question, so long as these exchange rates are consistent across the relevant pluralistic theories. Thus, for instance, the monistic theory according to which only pleasure is a non-derivative good and the monistic theory according to which only knowledge is a non-derivative good are made comparable by the agent's credence in one or more theories according to which both are non-derivative goods.

What I have said so far involves an important presupposition, namely that when two theories agree that some feature of the world is a non-derivative bearer of moral value, those two theories attribute the *same kind* and *same degree* of value to that phenomenon—e.g., that the value of a hedon according to Alice's hedonistic theory is equal to the value of a hedon according to the pluralistic theory that also values aesthetic goods. Clearly, this need not always be the case. For instance, consider an agent Alex who, like Alice, divides his moral belief between two theories, a hedonistic and a pluralistic version of consequentialism. But suppose that Alex also divides his *metaethical* beliefs between a robust moral realism and a fairly anemic anti-realism, and that his credence in hedonistic consequentialism is mostly or entirely conditioned on his credence in robust realism while his credence in pluralism is mostly or entirely conditioned on his credence in anti-realism. (Suppose he inclines toward a hedonistic view on which certain qualia have intrinsic value or disvalue entirely independent of our beliefs, attitudes, etc, which we are morally required to maximize. But if this view turns out to be wrong, he believes, then morality can only consist in the pursuit of whatever we contingently happen to value in some distinctively moral way, which includes pleasure but also knowledge, aesthetic goods, friendship, etc.) In this case, the sort of procedure that I have been describing no longer seems rationally compelling: Although the various moral theories to which Alex assigns positive credence all treat pleasure as a bearer of non-derivative moral value, there is no obvious reason to think that they assign the same *degree* of value to a given unit of pleasure.

But there is no reason to assume that such problems will always arise. An agent may be uncertain about first-order moral questions, like whether anything besides pleasure and pain has non-derivative value, without this reflecting any underlying metaethical uncertainties. An agent who divides her beliefs between various monistic and pluralistic theories might nevertheless be in no doubt as to the nature, basis, or degree of value possessed by some category of goods, like hedonic goods, that all the theories she entertains recognize as non-derivatively valuable.[9] The lack of any uncertainty concerning hedonic value makes it a constant feature of the various theories in which she has positive credence, and allows it to serve as a basis for normalization.[10]

To sum up what I have said so far, then, the suggestion is this: An agent who is certain that some form of (maximizing, agent-neutral) consequentialism is correct, but uncertain—even quite radically uncertain—about what sorts of things have value or disvalue nevertheless can find in the shared (or overlapping) assumptions of the various first-order theories to which she assigns positive credence sufficient basis for rationally perspicuous, non-arbitrary intertheoretic value aggregation. Nothing I have said establishes incontrovertibly that such agent is rationally required to maximize expected value, computed in the sort of way I have described. So far, I have only tried to establish that, contra pessimists about intertheoretic comparisons, there is a reasonable and non-arbitrary way of aggregating the value assignments of rival moral theories in the face of certain kinds of moral uncertainty.

Importantly, in the Easy Case and its close cousins, it seems that *both* PIVA and the grounding problem have easy and natural solutions. Alice, in the Easy

[9] Alternatively, if she does have metaethical uncertainties about the nature of value, these may be probabilistically independent of her first-order moral beliefs. Just as metaethical uncertainties need not pose any problem for an agent maximizing expected utility in the face of empirical uncertainties, so long as the metaethical and empirical uncertainties are probabilistically independent, so those metaethical uncertainties need pose no obstacle to hedging for first-order moral uncertainties, so long as the two are independent.

[10] See Appendix B, particular §B.2 and §B.5 for an extended defense of the claim that this sort of intertheoretic agreement or overlap can provide a basis for normalization between moral theories.

Case, can say quite easily what grounds the procedure she uses to aggregate her hedonistic and pluralistic theories: namely, the structure and content of the theories themselves. As both Hudson (1989) and Gracely (1996) point out, it cannot be the *truth* of these theories (i.e., the moral facts that these theories correctly describe) that grounds the correctness of the aggregation procedure, since at least one and possibly both of these theories will turn out to be false. Rather, what seems to make it rational for Alice to normalize her two theories at the value of a hedon and maximize expected value is her credal state, i.e., the fact that she distributes her moral credence over a pair of moral theories with particular contents, plus the truth of some principle of rationality (like the enkratic principle) that translates her beliefs about objective reasons into subjective reasons and rational requirements.

5.3.2 Second Case Study: Absolutist Deontology

Is the Easy Case an isolated exception with respect to intertheoretic aggregation? There is no use denying that it is an especially favorable case—a case of intertheoretic moral uncertainty that "behaves" helpfully like empirical uncertainty and seems to present only one salient option for normalizing the moral theories in which the agent has positive credence. Nevertheless, I will claim, there is an underlying idea that has the potential to generalize quite widely: namely, that sets of moral theories united by the right common content can be aggregated by procedures grounded in that shared content.

To illustrate this general idea (and more especially, to show that we were not in the last section simply taking advantage of an idiosyncratic feature of consequentialist moral theories), let's consider another case far afield from the first: namely, absolutist deontological theories that categorically prohibit certain kinds of actions like telling a lie, breaking a promise, or killing the innocent. No less than consequentialists, defenders of such theories must find something to say about how agents ought to act under morally relevant empirical uncertainty, e.g. uncertainty whether some act would violate a deontological constraint. If I am absolutely prohibited from breaking my promises, what should I do when I am unsure whether I made some particular commitment as part of a long-ago act of promising? If I am absolutely prohibited from killing the innocent, how am I to assess courses of action that carry some (perhaps moderate to vanishingly small) *risk* of killing the innocent, or when I have good reason to kill (in defense of myself or others) someone who might be either an innocent or a malevolent threat?[11]

Let's imagine, then, a deontologist David who has the following simple view of how to deal with uncertainty: An absolute moral requirement is incumbent on an agent whenever his degree of belief that the conditions of that requirement are

[11] This last possibility is suggested by Jackson and Smith (2006). They describe a case in which a skier is headed down a mountain slope in a direction that will trigger an avalanche, killing ten people, unless you shoot and kill him. You are uncertain whether the skier is ignorant of the danger he poses to the ten and therefore morally innocent, or intentionally trying to bring about their deaths. The deontological theory to which you subscribe permits (or requires) you to shoot the skier in the latter case but prohibits it absolutely in the former.

As Jackson and Smith point out, formulating the antecedents of deontological principles in terms of intentions—e.g. that we are prohibited not from breaking a promise or killing the innocent but from *forming an intention* to break a promise or kill the innocent—does not avoid the need to deal with uncertainty, both because (as even Kant admits) we can be quite radically uncertain about the contents of our own intentions and because it is far from obvious whether and in what cases forming an intention to perform some action that I believe I *may* have promised not to do, or to kill someone who *may* be innocent, should count as impermissibly forming an intention *to break my promise* or *to kill the innocent*.

satisfied is greater than .5. So, for instance, if it is in his power to perform some action that he might or might not have promised to perform, he is absolutely required to perform it so long as he believes it more likely than not that he did so promise; he is absolutely prohibited from uttering a particular sentence if he believes it more likely than not that his intention in doing so would be to deceive; and so on.

This sort of view is open to serious objections, but it seems to me as plausible as anything else deontological absolutists can say about the problem of uncertainty, so let us suppose that this is how some plausible class of deontological theories deals with the problem.[12] Now, just as the injunction to maximize expected value held in common by the consequentialist theories considered in the last section extends easily to encompass intertheoretic moral uncertainty (among theories of the relevant class) as well as empirical uncertainty, so too the probability threshold view of deontological obligations provides ready guidance to an agent like David who divides his beliefs between various deontological theories that have this threshold view in common.

Suppose, for instance, that David must decide whether to kill Thomas in order to prevent Thomas from killing ten innocent people. He is empirically uncertain whether Thomas is an innocent threat or a malevolent aggressor, but he is also

[12] Two prominent responses to Jackson and Smith's (2006) uncertainty-based objection to absolutist deontology, namely Hawley (2008) and Aboodi et al. (2008), both defend versions of this "threshold" proposal. In Tarsney (forthcoming b) (the penultimate draft of which is reproduced as Appendix A of this dissertation), I argue that, given a natural representation of the value assignments of deontological theories, deontologists acting under either empirical or moral uncertainty are *at least* rationally required to avoid options that carry a risk greater than or equal to .5 of violating a deontological moral constraint when there are morally safer options available, since those safer options will stochastically dominate their riskier alternatives. This leaves open the possibility that deontologists are subject to more stringent requirements of moral caution, e.g. on the basis of expectational reasoning.

morally uncertain, dividing his beliefs between one absolutist deontological theory that prohibits the killing of innocent threats in other-defense and another that permits it. David, then, must divide his beliefs among three salient possibilities: (i) Thomas is a non-innocent threat, and may therefore permissibly be killed to save a greater number. (ii) Thomas is an innocent threat, and innocent threats may permissibly be killed. (iii) Thomas is an innocent threat, and innocent threats may not permissibly be killed.

It seems quite natural to say, given that both moral theories to which David assigns positive credence agree on a probability threshold of .5 at which deontological prohibitions subjectively "kick in," that he is subjectively prohibited from killing Thomas iff the probability he assigns to possibility (iii)—that is, the conjunction of the empirical belief that Thomas is an innocent threat and the moral belief that it is absolutely impermissible to kill an innocent threat—is greater than .5. Absent some prior reason for David to treat moral and empirical uncertainty differently in his practical deliberations, the principle he accepts with certainty, that it is subjectively wrong to choose some practical option O whenever O carries a risk greater than .5 of violating an absolute moral constraint, seems to address the former kind of uncertainty as much as the latter.

Again, I have thus far not attempted to give any knockdown positive argument that this is the only rational way for an agent in David's position to deliberate. As in the last section, the point for now is simply that there is a rationally perspicuous, non-arbitrary way of responding to a particular kind of intertheoretic moral uncertainty that takes advantage of common principles shared by some class of moral theories over which an agent's credence is distributed.

5.4 Is Incomparability Anywhere a Threat to Comparability Everywhere?

So far I have suggested that an agent who divides her beliefs among a class of sufficiently similar moral theories can possess a rational basis for intertheoretic value aggregation and hence for moral hedging. But I have not addressed the (doubtless more realistic) case of an agent who assigns positive degree of belief to a much more diverse class of theories that share no obvious features to ground intertheoretic aggregation—e.g., an agent who assigns positive credence to both maximizing consequentialist theories and absolutist deontological theories.

In the next chapter we will discuss some possible approaches to these "hard cases" of intertheoretic value comparison. Some of these approaches will involve a retreat from the ambition of full-blooded quantitative comparison between any two moral theories, no matter how unlike, while the approaches that attempt to sustain this ambition will remain tentative and underspecified at best. Absent a solution to the problem of intertheoretic value aggregation in its full generality, therefore, a pessimist may argue as follows: "The principle of hedging under moral uncertainty requires that agents be able to compare and aggregate degrees of value across theories to which they assign positive degrees of belief, in general. Even if there are some easy cases where it looks like such aggregation is possible, as long as the hard cases remain unresolved, the problem of intertheoretic value aggregation as a whole is unresolved and so still poses a decisive obstacle to the principle of intertheoretic moral hedging."

My response is simply this: Someone who opposes any form of hedging for moral uncertainty, but shares the ordinary view that agents should be responsive to empirical risks and uncertainties, must justify drawing a line between the moral and the empirical. What cases like those of Alice and David show is that PIVA cannot justify drawing such a line. Rather, absent some *other* argument that agents should not be responsive to moral uncertainties (like those considered in Chapter 3), problems of incomparability suggest the need for special principles of rationality to handle *incomparability*, rather than special principles (or a special absence of principles) for responding to moral as opposed to empirical uncertainty.

To put it slightly differently, the existence of plausible procedures for intertheoretic value comparison and aggregation in certain cases of moral uncertainty, taken together with the various forms in *in*comparability that can arise in the absence of any moral (or other normative) uncertainty, suggests that the problem of incomparability is simply orthogonal to the distinction between moral and empirical uncertainty. Just as intertheoretic value comparisons do not *always* generate (even apparent) incomparability, so (apparent) incomparability of normative considerations does not arise only in the context of intertheoretic comparisons. Many philosophers, for instance, have held that certain pairs of moral values are genuinely incomparable—that is to say, they have advocated first-order moral theories that generate *intra*theoretic incomparability. An agent might believe, say, that patriotic and familial obligations are genuinely incomparable, such that when these values come into conflict (as when she feels called to volunteer for a war of national self-defense, but cannot do so without abandoning an ailing relative), there is no uniquely rational resolution. Incomparability may arise in more mundane, non-moral contexts as well—for instance, if I am shopping for home decor and must choose between satisfying my own aesthetic sensibilities and pleasing my spouse, whose tastes I regard as gauche.[13] Finally, even on a simple utilitarian theory of value that does not admit this sort of incomparability, purely empirical uncertainty can give rise to a different species of incomparability, as it does in the Pasadena Game (Nover and Hájek, 2004), for which the expected values of playing and not playing appear to be incomparable.

But it would be absurd, on the basis of such cases, to deny the rationality of quantitative value comparisons in general, to claim that the existence of one or more of these forms of incomparability shows that the procedure of expected value maximization is unreasonable in any context and hence to claim that there is no rational basis, say, for investors in the stock market to choose investments that maximize their expected financial returns. Likewise, it seems equally absurd to suggest that the (apparent) impossibility of quantitative comparison between utilitarian and deontological moral considerations should also preclude comparisons, e.g., among possible values in a consequentialist scheme under conditions of more limited moral uncertainty.[14]

[13] Thanks to an anonymous reviewer for this example.

[14] MacAskill (2013) points out that incomparability can be "infectious" in that an agent who has any positive credence, however slight, that incomparable values are at stake in a given choice situation and tries to calculate the expected value of her various options will in general find them to be undefined. But again, this worry is not unique to the problem of moral uncertainty, and is not solved by ignoring moral uncertainty. The investor who has some vanishingly small credence that

If we conclude that intertheoretic value aggregation is sometimes possible and sometimes not, though, what picture of rational choice under moral uncertainty does this leave us with? One possibility is that the sorts of views endorsed by pessimists about intertheoretic aggregation—like MFT, MFO, or irrelevantism—should be recast with *comparability classes* of moral theories in the role that the standard versions of these views give to *individual* moral theories. That is, if we conclude that comparison and aggregation are simply impossible for some pairs of moral theories, we might then accept a second-order view like "Choose an option that the comparability class of theories in which you have highest credence regards as permissible," or "Choose the option that a credence-weighted plurality of comparability classes in which you have positive credence regards as permissible," or "Choose an option that the *true* comparability class of moral theories (i.e., the comparability class that includes the true moral theory) regards as permissible." Alternatively, it may turn out that small comparability classes that offer precise, quantitative bases for intertheoretic aggregation can be grouped together into larger classes that offer progressively weaker or rougher bases for aggregation, in a kind of nested hierarchy. We will explore these possibilities at greater length in the next chapter.

her investment decisions implicate values like patriotism and family that she judges, as a matter of first-order moral theory, to be incomparable, will find herself in a similar predicament. The problem of infectious incomparability, in this respect, is structurally analogous to more familiar decision-theoretic problems like Pascal's Wager (a problem, one might say, of "infectious infinities"), and must likewise be resolved in a general way by decision theory.

5.5 "Top-Down" vs. "Content-Based" Approaches to Intertheoretic Aggregation

It may seem unparsimonious to hold that different cases of moral uncertainty should be treated differently—e.g. that some classes of theories, like the consequentialist theories in which Alice has positive credence, should be aggregated expectationally, while others, like the deontological theories in which David has positive credence, should be aggregated according to a threshold principles, while perhaps still others cannot and should not be aggregated at all. If one is sufficiently averse to such a piecemeal approach, then one might conclude on this basis that the existence of incomparability between *any* pair of moral theories is enough to render all intertheoretic comparisons moot, even when those comparisons seem quite natural. I will close this chapter, therefore, by describing what I take to be some general attractions of this piecemeal approach—what I will call the "content-based" approach to intertheoretic aggregation—before attempting to develop the approach in greater detail in Chapter 6.

The standard approach to intertheoretic aggregation, comprising nearly all the positive theories of rational choice under moral uncertainty that have been proposed in the literature to date, is what I will call the "top-down" approach. It is characterized by the following two features.

Top-Down Aggregation (1) There is a single general principle of intertheoretic aggregation and choice under moral uncertainty that applies to all first-order moral theories and to all agents regardless of their distribution of belief over first-order moral theories. (2) This general aggregation procedure is sensitive only to the value assignments of the first-order theories in which the agent has positive credence (as well as the degree of credence assigned to each theory), not to any of the other content of those theories.

This sort of top-down approach is exemplified by Lockhart (2000). On Lockhart's view, morally uncertain agents should always choose options that maximize expected moral rightness[15], where the expected moral rightness of options is determined by normalizing the value scales of all moral theories in which the agent has positive credence according to the following principle.

Principle of Equity among Moral Theories (PEMT) "The maximum degrees of moral rightness of all possible actions in a situation according to competing moral theories should be considered equal. The minimum degrees of moral rightness of possible actions in a situation according to competing moral 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)." (Lockhart, 2000, p. 84)

Thus on Lockhart's view both the principle for making intertheoretic com-

[15] More precisely, Lockhart's principle PR4 asserts that "In situations in which moral agents are uncertain of the degrees of moral rightness of some of the alternatives under consideration, a choice of action is rational if and only if the action's expected degree of moral rightness is at least as great as that of any other alternative" (Lockhart, 2000, p. 82). This is later supplemented by another principle, PR5, that governs situations in which an agent cannot be sure which option satisfies PR4 (p. 95).

parisons (viz., PEMT) and the procedure for aggregating rival theories in light of those comparisons for purposes of rational decision-making (viz., maximize expected moral rightness) apply to any agent regardless of her distribution of first-order moral belief. And PEMT, which treats the *range* of every theory's value assignment over options in a given choice situation as equal, is insensitive to any features of those first-order moral theories apart from their value assignments.[16]

The alternative approach suggested by the cases of Alice and David differs from this top-down approach in both respects.

Content-Cased Aggregation (1) Different aggregation procedures are appropriate for different sets of moral theories, or at least for agents who distribute positive degree of belief over different sets of moral theories. (2) These aggregation procedures are, at least in some cases, sensitive not just to the value assignments of the theories they aggregate, but to the metaethical and other theoretic claims that *ground*, *justify*, or *explain* those value assignments.[17]

Setting aside arguments for or against any specific top-down or content-based

[16] Other approaches to decision-making under moral uncertainty that are "top-down" in this sense include MFT, MFO, the "Conceivability PEMT" approach (proposed but not endorsed by Sepielli (2013)), and Bostrom's "parliamentary model" (Bostrom, 2009). The "background rankings" approach defended in Sepielli (2009) is top-down insofar as it is sensitive only to the value assignments of the theories it aggregates, but Sepielli does not claim that this approach to aggregation can be applied to *any* pair of first-order theories. The approach defended in MacAskill (2014) is difficult to classify, but includes a large top-down component that involves normalizing theories at the variance of their value assignments. (MacAskill stresses the distinction between cardinal and ordinal theories, but I will argue in the next chapter that the variance normalization approach effectively ignores this distinction.) As noted above, Ross (2006) briefly suggests the possibility of what I am calling content-based aggregation. To my knowledge, this is the closest anyone in the literature has come to defending a content-based approach.

[17] Greaves and Ord (forthcoming) also consider a "content-based" approach to intertheoretic value comparisons. But they use this term differently than I do, in particular to cover ideas like Sepielli's "background rankings" approach that are sensitive to features of a theory's value assignment that an approach like Lockharts PEMT ignores, but are not sensitive to any features of a theory beyond its value assignment.

theory of rational choice under moral uncertainty, there are two general reasons to prefer the content-based approach. The first is the intuitive relevance of the content underlying a first-order theory's value assignment, which top-down approaches by definition ignore. On the one hand, in cases like the Easy Case, top-down approaches cannot yield the intuitively correct normalization of rival theories. Lockhart's PEMT, for instance, will almost always imply that the value of a hedon according to Alice's hedonistic theory is either more or less than the value of a hedon according to her pluralistic theory, even though there seems to be every reason to think these values equal.[18] On the other hand, top-down approaches miss intuitive *in*equalities between theories with indistinguishable value assignments. For instance, consider two versions of hedonistic utilitarianism, one that derives its value assignment from a divine command theory of ethics plus an unusual reading of scripture, and another that derives its hedonism from the Humean theory of reasons plus an unusual theory of human psychology. The value assignments of these two theories are isomorphic (and therefore, prior to normalization, indistinguishable). Nevertheless, it seems quite plausible that an agent who has some credence in hedonistic utilitarianism should give those utilitarian reasons more weight if her credence is in the divine command theory rather than the Humean theory. A top-down approach like Lockhart's, however, will assign the same weight to both theories.

The second reason to prefer the content-based approach is more subtle. The content-based approach, I have suggested, offers a plausible response to the grounding problem, i.e. a plausible account of why an agent is rationally required to aggre-

[18] See §§B.2-B.3 for an extended defense of this claim.

gate her various first-order theories in a particular way: namely, that the aggregation procedure in some sense follows from the content of those theories, so that she rationally ought to accept it conditional on any of those theories being true. For instance, the principle that an agent ought never choose any option that carries a risk greater than .5 of violating an absolute moral constraint is endorsed by both the deontological theories in which David has positive credence, and implies a particular procedure for aggregating those theories and making rational choices in light of his uncertainty between them. Top-down approaches cannot claim this advantage, because the second-order principles they endorse apply to any agent regardless of her first-order credences, and therefore apply to agents whose first-order moral beliefs may be in deep tension with, if not actively contradict, the second-order principle proposed by the top-down account. That is, any top-down approach runs the risk of "begging the question," in the sense we have encountered in previous chapters, against first-order theories in which many agents may have justified positive credence.

Consider the ways in which Lockhart's approach might be in tension with an agent's first-order moral beliefs. First, as Lockhart himself acknowledges, the principle that uncertain agents should always maximize expected *moral* rightness assumes the lexical priority of moral over non-moral reasons (a principle Lockhart labels the "finality thesis"). This is a claim, however, that many first-order normative views reject (e.g. Wolf (1982)), and about which a reasonable agent might be substantially uncertain. Second, the application of PEMT requires that all the first-order theories over which an agent distributes her belief be in the business of assigning "degrees of moral rightness" to actions. Many deontological theories, of course, do not treat rightness and wrongness as exhibiting gradation, and though one can for instance simply treat such theories as assigning +1 to all morally obligatory actions, −1 to all morally prohibited actions, and 0 to all morally neutral actions, one arguably does such theories an injustice by cramming them into this sort of quantitative mold. But even more clearly, virtue-theoretical approaches to ethics will not fit neatly into Lockhart's box. Virtue ethics is arguably not in the business of classifying actions as right or wrong at all, let along right or wrong in degrees, and it is certainly at least reasonable to have substantial credence in a version of virtue ethics that conceives of itself in this way. If you do, then Lockhart's second-order view will be in some tension with your first-order beliefs.[19] Third, as Sepielli (2006) notes with respect to Lockhart's view, as Hedden (2016) notes with respect to expectational approaches to moral uncertainty generally, and as we saw in the last chapter, approaches like Lockhart's are in tension with first-order views that embrace features like agent-centered option or supererogation, denying in one way or another that one is rationally required to choose the option for which there is most reason. As I conceded in the last chapter, first-order theories according to which one is sometimes permitted to *ignore* the balance of moral reasons might conflict with and hence undermine second-order principles like Lockhart's.

This list of potential tensions is by no means exhaustive, but it will serve to illustrate the point. Lockhart's approach will look much more appealing to someone

[19] For useful discussion of the limited compatibility of deontological and virtue ethical theories with the sort of decision-theoretic framework Lockhart presupposes, see Colyvan et al. (2010).

who divides her first-order moral credence mainly among maximizing consequentialist theories than to someone who divides her credence mainly between Kantian deontology, Aristotelian virtue ethics, and "commonsense morality" views that incorporate large helpings of moral-prudential conflict and supererogation. The minimal point is that the latter sort of agent will likely be very uncertain about Lockhart's principle, suggesting either that she should adopt some alternative second-order principle that seems more plausible in light of her evidence, or that she should be in the market for some *third*-order principle of manage her uncertainty about second-order principles. It is less obvious that Lockhart's principle actually *contradicts* any of the first-order views this agent entertains—after all, it is not *logically inconsistent* to hold, for instance, that as a first-order matter rightness and wrongness don't come in quantitative degrees, but that as a second-order matter it is rational to treat them *as if* they come in degrees (and likewise for other incongruous pairings of first- and second-order principles). But in any case, there is something quite plausible about the thought that our first agent, who distributes her first-order belief mainly over forms of maximizing consequentialism, should appeal to an expectational principle like Lockhart's to guide her actions, given how well it jibes with her first-order beliefs, while there is something quite *im*plausible about the thought that our second agent should do the same, given how *poorly* these expectational principles (at least in their Lockhartian manifestation) jibe with her first-order beliefs.

The advocate of a top-down approach like Lockhart's, therefore, cannot plausibly claim that any agent, whatever her other beliefs and evidence, is rationally required to *believe* the same second-order principle. For any such principle (like "maximize expected rightness" + PEMT), there will be many agents who, in light of their justified first-order moral credences, have reason to disbelieve that principle. Rather, top-down theorist seems forced to claim that their preferred second-order principle has external, belief-independent normative force. As I will argue in Chapter 7, externalism itself is not an unacceptable price to pay—as Weatherson claims, we all must accept it in at least some weak form to escape the threat of infinite regress. But, as I will also argue in Chapter 7, the principles that can most plausibly be regarded as having external normative force are those that are constitutive of rationality as such (namely, I will claim, the enkratic principle). Particular intertheoretic aggregation procedures involving technical normalization principles like PEMT are not good candidates to play this sort of constitutive role in a theory of rationality. Thus it is more plausible to hold that at least some features of intertheoretic aggregation are relativized to the theories being aggregated. This will leave us with a piecemeal picture of rational choice under moral uncertainty, but it will also leave us with a picture that can be largely grounded in the agent's own normative beliefs, rather than imposed on those beliefs by force.

So far, however, we have only seen hints of what a content-based approach to intertheoretic value aggregation might look like, derived from examination of particular cases like those of Alice and David. The task of the next chapter will be to develop a more general theory that can go beyond these simple cases and address more difficult cases in which agents distribute their beliefs over a wide range of first-order theories.

Chapter 6: Outline of a General Theory

In the last chapter I argued that the existence of classes of moral theories within which precise and perspicuous intertheoretic value comparisons are possible means that the problem of intertheoretic value comparisons does not constitute a general objection to moral hedging. And I gestured at a general account of decision-making under moral uncertainty suggested by the existence of such comparability classes. The goal of this chapter is to make good on that suggestion and describe in greater detail a decision procedure based on aggregating the prescriptions of competing moral theories via a nested hierarchy of comparability classes.

The account given in this chapter will nevertheless be incomplete and tentative. The problem of identifying an ideally rational, fully general procedure for decision-making under normative uncertainty is enormously difficult, as demonstrated by the weaknesses of every extant attempt in the literature (some of which I have already criticized, and others of which will be addressed in this chapter). My aim is simply to offer a new candidate approach that avoids many of the drawbacks of extant proposals, but which will need a great many details filled in before it can be fully assessed.

To describe this approach, we must first give a more precise account of the problem it aims to solve. So in the next section I will distinguish three aspects of the problem of intertheoretic value aggregation, i.e., three barriers that an adequate approach to intertheoretic aggregation and decision-making under moral uncertainty must overcome. Then, in §§6.2-6.4, we will examine each of these three problems individually, describing and comparing several candidate approaches to each problem. Finally, §6.5 combines what seem like the most promising approaches to each aspect of the problem of intertheoretic value aggregation to arrive at a tentative, content-based account of rational decision-making under moral uncertainty.

6.1 Three Components of Comparability

The examples of apparent intertheoretic comparability given in the last chapter possess three advantageous features of which I implicitly or explicitly took advantage in arguing for the possibility of aggregation in these cases. In this chapter I will suggest, as a plausible hypothesis, that these three conditions are jointly sufficient for intertheoretic comparability and aggregation whenever they are fully realized by a pair of theories T_1 and T_2. The three features are as follows.

1. **Structural Compatibility** The value scales of T_1 and T_2 have identical structure (e.g., ordinal, interval, or ratio).

2. **Scale Normalizability** There is some set of practical options $\{O_1, O_2, ...O_n\}$ such that T_1 and T_2 are in full agreement concerning all the features of the world that ground their evaluations of O_{1-n}, so that the value scales of T_1 and T_2 can be normalized based on their agreement with respect to those options.

3. **Decision-Theoretic Compatibility** T_1 and T_2 endorse the same decision theory, i.e., the same principles for aggregating conflicting value assignments in the face of uncertainty.

For instance: In the Easy Case, where Alice divides her beliefs between hedonistic utilitarianism and a pluralistic theory that values both hedons and aesthetons, (i) structural compatibility is given by the fact that both theories have cardinal value scales; (ii) scale normalizability is given by the fact that both theories are in perfect agreement about the choiceworthiness of options that implicate only hedonic and not aesthetic value; and (iii) decision-theoretic compatibility is given by the fact that both theories endorse expected value maximization as the correct decision procedure under uncertainty.

In the case of David, who divides his belief between two deontological theories, some details must be filled in, but we can tell a similar story. For instance, perhaps both deontological theories share a lexical value structure with one ordinally-structured, lexically prioritized dimension representing deontological obligations and prohibitions and another, cardinally structured, lexically deprioritized dimension representing all other sorts of reasons. Scale normalizability comes from the presumed fact that both theories, despite disagreeing for instance about the permissibility of killing innocent threats, nevertheless agree about the strength of other moral obligations (for instance, against breaking promises or killing innocent non-threats) as well as the strength of non-moral reasons. And decision-theoretic compatibility is given by the fact, which we stipulated, that both theories endorse the same threshold approach to decision-making under uncertainty about deontological constraint violations.

Are these three conditions, in all cases, jointly sufficient to establish the possibility of intertheoretic comparison and aggregation under moral uncertainty? I'm not sure that any general proof of this claim is possible. But intuitively, structural compatibility and scale normalizability allow us to put the value assignments of T_1 and T_2 on a single, shared scale, and decision-theoretic compatibility gives us a decision procedure, the correctness of which we are certain of conditional on the disjunction of T_1 and T_2, for handling uncertainty about the values of options on that shared scale. So it seems reasonable to think that whenever all the theories in a given class satisfy these three conditions with respect to one another, it should be possible for an agent to aggregate the competing value assignments of those theories. On the tentative assumption that these three conditions are sufficient for full intertheoretic comparability and aggregation, let's call a class of moral theories that fully satisfies these conditions (that is, a class any two members of which have identically-structured and normalizable value scales and endorse the same decision theory) a *minimal comparability class.*

But if we take these three conditions as jointly sufficient for intertheoretic aggregation, we should next ask to what extent they are individually necessary—that is, how should an agent decide what to do when she distributes her credence over a class of theories that satisfies one or more of these conditions only partially, or not at all? If we start from the idea that the sorts of aggregation procedures described above, in simple cases where all three conditions are met, are rational when an agent distributes the entirety of her moral credence over a single minimal comparability class, then a natural way to develop a general theory of decision-making under moral uncertainty is to ask what happens as we relax each of these conditions. This is the approach I will adopt in this chapter: We will consider each of the three conditions in turn, discuss various ways in which that condition might be less than fully satisfied, and consider various ways of extending the account of decision-making under moral uncertainty from the simple case of minimal comparability classes to accommodate these increasing difficulties. My primary aim is to identify possibilities, i.e., reasonable candidate answers for how to accommodate weakening of each of the three conditions that can then be combined into reasonable candidate theories of decision-making under moral uncertainty in general. But in §6.5 I will identify the combination of responses, and hence the general theory of decision-making under moral uncertainty, that strikes me as most plausible.

6.2 Structural Compatibility

Perhaps the most striking challenge of intertheoretic value aggregation is the need to compare theories with differently structured value scales. For instance, anyone who wishes to assert that maximizing expected choiceworthiness provides a *universal* principle for decision-making under moral uncertainty must assume that all moral theories assign cardinal values to options that, suitably normalized, can be taken as inputs to compute intertheoretic expectations. But, as MacAskill in particular has emphasized (MacAskill, 2014, pp. 14-15, 52ff), it is simply a fact to be reckoned with that not all theories have cardinal structure. Some moral theories have *ordinal* structure, ranking options as better or worse without giving answers to (or recognizing as meaningful) the question "How *much* better/worse?" Since ordinal values cannot be multiplied by probabilities to yield cardinal values, an agent who has positive credence in merely ordinal theories cannot assign expected values to her options.

Faced with such structurally uncooperative theories, it is tempting to simply *demand* uniform structure from our moral theories. Lockhart's PEMT, for example, effectively forces cardinal structure on theories: (i) Lockhart holds that the maximal and minimal "degrees of rightness" assigned by each theory to any option in a given choice situation should be treated as equal *on a cardinal scale*, e.g., assigning the maximally right option value 1 according to that theory and the minimally right option value 0.[1] (ii) If a theory treats any option in a given choice situation as *intermediate* in rightness, value, or choiceworthiness between the best and the worst option, then it must be assigned some intermediate value on that cardinal scale.

But if an agent has some credence in a theory with a non-cardinal value scale, i.e., a theory according to which the choiceworthiness of practical options does not come in cardinal degrees, then an approach like Lockhart's is simply refusing to grapple with the phenomenon, i.e., with the agent's actual doxastic state. Thinking of theories, again, as maximal consistent sets of propositions, some theories include propositions like "The choiceworthiness of practical options can be compared ordi-

[1] Lockhart doesn't specify to *which* cardinal values the maximal and minimal degrees of rightness for each theory should be normalized, but this choice is of course arbitrary.

nally but not cardinally." To treat a theory T that includes this proposition as if it must secretly be assigning cardinal values to options that can be factored into a computation of expected value if we can only get at them is to solve the problem of moral uncertainty, not for an agent for has positive credence in theory T, but only for an agent who has credence in some cardinal surrogate of T.

If we want to know how an agent should act when she distributes belief over an arbitrary range of moral theories, therefore, we must grapple with the problem of structural diversity. MacAskill (2014) is the first to seriously attempt this, offering an account of rational choice under moral uncertainty that is bifurcated between cardinal and ordinal theories: the appropriate decision rule for cardinal theories, he holds, is to maximize expected choiceworthiness, while the appropriate decision rule for ordinal theories is a complex, probability-weighted version of a Borda count.

As I will argue at more length shortly, it seems to me that MacAskill's approach fails to genuinely respect the phenomenon of merely ordinal theories, since the Borda counting approach (and MacAskill's version of it in particular) is non-arbitrary only on the assumption that there are "hidden" cardinal values underlying the ordinal rankings of merely-ordinal theories. But a more immediate problem for MacAskill's account is that it acknowledges only a very small part of the potential structural diversity of moral theories—for many more structures are possible than the cardinal and ordinal structures on which MacAskill's account in focused.[2]

It is worth cataloging some of the ways in which moral theories might differ

[2] MacAskill distinguishes between interval-structured and ratio-structured cardinal theories, but given his view that we should always maximize expected choiceworthiness over cardinal theories, this distinction plays no role in his ultimate theory. He also briefly acknowledges the possibility of lexically-structured theories with two cardinal dimensions (MacAskill, 2014, pp. 47-50).

from one another in the structure of their value scales.

1. The cardinal structure of the real number line might be extended in various ways to accommodate the possibility of infinite or infinitesimal values. For instance, a theory that recognizes the possibility of infinite value or disvalue but does not countenance distinctions of degree between infinite values has a value scale with the structure of the *extended* real number line (the real numbers plus limit values $+\infty, -\infty$). Alternatively, if we wish to distinguish between degrees of infinite value and/or to represent the possibility of infinitesimal value, we might have some credence in a theory whose value scale has the structure of the hyperreal numbers or the surreal numbers. And of course many other such extensions of the real numbers as possible as well.[3]

2. A theory's value scale might have more than one dimension. MacAskill briefly considers a theory that represents degrees of choiceworthiness by two-dimensional vectors, lexicographically ordered (i.e., $(x_1, y_1) \geq (x_2, y_2) \iff (x_1 > x_2 \vee (x_1 = x_2 \wedge (y_1 \geq y_2))$. Such a two-dimensional structure could also be used to represent incomparable values ($(x_1, y_1) \geq (x_2, y_2) \iff (x_1 \geq x_2 \wedge y_1 \geq y_2)$). Either of these ideas, of course, could be extended to arbitrarily many dimensions (even infinitely many dimensions if, for instance, one thinks that there might be infinitely many incomparable forms of value). And the various dimensions of a multi-dimensional value scale could themselves display various

[3] Bostrom (2011) describes, without endorsing, one way of using the hyperreal numbers to compare the utilitarian value of infinite worlds. Several other approaches to the problem of infinite utility (e.g. those of Vallentyne and Kagan (1997) and Arntzenius (2014)) extend the structure of the utilitarian value scale beyond that of the real numbers is less drastic ways.

structures, e.g. ordinal, interval, ratio, extended reals, hyperreals, etc.

3. At the other end of the scale of complexity, a theory could simply have what we might call *binary structure*: classifying options as permissible or impermissible, but countenancing no ranking of permissible options as better or worse. This might be the structure, for instance, of an extreme libertarian view according to which morality consists entirely of a list of absolute, negative side constraints, which it is never permissible to violate, and such that all options that do not violate any side constraint are equally choiceworthy.

Is there any order underlying this chaos of possible structures? Perhaps not, but the following strikes me as a plausible hypothesis: What is fundamental to any normative theory is *binary* structure—more specifically, the identification of some subset of an agent's options as permissible or "eligible for choice." Since the fundamental job of any normative theory is to guide an agent's choices, to count as a normative theory it must "tell the agent what to do," i.e., it must declare one or more of her options in any given choice situation choiceworthy *simpliciter*.[4] Therefore, nothing can count as a normative theory (or, *a fortiori*, as a moral theory) which does not make the binary distinction between permissible and impermissible options. But the diversity of possible structures explored above is explained by the

[4] Of course, *pace* irrelevantism and MFT, the value assignment given by a particular moral theory is not fully action-guiding in that it does not tell an agent what to do in light of her moral uncertainties. If we consider the fully objective value assignments given by a complete (moral and non-moral) theory of the world, this is all the more true, since an objective value assignment does not guide an action's actions even in light of her *empirical* uncertainties. Nevertheless, the value assignments given by moral theories (whether fully objective or partly subjective) play an essential *role* in action guidance, both as inputs to an action-guiding, fully subjective value assignment, and insofar as an agent who is *certain* that a particular moral theory is true must be able to derive action guidance from its value assignment.

many possible extensions of binary structure: any structure that extends the binary structure of permissible/impermissible (or, equivalently, that can *induce* a binary structure) is a possible structure for a moral theory's value scale.

Since this idea of universal binary structure offers helpful suggestions for the general theory of decision-making under moral uncertainty, it is worth a brief exploration. First, is it true that all moral theories make a binary classification of options as permissible or impermissible?[5] Consider some examples: Classical utilitarianism ranks options on a cardinal value scale, and hence does not *simply* classify options as permissible or impermissible. But it holds that all and only the options that produce maximal utility are permissible, so it has a binary structure induced by its cardinal structure. A satisficing version of utilitarianism has the same cardinal structure as maximizing utilitarianism, but draws the line between permissible and impermissible options differently, regarding some options that don't produce maximal utility as nevertheless permissible.[6] Like satisfying utilitarianism, commonsense morality (which recognizes the supererogatory) and perhaps also Kantianism (which its idea of imperfect duties) recognize that some permissible options are more choiceworthy than others, while nevertheless preserving a distinction between permissible and impermissible.

One type of theory that might seem to challenge the universality of binary

[5] Here it is particularly worth noting that the following discussion is meant to generalize to *normative* theories, moral or otherwise, but I focus on moral theories for simplicity.

[6] This suggests that theories like utilitarianism have *more* than just cardinal structure, i.e., they have cardinal structure plus a "permissibility" predicate that does not logically supervene on the assignment of cardinal values to options. Some might prefer to treat the division of options as permissible or impermissible as something distinct from a theory's value assignment, but to the extent that the notion of universal binary structure proves helpful, it gives us reason to think of that binary structure as part of the value assignment itself.

structure comprises those theories that recognize value incomparability and/or moral dilemmas (i.e., genuine conflicts of obligation). Suppose I must decide whether to stay home and care for my ailing parent or go join the Resistance, and that this choice presents either a clash of incomparable values (familial obligation vs. patriotism, justice, etc) or simply a conflict of obligations. Does a moral theory that issues these verdicts tell me that I have permissible and impermissible options? I think so: It may be, especially on a "moral dilemmas" view, that both my options are simply impermissible—the universality of binary structure does not entail that at least one option in any given choice situation must be classified as permissible, any more than it entails that at least one option in any given choice situation must be classified as impermissible, though of course the former principle strikes many philosophers as extremely compelling. Perhaps it is more plausible to say that both of my options are *permissible*, despite in some sense violating moral obligations, on the grounds that I am in a situation where I cannot help violating a moral obligation. Incomparabilist views, though they likewise do not identify either option in this situation as *better* than the other or as *equal* in choiceworthiness (in the sense that any small addition of reasons one way or the other would make the newly favored option better overall), nevertheless must give me *some* guidance on what to do, and will presumably say that either option is permissible (though again, they have the option of saying that both options are impermissible).

Another kind of theory that might challenge the universality thesis is scalar utilitarianism (Norcross, 2006). Scalar utilitarianism explicitly disclaims the notions of obligation, prohibition, and permissibility, ranking actions only as better or worse (and consequently as supported by cardinally greater or lesser degrees of objective reason). It strikes me that there is something unstable about the scalar utilitarian position: An ideally informed agent who accepts scalar utilitarianism and faces a choice situation in which she knows with certainty that some option O is objectively best would exhibit a clear rational failure were she to choose any option other than O—which seems to provide a clear sense in which O, and only O, is a *permissible* option in this choice situation, i.e., O is the only option that is objectively rational, or that is rationally permissible for an ideally informed agent. The one way in which a scalar utilitarian could disclaim this implication, it seems to me, is by insisting that scalar utilitarianism offers only "a theory of the good" and not "a theory of the right"—i.e., that it is an *evaluative* theory (though evaluating actions as well as states of affairs) rather than a *normative* theory that describes or implies *reasons for action*. But of course if scalar utilitarianism is not a normative theory, then it is not a moral theory in the sense with which we are here concerned, and it need not be accommodated by a general theory of decision-making under moral/normative uncertainty.[7]

An interlocutor might complain, at this point, that I am doing exactly what I criticized Lockhart for doing, namely, imposing structural requirements on theories that the theories themselves disclaim, and hence generating a false representation of the credences of some agents (who have positive credence in theories that disclaim

[7] The version of scalar utilitarianism defended by Norcross (2006) is explicitly normative in the sense I have described, countenancing reasons for action but not obligation/prohibition, rightness/wrongness, or oughts. Norcross is focused on *moral* notions, so whether his view rules out the notions of objective rational requirement/prohibition/permission that I have described—understood in terms of what options an ideally rational and fully informed agent might choose—is not obvious.

any binary structure). But there must be *some* conceptual constraints on what counts as a moral or normative theory: Any agent might have positive credence in any number of sets of propositions (or at least, sets of sentence) that proclaim themselves to be moral theories but simply are not—e.g., "The only moral truth is that $1 + 1 = 2$" (a theory that says nothing about practical options) or "The only moral truth is that all choices made on a Tuesday or Friday are blue and all choices made on any other day of the week are green" (a theory that assigns properties to practical options, but not the *right* properties to count as a moral theory). That an account of choice under normative uncertainty fails to accommodate such theories is no strike against it—the goal of such an account is to aggregate the verdicts of those theories that *actually meet* the conceptual requirements to count as a normative theory. Cardinal structure is not plausibly such a requirement (though it may well be plausible as a requirement that the *true* normative theory must satisfy). But the binary classification of options as eligible for choice or ineligible for choice in light of objective reasons plausibly is such a requirement (and one that can be maintained even if we wish to shed much of the conceptual baggage associated with *permissibility* and *impermissibility*).

Are there other structural constraints besides binary structure that all normative theories must satisfy? One might think that the action-guiding role of normative theories requires that they identify not just some subset of an agent's options in a given choice situation as eligible for choice but *one* option as *uniquely* eligible for choice, as "the thing to do." A normative theory that merely identifies some set of options as choice-eligible leaves agents with a "liberty of indifference" problem: even if it is the case that, say, chocolate and vanilla ice cream are both equally choiceworthy, then a full normative theory should tell me how to choose between them, e.g. by telling me to choose on the basis of the first arbitrary whim that comes into my head, or to flip a coin. But it seems to me that the liberty of indifference is not so great a problem as to justify this "uniqueness" requirement as a conceptual constraint on normative theories. A commonsense theory of reasons, for instance, seems to tell me that I am permitted to choose chocolate straightaway, choose vanilla straightaway, or adopt any sort of cost-free mixed strategy (like flipping a coin), and which of those options I choose is simply up to me, underdetermined by my reasons. The fact that my normative theory does not direct me to a particular option (whether a simple action, a deterministic decision procedure, or a stochastic decision procedure) does not mean that I will fail to make a decision.[8]

It might also be suggested that binary structure is just a degenerate or impoverished limiting case of ordinal structure, and that we should in fact conclude that *ordinal* structure is the thing all normative theories have in common. It may be unproblematic to concede this claim.[9] But it strikes me as worthwhile to distinguish between theories (like the simple libertarian view given above) that recognize only two possible degrees of choiceworthiness—*permissible* and *impermissible*—and theories that countenance the possibility of arbitrary ordinal comparisons. It is a structural feature of the former class of theories that there are never three options

[8] It is worth remembering that an agent's options, in any given choice situation, will typically include various deterministic and stochastic procedures for selecting among other options.

[9] Though it has substantive implications if one thinks that a voting method like Borda counting is the right way to aggregate ordinal theories—in particular, it puts a great deal of pressure on the question of how to extend the simple Borda counting procedure to accommodate ties (i.e., normative theories classifying two or more options as equally choiceworthy).

O_1, O_2, O_3 such that $O_1 \succ O_2 \succ O_3$, while fully ordinal-structured theories allow this.[10] Binary structure may be a degenerate case of ordinal structure, but it is in any case distinct from *full* ordinal structure.

I conclude, therefore, that the value scales of moral theories can display a much greater diversity of structures than has previously been recognized: namely, every possible extension of binary structure, including ordinal, interval, ratio, extended real, hyperreal, surreal, multidimensional (lexical or incomparabilist), and surely many more.

6.2.1 Structural Enrichment and Structural Depletion

Having laid a deal of groundwork, we may now ask how a moral agent should deliberate when she divides her beliefs among moral theories with differently-structured value scales. It seems to me that there are three general options.

1. **Structural Enrichment** Give each theory all of the common structure necessary for intertheoretic comparison by enriching, in one way or another, its own self-proclaimed structure, e.g. mapping ordinal rankings onto a cardinal value scale in order to achieve comparability with other cardinal theories.

2. **Structural Depletion** Adopt a decision procedure that relies only on minimal structure, either the least structure had in common by all the theories in which the agent has positive credence, or the least structure had in common

[10] And of course, to further emphasize the possibility of structural diversity, we could imagine theories that recognize exactly three, four, or any finite number of possible degrees of choiceworthiness. More esoterically, ordinal theories might also differ in whether they allow or disallow infinite ordinals in comparing infinite classes of options, or confine their value scales to the finite ordinals.

by all normative theories full stop, namely binary structure.

3. **Multi-Stage Aggregation** Aggregate the value assignments of theories within a comparability class C_{1-1} in ways that make use, where appropriate, of their common structure. Then take the aggregate "value assignment" of the comparability class an input to a new aggregation procedure for a larger comparability class $C_{2-1} = \{C_{1-1}, C_{1-2}, C_{1-3}, ..\}$, where that new procedure makes use only of the structure common to all theories in C_{2-1}. Continue in this way until all theories in which the agent has positive credence have been aggregated into a single ranking of options in terms of subjective reasons.

The structural enrichment approach is typified by Lockhart's PEMT which (as described above) simply forces theories into cardinal form. But it is also implicitly exemplified, I believe, by MacAskill's approach. MacAskill holds that merely ordinal theories should be aggregated by means of a Borda count. Borda counting is a voting method which, in its simplest form, assigns each candidate a score equal to the sum over every voter's ballot of the number of candidates ranked below her on that ballot, and then selects the candidate with the highest score. On MacAskill's approach, moral theories play the part of voters and practical options the part of candidates. Ignoring for the moment the possibility of a theory ranking two options as equally choiceworthy, the Borda score of an option O is equal to the sum over every moral theory in which the agent has positive credence of the number of options ranked as worse than O according to that theory times the agent's credence in that theory.[11]

[11] MacAskill also proposes that Borda scores should be weighted by a measure on the set of options, so as to avoid problems with option individuation, but we can safely ignore this compli-

Borda counting holds some appeal in the social choice context because, in many cases, we might hope that the number of options/candidates that a given voter ranks above or below a given candidate offers a rough approximation of the cardinal utility she assigns to that option/candidate—that is, it is sometimes reasonable to presume that for most voters, the difference in strength of preference between the first- and second-ranked candidate is similar to the difference in strength of preference between the second- and third-ranked candidate, the third- and fourth-ranked candidate, and so on. Since it is plausible that, at least in many contexts, we would ideally like to maximize cardinal preference utilities if we could only get at them, the Borda method might look like an appealing approximation.

But in the context of choice under moral uncertainty, if we accept that ordinal theories are genuinely ordinal, and not simply cardinal theories that stubbornly refuse to reveal their cardinal value assignments, then this rationale for Borda counting is unavailable. Indeed, absent an assumption of implicit cardinality, the Borda counting approach appears arbitrary: Why, for instance, should being ranked third-from-last by a given theory contribute exactly twice as much to an option's score as being ranked second-to-last, if the ranking genuinely does not reflect any cardinal information?

Moreover, MacAskill defends a version of Borda counting that handles the possibility of ties in a theory's ordinal ranking of options by normalizing each theory's Borda score at its variance (MacAskill, 2014, p. 119).[12] But this amounts

cation.

[12] That is, MacAskill's approach first represents a theory's ordinal ranking with cardinal values 1, 2, 3, etc. It then calculates the variance of those cardinal values, i.e., the average squared distance from the mean. Finally, it "stretches" or "contracts" each theory's ranking in order to

to an explicit cardinalization of ordinal theories, for there is no such thing as the "variance" of an ordinal ranking. This makes it all the more clear that the Borda method relies on treating ordinal theories as implicitly cardinal.[13]

Approaches that implicitly or explicitly impose alien structure on a theory, e.g. by mapping ordinal rankings onto a cardinal value scale, strike me as *ad hoc* and as failing to grapple with the problem of what an agent ought to do when she has positive credence in theories with genuinely sparse structure. So it is worth looking for alternative approaches that allow us, e.g., to aggregate ordinal theories without treating them as cardinal.

The second approach, *structural depletion*, satisfies this demand. This approach involves appealing to decision procedures that rely only on the structure actually present in all theories in which an agent has positive credence. If an agent has positive credence only in a set of theories that display something less than the full diversity of possible structures, this approach is compatible with a wide variety of decision procedures, about which more will be said below. But for now let's focus on the simplest version of structural depletion, which assumes that an agent distributes her belief over all possible moral theories and therefore demands a decision procedure that relies only on the structure that all moral theories have in common—namely, as I have argued, binary structure.

equalize their respective variances. When none of the ordinal theories being aggregated treat any two options as equally choiceworthy, this amounts to a simple Borda count, since each theory will assign a number of ranks equal to the number of options, and the variance of every theory will be the same to begin with. But when some theories posit ties, the variance normalization method has a substantial impact on the outcome of the Borda method, giving comparatively greater weight to theories that posit a large number of ties.

[13] I discuss MacAskill's Borda method at much greater length in Tarsney (unpublished).

If we set aside all the surplus structure of every moral theory apart from its classification of practical options as permissible or impermissible, then there seems to be one salient option for intertheoretic aggregation: namely, the "My Favorite Option" approach that directs an agent to choose an option that has maximal probability of being objectively permissible.[14] In the absence of any information about rankings of relative choiceworthiness among permissible options or among impermissible options, or about cardinal differences in choiceworthiness, there can never be any basis for rejecting the option that has the best chance of being objectively permissible.[15] So the structural depletion view, at least in its simplest form, amounts to MFO.

I find MFO much more plausible and congenial than either MFT or irrelevantism. Nevertheless, it cannot accommodate any of our intuitions in favor of hedging, i.e., intuitions that the surplus structure of a theory's value assignment (and in particular, its cardinal structure) is often relevant to how we ought to act under moral uncertainty. We have described these intuitions before, but it is worth adding one more to drive the point home. MFO (and hence the simple version of structural de-

[14] Of course, irrelevantism and My Favorite Theory are still available as well, but these approaches do not engage in any form of intertheoretic aggregation. In principle, structural depletion is also consistent with a *threshold view*, a more lax alternative to MFO on which an agent is permitted to perform any option that has a probability greater than t of being objectively impermissible (or, if no option exceeds threshold t, the option that comes closest). This might be motivated by the desire to avoid the intuitive over-demandingness of MFO in cases where several options carry only a very small risk of being objectively impermissible.

[15] It seems plausible, moreover, that all plausible decision procedures that might be proposed in more enriched contexts will collapse to MFO when given only binary information—for instance, probability-weighted Borda counting or approval voting do so straightforwardly, as does stochastic dominance, as does expectational reasoning (regardless of risk attitudes) if one represents permissibility and impermissibility as arbitrary cardinal values such that the value of permissibility is greater than the value of impermissibility. Some relatively weak methods, like statewise dominance reasoning or Condorcet methods that simply issue no verdict when there is no option that a probability-weighted majority of theories prefer to each alternative, will not converge with MFO, but only because they are silent in many cases where MFO is not.

pletion) disallows hedging even in the cases where intertheoretic comparisons seem easiest. For instance, in the Easy Case from Chapter 5, where Alice's certainty with respect to every proposition concerning hedonic value makes it natural to normalize her two theories by assuming that they value hedons equally, we might imagine that Alice assigns credences .51 to hedonism, .49 to pluralism, and is faced with options $O \Rightarrow$ *100 hedons, 0 aesthetons* and $P \Rightarrow$ *99 hedons, 1000 aesthetons*. MFO tells us that Alice should choose O, since it has the greater likelihood of being objectively permissible (.51 to .49), contrary to what seems like overwhelming reason for thinking that she should choose P.

Some versions of the structural depletion view might allow Alice to hedge so long as she distributes her belief only over theories with the right structure (presumably, only over theories with cardinal structure of one sort or another), but structural depletion *must* deny that Alice ought to hedge if we modify the case to give her vanishingly small but positive credence in every moral theory (such that her credences in hedonism and pluralism are reduced to $.51 - \varepsilon$ and $.49 - \varepsilon$ respectively, for some very small ε). I conclude, therefore, that the structural depletion view is unsatisfactory.

6.2.2 Multi-Stage Aggregation

This leaves us with the intermediate option that I have called *multi-stage aggregation*.[16] This approach is best illustrated by means of an example. Suppose

[16] This idea is introduced by MacAskill, who calls it a "multi-step procedure" (MacAskill, 2014, p. 118).

that Dorothy divides her credence between four theories: the two consequentialist theories, hedonism and pluralism, from the Easy Case, and two ordinallly structured deontological theories that disagree about whether one ought to kill innocent threats, but are otherwise in complete agreement. Dorothy must decide whether to kill Eustice: Eustice was innocently driving his truck down the street when his brakes failed, having been sabotaged by some nefarious wrongdoer, and his truck is now careening unstoppably toward five innocent people trapped in a narrow alleyway. The only way to save the five is for Dorothy, perched on a nearby rooftop, to destroy Eustice's truck with a bazooka. In the truck with Eustice, moreover, are a dozen priceless works of art and the only extant score of a newly rediscovered Beethoven symphony. If Dorothy opens fire, therefore, both Eustice and his truckload of aesthetons will be destroyed, but five innocent lives will be saved.

The death of an innocent person, let's assume, amounts to a loss of 20 hedons, while the loss of the artistic contents of Eustice's truck would amount to a loss of 10 hedons.[17] However, their destruction would also amount to a loss of 200 aesthetons, so while Dorothy's hedonistic theory (T_1) supports killing Eustice to save the five, her pluralistic theory (T_2) opposes it. Of Dorothy's deontological theories (T_{3-4}), meanwhile, one claims that it is permissible (indeed, obligatory) to kill Eustice and impermissible not to, while the other claims that it is permissible ("") not to kill Eustice and impermissible to kill him. (For now, we leave it unspecified whether the structure of these theories is binary or ordinal.) Table 6.1 represents Dorothy's

[17] Suppose that, although people will derive enjoyment from these works, that enjoyment is highly substitutable, so that if Eustice's truck is destroyed, the would-be appreciators of the art it carries would be able to derive nearly as much pleasure from the appreciation of other artworks.

	Credence	O_1 (kill)	O_2 (don't kill)
T_1 (hedonism)	.3	-30	-100
T_2 (pluralism)	.3	-230	-100
T_3 (Kill IT's.)	.25	permissible	impermissible
T_4 (Don't kill IT's.)	.15	impermissible	permissible

Table 6.1: Dorothy's Dilemma, Stage 1

credences and the value assignments of each theory to the options of killing and not killing Eustice.

T_1 and T_2 have the same cardinal structure and both endorse expected value as the correct aggregation procedure in the face of uncertainty, so on a multi-stage approach, they should be aggregated expectationally[18]: The expected value of O_1, relative to the comparability class of theories that includes T_1 and T_2, is $(.3 \times -30) + (.3 \times -230) = -78$, and the expected value of O_2 is $(.3 \times -100) + (.3 \times -100) = -60$.[19] Therefore, as far as Dorothy's consequentialist theories are concerned, she ought to choose O_2 (i.e., not kill Eustice).

But how do we aggregate these consequentialist theories with the deontological theories T_3 and T_4? If we assume that the deontological theories have merely binary structure, the multi-stage view gives a straightforward answer: We take the comparability class $C_1 = \{T_1, T_2\}$, along with the theories T_3 and T_4, as inputs to

[18] What the multi-stage approach entails, strictly, is that comparability classes of theories should be aggregated according to whatever aggregation procedure is appropriate given their shared structure and content. For now I assume that the appropriate aggregation procedure for cardinal theories is expectational, at least when the theories themselves endorse an expectational decision procedure. But we will return to this question in §6.4 and Chapter 7.

[19] It makes no difference whether we use the agent's unconditional credences or her credences conditional on the comparability class (that is, conditional on $T_1 \vee T_2$), though the latter would conveniently allow the credences to sum to 1.

	Credence	O_1 (kill)	O_2 (don't kill)
C_1 (consequentialism)	.6	impermissible	permissible
T_3 (Kill IT's.)	.25	permissible	impermissible
T_4 (Don't kill IT's.)	.15	impermissible	permissible

Table 6.2: Dorothy's Dilemma, Stage 2

a binary aggregation procedure. Since, I have claimed, the aggregation procedure appropriate to binary-structured theories is MFO, that means taking the aggregated evaluation of C_1, reducing it to only its binary substructure, and taking that as an input to MFO. Table 6.2 represents this second stage of the aggregation procedure. Since Dorothy's aggregate credence in views according to which it is permissible to kill Eustice is .25, while her aggregate credence in views according to which it is permissible not to kill Eustice is .75, we conclude that she ought not kill Eustice.[20]

Note that the two-stage procedure we have just described is *not* equivalent to MFO *simpliciter*: According to MFO, Dorothy ought to choose the option that has the greatest subjective probability of being objectively permissible, and this is O_1, i.e., killing Eustice: The fine-grained theories that prefer this option, T_1 and T_3, command a combined credence of .55 while the theories according to which Dorothy ought not kill Eustice, T_2 and T_4, command a combined credence of only .45. The source of the divergence is that the inputs to the final, MFO-like aggregation procedure are not all maximally fine-grained theories: C_1, treated in this second stage

[20] I assume that T_3 and T_4 have merely binary structure for the sake of simplicity. However, given the stipulation that there are only two options, any plausible decision procedure for ordinal-structured theories (Condorcet methods, Borda counting, approval voting, etc) will work just like MFO when it is applied at the second stage of the decision procedure. So all that the example really requires is that T_3 and T_4 have the *same* structure, either binary or ordinal.

as a single theory, subsumes the already-aggregated value assignments of two fine-grained theories. From here on, let's call the pure version of MFO that I claimed was entailed by structural depletion *fine-grained MFO*, MFO-like aggregation procedures like the one just described in the case of Dorothy *coarse-grained MFO*, and allow "MFO" without qualification to refer to both fine-grained and coarse-grained MFO.[21]

The attraction of the multi-stage approach is that it preserves sensitivity to the relevant structure of richly-structured (e.g., cardinal) theories without artificially imposing this structure on less-structured (e.g., ordinal) theories: In Dorothy's case, this means that the cardinal values assigned by T_1 and T_2 are allowed to make a difference, without implicitly or explicitly cardinalizing the non-cardinal theories T_3 and T_4. On the other hand, in order to capture these virtues of structural enrichment and structural depletion respectively, the multi-stage view must open itself up to

[21] Note that MFO conceived as a decision procedure *only* for dealing with moral or normative uncertainty is already coarse-grained. For instance, suppose an agent A has .6 credence in utilitarianism and .4 credence in Kantianism, and must choose between two options, O and P. Kantianism implies that O is morally prohibited and P is morally obligatory. O has greater expected utility than P, and hence would be the more rational choice if A were certain of utilitarianism. But A has empirical uncertainties as well as normative uncertainties, and conditional on the truth of utilitarianism, she judges that there is a .3 probability that P would in fact produce more utility than O (as compared with a .7 probability that O would produce greater utility than P). Thus the probability that P is objectively permissible is .58 while the probability that O is objectively permissible is just .42. But A's total credence in *moral theories* that "prefer" O is .6 while her total credence in moral theories that prefer P is only .4.

Fine-grained MFO, which instructs A simply to choose the option with the highest probability of being objectively permissible, would thus tell her to choose P. This may seem counterintuitive, especially given certain specifications of the possible utilities that might result from A's choice, since it simply disregards relevant cardinal information (for instance, suppose there's a .3 probability that P would produce just one more utile than O, but a .7 probability that O would produce 1000 more utiles than P). But on the other hand, coarse-graining at the level of moral theories sacrifices the attractive simplicity of fine-grained MFO, and will seem arbitrary absent some motivation for cutting off cardinal aggregation at the boundary between empirical and normative uncertainty: If we are willing to make cardinal comparisons between the objective value of acts given different states of the empirical world, conditional on a particular moral theory, why not also make cardinal comparisons between the objective value of acts given different moral theories, at least so long as such comparisons are possible (e.g., within minimal comparability classes)?

two corresponding lines of attack.

First, it may seem that the multi-stage view is insufficiently sensitive to the relatively enriched structure of theories like T_1 and T_2. MacAskill, who previously endorsed a version of multi-stage aggregation (Crouch, 2010), repudiates the view for this reason in his (2014). He describes a case in which an agent has (i) slightly more than .5 credence in a pair of cardinal theories that disagree over which of options O_1 and O_2 has greater value but, when aggregated, assign slightly greater expected value to option O_1, and (ii) slightly less than .5 credence in an ordinal theory that prefers O_2 to O_1. A multi-stage aggregation approach yields the conclusion that the agent should prefer O_1. MacAskill finds this implausible. He points out in particular that the agent might come to prefer O_2 by *reducing* her credence in the cardinal theory that prefers O_2 and redistributing that credence proportionately over her other moral theories, if this gave her a credence in the ordinal theory greater than .5. This violates a criterion MacAskill calls "updating consistency," to the effect that increasing one's credence in a theory T that regards option O as maximally choiceworthy, while proportionately reducing one's credence in all other theories, should never make O *cease to be* the most rational option, if it was so initially (MacAskill, 2014, pp. 117-9).

There is an intuitive cost here, to be sure, but to my mind not as great a cost as MacAskill supposes. The fact that, in the initial description of the case, O_1 has only "slightly greater" expected value than O_2 according to the comparability class of cardinal theories *should* make no difference when the cardinal theories are aggregated together with ordinal theories: Since ordinal value differences don't come in degrees, any sense that the value difference between O_1 and O_2 is greater according to the ordinal theory than according to the aggregated cardinal theories is simply misleading. And the principle of updating consistency, though intuitive, is not sacrosanct: Multi-stage aggregation does seem to generate threshold effects, e.g. that whenever an agent has credence greater than .5 in a merely-ordinal theory that regards option O as most choiceworthy, then she is rationally required to choose option O, and these threshold effects can create violations of updating consistency, as MacAskill's case illustrates. But the multi-stage view explains why these threshold should exist and why we should expect updating consistency to sometimes be violated. Finally and most importantly, the costs of the multi-stage approach must be compared with the costs of structural enrichment and structural depletion, which strike me as considerably greater.

A second objection to multi-stage aggregation is as follows: It seems arbitrary that the fine-grainedness of theories or classes of theories taken as inputs to an aggregation method should effect the outcome of that aggregation method. To avoid such arbitrariness, we should therefore either limit ourselves to aggregation methods that are insensitive to "theory individuation" (that is, insensitive to how we group sets of fine-grained theories into coarser-grained theories) or else, whenever we are forced to rely on such methods, take only maximally fine-grained theories as input.

Two replies: (1) The coarse-graining of theories into comparability classes is not arbitrary. T_1 and T_2 are grouped together by the multi-stage aggregation procedure described above because they share common structure that T_3 and T_4 do not.

Dropping for a moment the technical sense of "theory" as a maximal consistent set of propositions, one might regard them as variants of a single theory, viz., consequentialism. More particularly, they are aggregated *before* the second, coarse-grained MFO stage of the aggregation procedure because they possess relevant structure (viz., cardinal structural) to which MFO is insensitive, while the binary-structured T_3 and T_4 do not.

(2) There is, as far as I can see, no way of satisfy the demands of this response without falling into the greater difficulties associated with structural enrichment or structural depletion: If there are theories that genuinely have only binary structure, then the only procedure available for aggregating such theories is MFO, which is sensitive to theory individuation.[22] But if we take maximally fine-grained theories as inputs to such a decision procedure, then we are left with fine-grained MFO or some other method that disregards cardinal information entirely and therefore precludes moral hedging.[23]

I have no general proof to offer that structural enrichment, structural depletion, and multi-stage aggregation are the only options for solving the problem of

[22] Likewise, any ordinary voting method that might be used to aggregate merely-ordinal theories will be sensitive to theory individuation—for instance, since Borda counting in a two-option case is equivalent to MFO, a probability-weighted Borda count of T_1, T_2, T_3, and T_4 will choose O_1 while a probability-weighted Borda count of C_1, T_3, and T_4 will choose O_2. MFO itself might be used as a method of aggregating ordinal as well as binary theories, since it is the analogue of *approval voting* in the social choice context.

[23] Note again that the "standard" version of MFO is not maximally fine-grained, since it aggregates by non-MFO-like procedures *within* moral theories (e.g., aggregating different possible consequences of a course of action expectationally within consequentialist theories) before engaging in MFO-like aggregation *across* theories. This, I have argued, is at least as arbitrary a form of coarse-graining as that involved in multi-stage aggregation. But the alternative, for the principled advocate of fine-grained MFO, is a complete insensitivity to scale of risk: For instance, if an agent A is almost certain that utilitarianism is true, and must choose between options O and P, where O has a .51 probability of producing slightly bettter consequences than P but a .49 probability of producing catastrophically worse consequences (say, ending all life on Earth), maximally fine-grained MFO requires A to choose O, which is quite implausible.

structural diversity. Nevertheless, in the absence of a better options, and given the difficulties of structural enrichment and structural depletion, I conclude that multi-stage aggregation is the right approach to take: Different classes of moral theories with differently structured value scales should be aggregated separately by methods appropriate to the shared structure of each class, and then aggregated with one another in ways that make use of only the (perhaps very minimal) structure that all the classes being aggregated have in common. I have only described one simple example of how such a multi-stage aggregation procedure might go—in more complex cases, the procedure might involve several stages, and many distinct comparability classes of theories (individuated, for instance, by cardinal vs. ordinal structure, perhaps interval vs. ratio structure, transfinite structure, dimensionality, etc) that are aggregated independently before the final stage.

6.3 Scale Normalizability

When the value scales of two theories share the same structure, or can be justifiably re-represented as sharing the same structure (by enriching and/or depleting the structures of one or both theories, either straightaway or in the course of a multi-stage procedure), the next question is how to *normalize* those scales—that is, which points and/or intervals on one scale correspond to which points and/or intervals on the other. Consider, for instance, hedonistic and preference utilitarianism, two straightforward maximizing consequentialist theories that agree on every feature of morality, except that hedonistic utilitarianism regards pleasure and pain as the sole non-derivative bearers of moral value while preference utilitarianism regards satisfied and dissatisfied preferences as the sole non-derivative bearers of moral value. Both theories, we may stipulate, have the same cardinal structure. But this structure does not answer the crucial question for expectational reasoning, how the value of a hedon according to hedonic utilitarianism compares to the value of a preference utile according to preference utilitarianism—that is, for an agent who divides her beliefs equally between the two theories and wishes to hedge when they conflict, how much hedonic experience does it take to offset the dissatisfaction of a preference of a given strength (or vice versa)? Likewise, of course, in trolley problem situations that pit consequentialist and deontological theories against one another, even if we could overcome the apparent structural incompatibility of these rival theories, the thorniest question seems to be: How many net lives must be saved, according to some particular version of consequentialism, to offset the wrongness of killing an innocent person, according to some particular version of deontology? It is this problem of scale normalization that has attracted most of the attention in the extant literature on intertheoretic value comparisons.

In fact, there are several distinct challenges involved in scale normalization that are rarely distinguished in the literature. Though I doubt that this accounting is exhaustive, three problems of particular importance are *unit normalization*, *level normalization*, and *dimensional normalization. Unit normalization* consists in a mapping between *intervals* on the value scales of two theories that designates pairs of intervals as *equal in magnitude*—for instance, designating any interval of one hedon on the value scale of hedonistic utilitarianism as equal in magnitude to any interval of three preference utiles on the value scale of preference utilitarianism. *Level normalization* consists in a mapping between *points* on the value scales of two theories that designates pairs of points as *equal in absolute value*—for instance, designating simple omissions (e.g., refraining from any bodily motion) as having the same value according to a consequentialist and a deontological theory. *Dimensional normalization* consists in a mapping between dimensions of the value scales of two multi-dimensional (i.e., lexical or incomparabilist) theories—for instance, mapping the sole dimension of a hedonistic utilitarian value scale onto one of the dimensions of a "leximin" egalitarian theory (on which the wellbeing of less well off social groups is represented by lexically prioritized dimensions on a multi-dimensional value scale).[24]

These components of scale normalization are not independent, of course. Where multi-dimensional theories are in play, unit or level normalization requires dimensional normalization. Given two theories with straightforward cardinal structure, establishing two or more points of level normalization between the theories (i.e., two or more pairs of corresponding points on the value scales of the two theories

[24] MacAskill (2014) discusses another such case: Suppose I divide my beliefs between classical utilitarianism and a lexically structured deontological theory on which certain moral reasons (e.g. to keep one's promises or refrain from telling lies) are lexically stronger than other kinds of reasons (e.g. prudential reasons or reasons of beneficence). Suppose further that both dimensions of the deontological theory's value scale have the same cardinal structure as the single dimension of the utilitarian theory's value scale—in order to represent the judgment, for instance, that telling two lies is twice as bad as telling one lie. In attempting to normalize the utilitarian and deontological value scales, then, the first question we must ask is whether utilitarian moral values should be normalized with the lexically prioritized or the lexically deprioritized dimension of the deontological value scale. The latter approach may seem natural if one thinks of utilitarian reasons as corresponding to reasons of beneficence in the deontological schema, but this approach also has the counterintuitive consequence of giving potential deontological reasons, e.g. for promise-keeping, absolute priority over all utilitarian moral considerations under moral uncertainty, even if my credence in deontology is quite low. MacAskill claims, and I agree, that it is non-obvious which dimension of the deontological value scale the utilitarian value scale ought to be normalized with in this case, and suggests that an agent in such circumstances should be represented as dividing her beliefs between two versions of utilitarianism that normalize with the upper and lower dimensions of the deontological value scale respectively (MacAskill, 2014, pp. 47-50).

that are equal in absolute value) will suffice to establish unit normalization as well. But the various aspects of normalization can also come apart in interesting ways. For instance, in the Easy Case from Chapter 5, the hedonic element common to Alice's hedonistic and pluralistic theories allowed us to establish unit normalization between the theories (both theories assigning equal value to an increment of one hedon) without in any obvious way yielding a level normalization of the two theories. On the other hand, one plausible way of establishing a partial level normalization between (at least) aggregative consequentialist theories is to hold that all such theories assign the same value (namely, 0) to an "empty universe" that contains no value-bearers at all. If true, this establishes *one* point of level normalization between apparently disparate consequentialist theories like hedonistic and preference utilitarianism, but in the absence of another such point, does not yield unit normalization.

The significance of these various forms of normalization depends to some extent on what decision procedures we use to aggregate theories. Risk-neutral expectational reasoning requires only unit normalization, not level normalization.[25] Risk-weighted expectational reasoning (whether risk-seeking or risk-averse) requires both unit and level normalization.[26] Stochastic dominance reasoning, on the other hand, requires

[25] This is equivalent to saying that it requires only *interval-scale comparability*, i.e., comparability with respect to the interval-scale properties of the theories' value scales, without regard to any possible ratio-scale properties.

[26] Suppose A has equal credence in two theories, T_1 and T_2, and is faced with two options, O and P. T_1 assigns O a value of x and P a value of $x + 5$. T_2 assigns O a value of y and P a value of $y - 5$. Assume that these values are unit-normalized, so that the difference in value between O and P is the same according to the two theories, though they disagree about which is better. Now suppose that A wishes to hedge for her moral uncertainty, not by simply maximizing expected value, but by maximizing *risk-weighted* expected value, with a slight aversion to risk. To do this she will need to make level comparisons as well as unit comparisons between T_1 and T_2: If $x = y$, for instance, then O is risk-free while P is risky, so A should choose O. But if $x = y - 10$, then O

only level normalization—and though in the context of cardinal theories, the full level normalization required for stochastic dominance implies unit normalization as well, this point becomes relevant in the context of ordinal theories for which unit normalization is impossible (*pace* MacAskill's Borda counting approach) but level normalization is possible.[27]

Beginning in the last chapter, I have argued for the idea of *content-based* aggregation, including content-based normalization: e.g., that the hedonistic and pluralistic theories in the Easy Case can be unit-normalized in virtue of their shared understanding of hedonic value, or that hedonistic and preference-based utilitarianism can be level-normalized at a shared zero (the empty universe) based on the the aggregative character of each theory. But even if such an approach is correct in the cases so far described, it is unclear how far this approach can generalize: e.g., can anything in the content of classical utilitarianism and Kantianism provide a basis for an adequate normalization of their respective value scales? Treating this as an open question, there are several candidates for a general approach to the normalization problem.

1. **Universal Content-Based Normalization (UCBN)** First appearances notwithstanding, there are in fact hidden bases for content-based comparisons between even such disparate moral theories like Kantianism and utilitarianism—bases analogous to, but presumably much subtler than, the shared understand-

is risky while P is risk-free, so A should choose P.

[27] Thus, while in the context of cardinal theories the most salient problem seems to be that of unit normalization (given a background commitment to risk-neutral expectational reasoning), in the context of ordinal theories the most salient problem is likely to be level normalization.

ing of hedonic value that provided a basis for normalization in the Easy Case.

2. **Incomparability** In some cases (like, perhaps, Kantianism and utilitarianism), normalization is simply impossible. Rather than attempting to solve the normalization problem in these cases, we should instead ask what decision procedures are appropriate in the face of incomparability.

3. **Binary Normalization** Sufficiently disparate theories can be normalized only in the most coarse-grained manner: namely, by treating *permissibility* and *impermissibility* as equivalent designations across all theories, and thus as providing a very weak form of level normalization. Thus, as per the structural depletion and multi-stage views discussed in the last section, the aggregation of a maximally diverse set of moral theories must in one way or another involve some form of MFO (or threshold procedure), since these are the only options left open by the minimal form of normalization possible among a full range of moral theories.

4. **Top-Down Normalization** When nothing in the content of theories T_1 and T_2 grounds a particular normalization between them (and perhaps even when the content of T_1 and T_2 *seems* to ground a particular normalization), the theories should be normalized according to some general, content-independent normalization principle like Lockhart's PEMT.

As in the last section, our task is now to spell out and compare these options. My conclusions with respect to the normalization problem will remain more agnostic, however, than my conclusions with respect to the problem of structural diversity.

The picture that I will argue for is this: (i) There is at least some reason to hope that the UCBN strategy will succeed, i.e., that a sufficiently careful examination of a rational agent's credal state will generally, and perhaps always, reveal bases for some form of normalization between any two normative theories in which she has positive credence. Even if content-based normalizability proves to be not-quite-universal, its failures might still be confined to a small enough number of cases that incomparability/binary normalization (and hence, some decision procedure like MFT, MFO, or a threshold view) would yield intuitive acceptable results if taken as the final stage of a multi-stage decision procedure, to be invoked only when content-based normalization gives out. (ii) Given this hope, and our intuitions in favor of moral hedging even between very disparate theories like Kantianism and utilitarianism, we should be reluctant to adopt incomparability or binary normalization as general responses to the normalization problem (i.e., as the default state of affairs once we move beyond minimal comparability classes). (iii) If our desire to vindicate hedging even in difficult cases is strong, and our progress in finding plausible content-based grounds for normalization in those difficult cases turns out to be disappointing, then we may have reason to adopt a top-down normalization principle. The most plausible such principles can be understood as explications of what MacAskill calls the "principle of equal say" (PES) (MacAskill, 2014, pp. 101ff), according to which the weight or influence of a moral theory on an agent's choices should be in some sense proportionate to her credence in that theory (at least, where no other basis for normalization exists). However, though it has some intuitive appeal, PES is less well motivated than content-based normalization methods like that illustrated by the Easy Case. Additionally, as we will see, there are many possible precisifications of the idea of "equal say," varying along at least two dimensions, and the apparent arbitrarity of any particular choice of top-down normalization method further diminishes the attractiveness of this approach. Thus, I will tentatively conclude that we should prefer to extend the reach of content-based normalization methods insofar as this proves possible, but that it is worth continuing to investigate top-down normalization methods as well in case content-based normalization proves unable to yield intuitively satisfactory results in too great a number of difficult cases.

6.3.1 Extending Content-Based Normalization

The Easy Case from Chapter 5 illustrates a simple basis for normalization that we may call *content overlap*: Alice's hedonistic T_1 and pluralistic T_2 include all the same propositions concerning the nature, explanation, and grounding of hedonic value, and from this we are able to conclude that they assign the same degree of value to a given unit of hedonic experience.[28]

On its face, content overlap does not seem to solve much of the normalization problem, since most pairs of moral theories do not seem to overlap in the neces-

[28] See Appendix B for a more complete explanation of the Easy Case and in particular of how Alice's doxastic state grounds the intuitive normalization of her two moral theories.

In conversation, William MacAskill has described the approach that this case exemplifies as a "shared rightmakers" approach. But it is important to remember, as I noted in Chapter 5, that the correctness of normalizing Alice's theories at the value of a hedon depends on their agreeing, not just on the *identity* of the rightmakers for certain acts (viz., the hedonic consequences of acts that have no positive or negative aesthetic consequences) but also on the *explanation* of their status as rightmakers (e.g., whether the moral significance of hedonic experience is a basic, mind-independent evaluative fact or a consequence of conventions within some particular moral community). When two theories both treat hedonic consequences as rightmakers, but give different explanations of their status as rightmakers, there is no longer any reason to normalize the theories at the value of a hedon.

sary way, sharing a complete explanation of some category of moral considerations or reasons that can serve as a basis for normalization. But there are a few ways in which the content overlap approach can be generalized. The first is by finding pluralistic theories to serve as "bridges" between non-overlapping theories. This is illustrated by a simple variant of the Easy Case in which Alice has some credence in a monistic theory that only values aesthetic goods, alongside her hedonistic and pluralistic theories. Given that the pluralistic theory fully overlaps with the hedonistic theory with respect to hedonic goods, and fully overlaps with the aesthetic theory regarding aesthetic goods, its value scale can be normalized with each of these monistic theories, and by extension, the monistic theories can be normalized with one another.[29] By the same sort of reasoning, an agent who has positive credence in both hedonistic and preference utilitarianism might have content-based grounds for a particular normalization of the two theories, if she *also* has positive credence in a pluralistic theory that attaches moral significance both to hedonic experience and the satisfaction of preferences and that regards these goods as comparable with one another.

In principle, the bridging approach seems to have extremely wide applicability. Let's introduce a new piece of terminology and call the propositions that identity and explain a particular feature of the world (like hedonic experience or the non-universalizability of an agent's maxims) as a source of moral value or reasons a *value claim*. Any non-deflationary moral theory (i.e., anything other than nihilism, error theory, and the like) will include one or more such value claims. For any

[29] See Appendix §B.4.1 for a more complete defense of this sort of "indirect comparability."

pair of theories T_1 and T_2, if T_1 includes a value claim C_1 and T_2 includes a value claim C_2 such that C_1 and C_2 are logically compatible (i.e., don't jointly imply a contradiction), then there is some logically possible pluralistic theory T_3 that includes their conjunction, $C_1 \wedge C_2$. In fact, there are almost certain to be many such theories, characterized by different exchange rates between C_1-value and C_2-value. But if an agent can aggregate her credence in such theories (i.e., her credences conditional on $C_1 \wedge C_2$) to yield a single exchange rate between the two sorts of value, then this may enable comparisons between T_1 and T_2 by providing a basis, in the agent's doxastic state, for normalizing their value scales (at the aggregated exchange rate between C_1-value and C_2-value).[30]

There are a few reasons to worry about how far this approach can really get us, however. First, some value claims may simply be logically incompatible, such that no consistent pluralistic combination of them is possible. For my part, I find it hard to think of many plausible cases of outright incompatibility between value claims, but as a toy example, consider the Kantian value of autonomy and the Buddhist value of recognizing the non-existence of the self (*anattā*, one of the seven "beneficial perceptions" (Bhikkhu, 2013)). If the Kantian explanation for the value of autonomy presupposes the existence of the self and the Buddhist explanation for the value of anattā presupposes the nonexistence of the self (and if we set aside the possibility that this is a merely-verbal dispute), then no pluralistic combination of the two value claims is possible.[31]

[30] See §B.4.2 for an explanation of how such aggregation over possible normalizations of two theories might be carried out.

[31] Nonetheless, indirect normalization between theories that value Kantian autonomy and theories that value the recognition of anattā might still be possible, e.g. if each of these value claims

Second, for some value claims, the most plausible pluralistic theories that combine them may treat them as incomparable. For instance, if one were persuaded by the positive value claims of both Nietzschean perfectionism and classical utilitarianism, one might do best to conclude that these two value schemes represent equally valid but irreconcilable normative perspectives rather than weighable inputs to a single, cardinal value function (a conclusion along the lines of Sidgwick's (1874) "dualism of practical reason").

Third and finally, the most plausible combination of certain value claims might be lexical. Certain value claims involve, in their self-justificatory content, assertions of lexical priority over other sorts of value claims. The Kantian explanation of the value of conformity with the categorical imperative, for instance, seems to essentially involve an assertion of lexical priority over all instrumental reasons ("hypothetical imperatives") and hence over the value claim of any moral theory whose value claims reduce to instrumental reasons (e.g. the contractarianism of Hobbes (1651) and Gauthier (1986) or the relativism of Harman (1975)). Thus, someone who was fully convinced by all the positive claims in both the *Groundwork* and *Morals by Agreement*, and therefore accepted a pluralistic hybrid of Kant's moral theory and Gauthier's, would presumably conclude that Kantian reasons, where operative, take lexical priority over contractarian moral reasons (as well as other instrumental reasons). The value claims of a divine command theory of morality, plausibly, involve claims to lexical priority over all other value claims whatsoever, so that if any

figures in a different pluralistic theory along with come third value claim C_3 that allows the two theories and hence the incompatible Kantian and Buddhist values to be normalized.

pluralistic hybridization of divine command theory with other ethical perspectives is logically possible, the result would be a lexically structured theory within which God's commands have absolute force.

The first two of these worries (incompatibility and incomparability) suggest that the bridging approach may have limited reach. The third worry suggests a broader objection, that content-based normalization approach in general is vulnerable to *fanaticism*. Suppose we conclude that a pluralistic hybrid of Kantianism and contractarianism would give lexical priority to Kantianism, and on this basis conclude that an agent who has positive credence in Kantianism, contractarianism, and this pluralistic hybrid ought to give lexical priority to Kantianism as well. This may seem intuitively implausible—one is tempted to say, it seems "unfair" to contractarianism to grant it no weight at all when it conflicts with Kantianism. Certainly it is intuitively implausible that an agent who has .999 credence in contractarianism and .001 credence in Kantianism, or divine command theory, or some other similarly self-assured moral theory, should give absolute priority to the theory she regards as almost certainly false over the theory she regards as almost certainly true.

I am willing to bite the bullet on this objection, up to a point: Some value claims may simply be more intrinsically weighty than others, and in some cases absolutely so. In cases where the agent's credence in the lexically prioritized value claim approaches zero, however, the situation begins to resemble Pascal's Wager (Pascal, 1669), the St. Petersburg Lottery (Bernoulli, 1738), and similar cases of extreme probabilities and magnitudes that bedevil decision theory in the context of merely empirical uncertainty. It is reasonable to hope, then, that the correct decision-theoretic solution to these problems (e.g. a dismissal of "rationally negligible probabilities" (Smith, 2014, 2016) or general rational permission for non-neutral risk attitudes (Buchak, 2013)) will blunt the force of the fanaticism objection.

Another way of extending the content overlap approach to normalization, apart from pluralistic bridging, is to consider the "sub-moral" content of moral theories: that is, the underlying theories of reasons or normativity on which they are built. One region of logical space where this idea seems particularly promising is the Humean theory of reasons and the various moral/normative theories built on top of it. These latter include the aforementioned views of Hobbes, Gauthier, and Harman, many "commonsense" moral views like that of Bernard Williams, Philippa Foot's erstwhile "hypothetical imperatives" view (Foot, 1972), and most anti-realist or "quasi-realist" approaches to morality. Indeed, if I was right to claim in Chapter 1 that the Humean and enkratic conceptions of rationality exhaust the space of possibilities, and if (as seems plausible) the Humean view commands at least as many adherents as the enkratic view, then a *majority* of the moral theories actually embraced by philosophers may turn out to be, in the last analysis, Humean theories, in which an agent's moral reasons derive in one way or another from the contents of her motivational set.

There is clear potential for a general content-based normalization of Humean moral theories, since all such theories share a "common currency," viz., Humean reasons and the desires/motivational states from which such reasons arise. *Ceteris paribus*, it seems reasonable to normalize such theories by assuming that they all assign the same degree of normative significance to the direct satisfaction of a given non-moral desire (at least, when it neither conflicts with nor results in the satisfaction of any other element of the agent's motivational set). Since Humean moral theories presumably will not treat moral reasons as lexically or infinitely stronger than non-moral reasons, normalizing these theories at the strength of the non-moral reasons about which they have no disagreement should yield a satisfactory normalization of their moral value scales as well. For instance, suppose an agent A divides her beliefs between Gauthier's contractarianism and Harman's relativism, and faces a situation in which a Gauthier-norm N conflicts with a Harman-norm M. Suppose that Gauthier's view and Harman's view are in full agreement, say, about the strength of A's entirely non-moral reason to accept a free ice cream cone. And suppose the contractarian view says that her reasons to comply with N in this situation are 150 times stronger than her reason to accept a free ice cream cone, while the relativist view says that her reasons to comply with M in this situation are just 50 times stronger than her reasons to accept a free ice cream cone. Then it appears that the value of complying with norm N according to contractarianism is three times the value of complying with norm M according to relativism.[32]

[32] In some cases of intra-Humean disagreement, it may seem that an apparent disagreement between "moral theories" is actually an empirical disagreement, namely, a simple disagreement about what course of action would best satisfy an agent's desires. In some simple cases this may be true: A simpleminded Humean moral theory can be constructed from the Humean theory of reasons plus the observation that a typical human being has some broadly moral motivations, like empathy, which on the Humean theory of reasons she can some reason to satisfy. "Moral" disagreements can then arise from rival theories of moral psychology, i.e., of what the exact content of those moral motivations or concerns amounts to. But many intra-Humean disagreements are genuinely normative, in that they involve disagreements about exactly how an agent's reasons relate to her desires and other motivational states (e.g. what forms of practical deliberation from an agent's existing motivational state result in the *discovery of preexisting reasons*, as opposed to alterations in an agent's motivational state that simply give her new reasons she didn't have before) or about the relationship between *rationality* and objective reasons (e.g. the question of "straightforward maximizing" vs. "constrained maximizing" taken up by Gauthier (1986)).

Can we say anything similar about non-Humean views—i.e., is there some "common currency," analogous to the currency of desire, in which non-Humean views must trade, that would enable a general content-based normalization of their value scales? This is non-obvious. I claimed in Chapter 1 that the Humean view identifies objective reasons with the satisfaction of desires or other motivational states while the enkratic view identifies objective reasons with goodness, value, or choiceworthiness, where these concepts are at least not *conceptually* linked to desire satisfaction. To simplify, let's say that enkratic views identify objective reasons with *desire-independent choiceworthiness.* On face, this conception of reasons does not offer the same potential for a general normalization of moral theories as the Humean conception. For while the strength of desires is an empirically measurable feature of the world, and therefore all normative theories that identify reasons with desires and strength of reasons with strength of desires can be put on a common scale, there is no similar way of measuring degrees of desire-independent choiceworthiness that would generate a common scale of reason strength for non-Humean theories.

One strategy that might allow a general normalization of non-Humean theories is suggested by MacAskill (2014). MacAskill argues that all moral theories can be understood as having as the range of their value assignments the same "universal scale" of degrees of value. He explicates this idea of a universal scale by analogy with the metaphysics of quantity: There is a debate among metaphysicians as to whether mass properties (*5 kg* etc) or mass relations (*more massive than*, *less massive than*, *equally as massive as*) are more fundamental. One reason to prefer the former, "absolutist" position is its ability to make sense of intuitive trans-world mass relations (e.g., the possibility of a world identical to ours in every respect except that everything is twice as massive). Similarly, MacAskill suggests, if we suppose that there is a scale of absolute value properties analogous to the mass-absolutist's scale of absolute mass properties, then those value properties (being abstract entities) exist at all possible worlds, including the possible worlds postulated by rival moral theories. Thus, it seems, all moral theories should be comparable by virtue of the comparability of the absolute value properties (all belonging to a single "universal scale") instantiated at the corresponding possible worlds.

To my mind, however, MacAskill's proposal moves too quickly from an *ontological* claim about the existence of certain properties to a *semantic* claim that all moral theories can be understood as making reference to those properties. Let's suppose that practical options in the actual world instantiate absolute value properties arranged on, say, an interval scale. Does this mean that *all moral theories*, including merely-ordinal theories, multidimensional theories, hyperreal theories, etc, in fact assign interval-scale degrees of choiceworthiness to options, despite their own protestations to the contrary? Presumably not. Even given that a particular set of interval-scale value properties are the ones instantiated in the actual world, it is still possible for a (mistaken) moral theory to attribute a *different* set of value properties to options, properties that are *not* instantiated in the actual world. Further, as MacAskill acknowledges (pp. 154-5), some moral theories may be explicitly comparativist, and not attribute any absolute value properties at all. Finally, even a moral theory that attributes absolute, interval-structured value properties to options may not attribute properties from *the same* interval scale as those properties that are instantiated in the actual world—that is, mere structural agreement between theories does not obviously guarantee that the theories make reference to the same scale or that their scales can be normalized.[33] Of course, we might always insist that any theory that does *not* attribute absolute value properties from the particular scale instantiated at the actual world simply doesn't count as a moral theory, but given the structural diversity of (apparent) moral theories, if nothing else, this stipulation would force us to implausibly exclude the large majority of what we take to be moral theories.[34]

[33] The analogy to mass properties is importantly misleading in this respect. Suppose we accept the absolutist theory of mass properties, and now wish to know whether "intertheoretic mass comparisons" are possible: For instance, can we compare the mass attributed to a given object by the Aristotelian theory of mass with the mass attributed to the same object by the Einsteinian theory of mass? The answer seems to be "yes," since we are able to make uncontroversial translations between Aristotelian and Einsteinian mass ascriptions (e.g., 1 obol = .72 grams). But these translations assume a particular normalization of the Aristotelian and Einsteinian mass scales: that both theories *agree on* the mass of a certain class of objects, e.g. small metal coins. (The Aristotelian measures the coin and concludes that it has a mass on 1 obol, the Einsteinian measures the coin and concludes that it has a mass of .72 grams, and since we assume that they don't disagree about how massive the coin is, we translate obols to grams at a rate of 1:.72.) But suppose we took as our reference point, not small metal coins, but air-inflated party balloons. This would give us a different normalization of the two mass scales. (The Aristotelian measures the balloon and concludes that it has a mass of 1 obol, the Einsteinian measure the balloon and concludes that it has a mass of, say, 2 grams.) If we assume that the Aristotelian and the Einsteinian agreed about the mass of air-inflated party balloons, then we have to conclude that they *disagree* about the mass of small metal coins, and vice versa. Intertheoretic mass comparisons between the Aristotelian and Einsteinian theories seem possible only because of the contingent fact that there is one particularly natural way of normalizing the scales: namely, by assuming that they agree on the mass of small metal coins and other similarly much-denser-than-air objects. Although I have argued that there is *often* a similarly natural way of normalizing value scales in a moral context (e.g., assuming that Alice's hedonistic and pluralistic theories agree on the value of a hedon in the Easy Case), the mass analogy should not convince us that such a natural normalization method exists in *all* cases of intertheoretic moral uncertainty. Just as the possibility of intertheoretic mass comparisons does not follow *simply* from the existence of absolute mass properties, so the possibility of intertheoretic value comparisons would not follow *simply* from the existence of absolute value properties.

[34] The argument for the universal scale view, as MacAskill presents it, relies on the claim that absolute value properties, *qua* abstract objects, must exist at all possible worlds. As he subsequently acknowledges, the possible worlds corresponding to false moral theories are presumably *epistemically* but not *metaphysically* possible worlds, and it is unclear why we should think that all actual abstract entities exist at all epistemically possible worlds. (In fact, this seems clearly wrong, since for any abstract entity, I can rationally have non-zero credence that it does not exist, i.e., assign some credence to epistemically possible worlds where it does not exist.) MacAskill interprets this point as an objection to absolutism about value properties, but this strikes me as the wrong lesson. Rather, the point is that even if practical options instantiate absolute (i.e., unary,

It is unclear, therefore, why accepting (as a first-order matter) the existence of a scale of absolute value properties should lead us to conclude that all or even most moral theories assign values on that scale. Nevertheless, the idea that *many* (though not all) moral theories might make reference to the same set of abstract entities (absolute value properties arranged on a value scale with a particular structure) does suggest the possibility of content-based normalizations that go beyond the overlapping-content form of normalization illustrated by the Easy Case. It seems to me that this approach is more promising with respect to non-Humean theories than with respect to Humean theories, since it is more natural to understand Humean theories as comparativist rather than absolutist (especially in light of their agent-relative character), but it is at least conceivable that the universal scale approach might allow us to bridge the gap between (some) Humean and non-Humean theories as well.[35]

I have so far explored some positive conclusions that might be drawn about the potential for extending content-based normalization beyond small comparability classes like the one illustrated by the Easy Case. Are there any general *negative* conclusions we can draw with confidence—i.e., definite barriers to content-based normalization? The clearest route to such negative conclusions, as far as I can see,

non-relational) value properties at the actual world, those properties need not feature in every moral theory—both because they may not *exist* at the epistemically possible worlds corresponding to certain moral theories, but also because, even if they exist at a given possible world, they may not be *instantiated* at that world.

[35] MacAskill concedes that the universal scale approach will not establish comparability between all moral theories, but claims that it establishes comparability between all theories that accept an absolutist rather than comparativist understanding of value (MacAskill, 2014, p. 155). But the arguments I have given above suggest that this is too optimistic, since absolutist moral theories may attribute *different* sets of absolute value properties, from different and potentially incomparable value scales.

is the hybridization test, that is, looking for pairs of value claims that are either logically incompatible or that seem to yield incomparability when combined to form a pluralistic theory. As I have said, I don't think there are many *clear* examples of either kind, but I will close this discussion by noting the possibility of one very general barrier to normalizability: the boundary between Humean and enkratic moral theories. In Chapter 1, I postponed the question whether the Humean and enkratic views offer rival *conceptions* of shared underlying concepts of reasons, rationality, etc, or simply attach labels like "reasons" and "rationality" to fundamentally different *concepts*. Now as then, I won't attempt to answer this question, but I raise the question in order to observe that, to the extent one understands these views as dealing in entirely different concepts, one might expect that content-based normalization between the two categories of theory will prove impossible. For if Humean, desire-based "reasons" and enkratic, desire-independent "reasons" are not both articulations of *some* underlying normative concept, then what sense could it make to compare them? To say that "My reason to choose option O according to [enkratic moral theory] T_1 is twice as strong as my reason to choose option P according to [Humean moral theory] T_2" would be a simple category mistake, on a par with saying "The gravity of the situation on the Korean peninsula is twice as great as the surface gravity of Ganymede." That is, quantitative comparisons only make sense when there are two quantities (or possible quantities) *of the same something* to be compared, and unless there is some underlying concept of reasons, normativity, or the like that unites anything labeled a "moral theory," then there will be some insuperable barriers to meaningful intertheoretic comparison.[36]

As the preceding discussion has illustrated, content-based normalization and the comparability class approach to decision-making under moral uncertainty seems to create a patchwork theory in which aggregation procedures for particular classes of theories depend, perhaps in non-obvious ways, on the content of those theories—not just their value assignments, but also the *arguments* or *explanations* they offer for those value assignments, the metaethical and other philosophical premises on which those arguments rely, and perhaps other sorts of propositions that might figure in a complete theory of the world, in ways that we have not yet anticipated. This contrasts to the neat simplicity of a principle like Lockhart's PEMT. But such simplicity is purchased at the price of a Procrustean indifference to the actual content of the moral theories being compared and aggregated, a desire to fit those theories into an attractive formalism even if this requires misrepresenting some of their features (e.g., representing ordinal structure as cardinal) and ignoring others (e.g., ignoring the overlapping content that produces natural normalizations at odds with the general normalization principle). Thus it seems to me that content-based normalizations should be adopted wherever possible, even if it turns out that many sets of moral theories *cannot* be normalized by reference to their content. What

[36] This consideration raises a possibility, which I will not explore at any length, that the resolution of the debate between the Humean and enkratic theories of reasons should be taken as having "external" normative significance in Weatherson's sense: If the Humean and enkratic views simply attach the label of "reasons" to different concepts, then plausibly rationality does not require us to account for our uncertainty between the Humean and enkratic views, but simply to account for our uncertainties about our *reasons* (that is, the normative concept fundamental to whichever of the Humean or enkratic views turns out to be correct), while ignoring whatever beliefs we might have about *schmeasons* (the wrongheaded, perhaps incoherent concept that goes by the name of "reasons" in whichever of the Humean or enkratic views turns out to be mistaken).

we should do in these latter cases—in particular, whether we should simply accept defeat, or should adopt a top-down normalization procedure like PEMT—is the question to which we now turn.

6.3.2 Normalization Failure: Incomparability/Binary Normalization

One possibility suggested by the multi-stage aggregation procedure described in the last section is to aggregate comparability classes of theories up to the point where content-based comparison gives out, and then perform the last stage of the aggregation procedure by means of a norm like MFT or MFO that does not rely on any precise intertheoretic normalization. For simplicity, let's confine our attention to these two possibilities: multi-stage aggregation with either MFT or MFO as the final stage.

The latter procedure was introduced in the last section, where we distinguished this sort of "coarse-grained" MFO from the more general "fine-grained" view on which an agent should simply choose the option that has the highest probability of being objectively permissible. We should now draw a parallel distinction between "coarse-grained" and "fine-grained" MFT. *Fine-grained MFT* says that an agent should always choose the option preferred by the moral theory in which she has highest credence, where "moral theory" is understood in our familiar sense, as a maximal consistent set of moral propositions.[37] *Coarse-grained MFT* says that an

[37] It's important to clarify, in this context, that a moral theory as opposed to a theory of the world generally should be understood as a maximal consistent set of "purely moral" propositions, or of general moral principles, that does not include particular injunctions like "A objectively ought to choose option O at time t." Otherwise "moral theories" would embed empirical information, e.g. about which acts have the best consequences, resulting in a view more objectivist than the proponents of MFT typically intend. To say that a moral theory "prefers" a particular option O,

agent should always choose the option preferred by the *maximal comparability class* of moral theories in which she has highest credence, where a maximal comparability class is a set of theories among which some form of content-based comparison and aggregation are possible, but such that no form of content-based comparison or aggregation is possible between theories inside the set and any theories outside the set. This view says, in effect, "Aggregate theories as far as content-based comparisons will allow, and when no further content-based comparisons are possible, do what the most plausible comparability class of theories says, in aggregate."

Coarse-grained MFT is, arguably, an improvement on fine-grained MFT insofar as it offers a more plausible answer to the problem of theory individuation. Gustafsson and Torpman (2014), in the most thoroughgoing defense of MFT to date, embrace a view very close to what I have called "fine-grained MFT," claiming that we should "[r]egard moral theories T and T' as versions of the same moral theory if and only if you are certain that you will never face a situation where T and T' yield different prescriptions" (Gustafsson and Torpman, 2014, p. 13). But as MacAskill (2014) has pointed out, this has the implausible result that a typical agent ought to be guided exclusively by a moral theory in which she may have almost no credence. For instance, suppose an agent has 99% credence that some form of commonsense morality is correct and only 1% credence that classical utilitarianism is correct. Nevertheless, among possible commonsense moral views, she is uncertain of a great many particular, logically independent moral questions.

then, is to say that it implies that the agent has more subjective reason to choose O than to choose any alternative, given her non-moral belief state.

(Exactly how wrong is it to break a promise? Is it wrong to have an abortion at 5 months? How about at 4 months 28 days? Are my obligations to my children twice as strong as my obligations to distant strangers, or five times as strong, or what?) Hence, her credence in any *fully specified* version of commonsense morality is less than 1%. Classical utilitarianism, meanwhile, is a fully specified theory to begin with, i.e., she has no uncertainties about which form of classical utilitarianism is true that might result in conflicting practical guidance. So on Gustafsson and Torpman's view, she ought to always do what classical utilitarianism recommends, even when *every* version of commonsense morality in which she has positive credence tells her to do the opposite. Coarse-grained MFT has the potential, at least, to correct this implausible result, for if some or all of the commonsense views in which she has positive credence admit of content-based comparisons (say, based on their overlapping content in the many moral cases about which they are in full agreement), then her "favorite" maximal comparability class will plausibly turn out to be a less-than-maximally-specified version of commonsense morality in which she has substantial (up to 99%) credence (assuming for the sake of argument, of course, that commonsense morality and classical utilitarianism turn out to be genuinely incomparable).

Coarse-grained MFT seems a natural option if one holds that, between some pairs of moral theories, no form of normalization whatsoever is possible. But there is also a more moderate position, holding that a kind of rough normalization is possible even between the most disparate moral theories: namely, that the shared concepts of *permissibility* and *impermissibility* allow us to treat these designations as equivalent across theories, where other forms of normalization fail. This claim, it seems to me, is implicit in any version of MFO, fine-grained or coarse-grained, since the basic principle of MFO is to sum up an agent's credence in theories (of whatever coarseness of grain) that treat a given option as *permissible*, giving equal weight to the designations of each theory. Absent something like a normalization claim, we might be left to wonder, for instance, why an option's permissibility by a more demanding moral theory (that regards only a few options as permissible) should not count for more than its permissibility according to an un-demanding moral theory (that regards many options as permissible). Thus, while this claim is not crucial to anything else in this chapter and I am not deeply attached to it, it seems to me that MFO involves at least a *kind* of normalization, namely, *binary* normalization of the designations *permissible* and *impermissible*.

So we have at least two relatively pessimistic options for what to do when content-based normalization proves impossible: (1) accept incomparability and use coarse-grained MFT as the last stage of a multi-stage aggregation procedure or (2) take binary normalization as the best we can get and use coarse-grained MFO as the last stage of a multi-stage aggregation procedure. Both of these views involve moral hedging—perhaps a great deal of hedging—within comparability classes, but simply declare that there are some kinds of moral uncertainty for which hedging is impossible. I have nothing to say in assessment of these views that I have not already said: Their chief drawback lies in our intuitive judgment that it is rationally appropriate to hedge even between widely disparate moral theories, e.g., to take account of the utilitarian stakes when one faces a choice situation like the trolley problem with beliefs divided between utilitarianism and Kantianism. Thus, we should at least continue to explore the potential for other views—content-based or top-down normalization approaches—to vindicate these intuitions.[38]

6.3.3 Top-down Normalization

If we conclude that there are some pairs of theories whose content does not ground any particular normalization between them, but between which hedging is nevertheless rationally appropriate, then we must have some method of normalization that does *not* derive from the content of the theories being normalized. As I have noted, the archetype of this sort of normalization method is Lockhart's PEMT, according to which in any choice situation, "[t]he maximum degrees of moral rightness of all possible actions...according to competing moral theories should be considered equal" and "[t]he minimum degrees of moral rightness...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)" (Lockhart, 2000, p. 84).

As I have also noted, PEMT is subject to apparently decisive objections (Sepielli, 2013). Nonetheless, it offers a useful template from which we can generate other, more plausible normalization principles. As MacAskill (2014) observes, there are two salient features of PEMT that can be varied to produce alternative

[38] It is worth noting that the multi-stage aggregation approach that I defended as a response to the problem of structural diversity does not commit us to accepting one of these defeatist views: We might hold that some form of hedging, or at least something richer than MFT/MFO, is possible even at the last stage of the aggregation procedure. For instance, we might hold that the last stage of aggregation combines cardinal, ordinal, and binary theories using a voting method richer than MFO, e.g. a Condorcet method or (*pace* my objections) a Borda count.

views.[39] First, it considers only the options available to a single agent in a single choice situation. Second, it normalizes the value assignments of rival moral theories in that choice situation by treating the *ranges* of those value assignments (i.e., the difference in value between the best and worst options) as equal. We can then construct a wide variety of alternative principles for normalizing rival value assignments by varying either the *set of options* relative to which the normalization method is defined (i.e., the domain of the value assignments being normalized) or the *feature of the value assignments* to be equalized. Let's call this first feature the *scope* of a normalization method and the second feature its *normalization point*. Scope and normalization point, then, are two dimensions along which top-down normalization methods can vary.

Potential values of the scope variable include the following:

1. The options that make up a single choice situation (as in Lockhart's PEMT).

2. All conceivable options over which a theory's value assignment is defined (as in Sepielli's "Conceivability PEMT").

3. The options that make up all (past/future/lifetime) choice situations faced by a particular agent.

4. The options that make up all (past/future/past-present-and-future) choice situations faced by *all* actual agents.

Possible normalization points include:

[39] The following discussion is heavily indebted to MacAskill's analysis.

1. The *range* of each theory's value assignment (as in Lockhart's PEMT and Sepielli's Conceivability PEMT).
2. The difference between the mean value of a theory's value assignment over the specified domain and either the *maximum* or *minimum* value of the assignment over that domain (approaches MacAskill calls *Max-Mean* and *Mean-Min*).
3. The *variance* of each theory's value assignment (the approach endorsed by MacAskill).
4. The *standard deviation* of each theory's value assignment.
5. The *mean absolute deviation* of each theory's value assignment, relative to some specified measure of central tendency (mode, median, arithmetic mean, geometric mean...).

Neither of these lists is exhaustive, of course. With respect to normalization points in particular, any possible measure of dispersion, including any measure of typical or maximal deviation from any measure of central tendency, constitutes a feature of value assignments that might be treated as equal to achieve a normalization of rival theories. Some combinations of scope and normalization point are open to fairly decisive objections, and can be easily set aside: For instance, any view with the extremely narrow scope of Lockhart's PEMT will have the consequence that normalizations between theories change from choice situation to choice situation and hence that an agent may be rationally required to choose a course of action that she knows in advance is strictly worse than some available alternative according to *every* moral theory in which she has positive credence (Sepielli, 2013). On the other hand, the extremely wide scope of Sepielli's Conceivability PEMT coupled with its use of range normalization yields disastrous consequences since for many theories (e.g. nearly any form of aggregative consequentialism), the range of their value assignments over all conceivable options is either undefined (if the theory regards all and only finite degrees of value as possible) or infinite (if the theory regards infinite degrees of value as possible). Nevertheless, there is no obvious reason to think that such decisive objections will narrow the set of plausible options down to one.[40]

This surfeit of possibilities is, paradoxically, an important reason for skepticism concerning top-down normalization approaches in general. The principal motivating idea behind any of these approaches is what MacAskill calls the "principle of equal say" (PES), the idea that an agent who assigns equal credence to two moral theories should give them equal weight or influence over her choices, or more generally that the weight or influence of a moral theory on an agent's choices should be in some sense proportionate to her credence in that theory (MacAskill, 2014, pp. 101ff). But as the preceding discussion illustrates, this principle is subject to a great many

[40] MacAskill argues for variance voting by a process of elimination, showing that range normalization, Max-Mean, and Mean-Max have highly counterintuitive consequences in certain cases. But variance normalization is not immune from such counterexamples: For instance, suppose an agent A has equal credence in two theories T_1 and T_2 and faces a set of ten options, O_{1-10}, of which both theories agree that only O_1 and O_2 are reasonable options. T_1 (on its pre-normalized value scale) assigns O_1 a value of 100, O_2 a value of 95, O_{3-9} values of 0, and O_{10} a value of $-10,000$. T_2 (on its prenormalized value scale) assigns O_1 a value of 95, O_2 a value of 100, and O_{3-10} values of 0. The two theories, then, are quite similar except that they disagree about which of O_1 and O_2 is more choiceworthy, and T_1 has some particular, extremely strong moral objection to O_{10} that leads it to assign an extreme negative value to this option. Because T_1's extreme assignment to O_{10} makes its variance, on the prenormalized value scales, much greater than that of T_2, MacAskill's variance normalization approach tells us that A should give greater weight to T_2's preference for O_2 over O_1 than to T_1's preference for O_1 over O_2, and is therefore rationally required to choose O_2. But given that T_1 and T_2 *both* regard O_{3-10} as obviously bad options, it is at least strange that the availability of O_{10}, and the disagreement between T_1 and T_2 about *how* bad this particular bad option is, should be the decisive factor in A's choice between O_1 and O_2.

interpretations of which none is uniquely salient or obviously preferable to the rest. Just as Bertrand-style paradoxes weaken the intuitive appeal of the principle of indifference in probability theory (Van Fraassen, 1989, pp. 303ff), so the multiplicity of "equalization" approaches to intertheoretic normalization weakens the intuitive appeal of the principle of equal say.

It is also unclear to what extent PES is well-motivated in the first place. As both Sepielli and MacAskill have pointed out, it is not as if I would treat a theory "unfairly" by giving it less weight than another theory in which I have equal credence (Sepielli, 2006, p. 602; MacAskill, 2014, p. 101). Since it seems at least intuitively clear that some first-order moral theories do in fact propose weightier moral reasons than others (say, divine command theory vs. Harman-style relativism), why should there be even a presumption of equal weight or equal say among theories? And in the absence of such a presumption, what principle could we appeal to in place of PES to motivate any particular top-down normalization principle?[41]

Finally, there are the various technical problems that afflict all extant interpretations of PES, including Lockhart's PEMT and MacAskill's variance normalization approach: sensitivity to irrelevant alternatives, an apparent inability to accommodate theories whose range/variance is infinite or undefined, and the difficulty of generalizing beyond simple cardinal structures (e.g. to ordinal or hyperreal struc-

[41] One option I haven't discussed, but that deserves mention here, is a *subjectivist* approach to intertheoretic normalization. On such an approach, at least when there are no content-based grounds for any particular normalization between a pair of theories, the agent should simply normalize the two theories in whatever consistent manner she chooses, or in whatever consistent manner seems most appropriate to her. On this view, what grounds the appropriateness of a particular normalization is a preference, a "seeming," or some other such mental state of the agent. This view may appeal to some, but of course will not satisfy those who think there is a more robust fact of the matter about how, say, utilitarianism and Kantianism ought to be normalized.

tures).[42]

For these reasons, I regard top-down normalization as less attractive than content-based normalization. Nevertheless, these worries are far from decisive, and given the difficulty of foreseeing how far content-based normalization can be taken along with the intuitive appeal of moral hedging, it is worth continuing to investigate top-down normalization approaches in the hope of finding one that is uniquely well-motivated and yields intuitively acceptable results. My own tentative view, then, is that there is reasonable hope of finding universal or near-universal content-based grounds for intertheoretic comparisons that reduce incomparability (or merely binary normalization) to an acceptable minimum, but that if this hope cannot be realized, then some top-down normalization principle will likely turn out to be the next most attractive option.[43]

[42] Lockhart's PEMT approach, as I have already said, simply assumes that all theories have or can be appropriately represented as having cardinal structure, and so seems incapable of genuinely accommodating ordinal structure. MacAskill attempts to make his Borda counting approach to ordinal aggregation continuous with his variance normalization approach to the aggregation of (not-otherwise-comparable) cardinal theories by handling ties in ordinal rankings in a way that amounts to normalizing ordinal theories "at the variance of their Borda scores" (MacAskill, 2014, p. 247), which I have argued amounts to tacit cardinalization as well. Any attempt at *unit* normalization of ordinal theories will, I think, be subject to the same objection. However, *level* normalization of ordinal theories—i.e., a normalization stipulating that a certain ordinal rank on the value scale of T_1 represents the same degree of absolute choiceworthiness as a certain ordinal rank on the value scale of T_2—does not involve any such implicit cardinalization, and is necessary if we wish to use statewise or stochastic dominance as principles of decision-making under uncertainty over merely ordinal theories. (An alternative is to rely on voting methods, e.g. Condorcet methods, that do not in any obvious way require either unit or level normalization of ordinal theories.)

[43] One approach to top-down normalization that I regard as particularly promising, though it has yet to be substantially developed or explored in the literature, eschews normalization points like range or variance in favor of a broadly game-theoretic approach on which moral theories are represented as agents with utilities corresponding to their value assignments and "bargaining power" of one kind or another corresponding to their subjective probabilities. One version of this approach is Bostrom's "parliamentary model," described in a brief online article (Bostrom, 2009), on which rational deliberation under moral uncertainty is modeled as a legislative body, with theories represented by a number of delegates proportionate to the agent's credence, and in which decisions are made by a deterministic voting method. (To give greater room for moral hedging in this model, Bostrom adds the somewhat ungraceful epicycle that the delegates in the moral parliament *believe* the voting method to be stochastic, so that even when a majority coalition favors a particular op-

6.4 Decision-Theoretic Compatibility

Combining the value assignments of different moral theories requires not only a means of reconciling their potentially diverse structures and normalizing their value scales but also an *aggregation procedure* or *decision procedure*: a procedure, like dominance reasoning or expected value maximization, that tells us *how to deal with uncertainty* over several different representations of the value of options, given that we know how to compare those representations.[44]

To show that disagreement over aggregation procedures (just like differences

tion, they have incentives to bargain for additional support, e.g. by offering their support in other choice situations to delegates who regard those choices as more urgent.) Other possibilities in a similar spirit include the use of axiomatic bargaining solutions, taking moral theories as bargaining agents with some appropriate, probability-sensitive outcome (say, MFO) as the threat point, or modeling a set of choice situations as a multi-item auction in which theories are endowed with capital proportionate to their subjective probabilities with which they can "bid" for control of particular choice situations. Such approaches have the advantage of at least largely avoiding the "irrelevant alternatives" problems to which all the normalization-point methods described above appear to be subject (see note 40), though this must be examined on a case by case basis (e.g., not all axiomatic bargaining solutions that have been defended in the literature satisfy the standard Independence of Irrelevant Alternatives axiom (Kalai and Smorodinsky, 1975)). Such game-theoretic approaches may also allow for a more natural integration of theories with unbounded value scales, theories that recognize infinite values, theories with merely ordinal structure, etc (insofar as agents with analogously structured utility functions can still engage in rationally guided, strategic interactions with other agents with variously-structure utility functions). For these reasons, these approaches strike me as more promising than Lockhart's PEMT or MacAskill's variance normalization.

On any parliamentary, bargaining-theoretic, auction-theoretic, or similar approach, the question of *scope* remains important, though all these approaches seem to require scope wider than a single choice situation if they are to yield interesting conclusions. My own hunch is that the most attractive top-down normalization approaches will take as their scope all the lifetime choices of either a single agent or all actual agents, and then model the deliberative process either as a cooperative bargaining situation with the preferred outcome given by some axiomatic bargaining solution or as a multi-item auction. Since I regard top-down normalization methods as a second-best option, however, I will not attempt to develop or evaluate such an approach here.

[44] It is potentially misleading to speak of "decision procedures" in the context of a multi-stage approach, because except at the final stage of the aggregation process, these procedures will not issue a *decision* but rather simply an aggregation of the reasons offered by some comparability class of moral theories. So it is perhaps more proper to call these rules "aggregation procedures." Nevertheless, the procedures themselves are the sorts of rules familiar from decision theory for weighing objective reasons that would exist given various states of the world, under conditions of uncertainty, to arrive at a decision: rules like statewise dominance, stochastic dominance, risk-neutral or risk-weighted expected value maximization, maximin, and so on.

	Credence	O_1 (risky)	O_2 (risk-free)
S_1	.9	0	100
S_2	.1	2500	100
T_1 (risk neutral)	.3	250	100
T_2 (risk averse)	.7	5	10

Table 6.3: George's Gamble, Pre-Normalization

in structure and the difficulties of normalization) presents a difficulty for intertheoretic value aggregation and hence for a complete theory of decision-making under moral uncertainty, let's begin with an example: Suppose that George is faced with two options, one risky and the other risk-free, and divides his belief between two moral theories, one of which is risk-neutral and the other of which is risk-averse—specifically, measuring the expected value of an option by the probability-weighted sum of squares roots of its possible objective values.[45]

Table 6.3 represents George's situation, where O_1 and O_2 are his options, T_1 and T_2 are moral theories, and S_1 and S_2 are possible states of the world. We see that T_1 prefers the risky option, since it has higher expected value, while T_2 prefers the risk-free option, which has higher risk-weighted expected value according to T_2's risk aversion function.

There is, to begin with, a normalization problem here: If we take T_2's square root risk aversion at face value, then T_1 will end up carrying much greater weight in the aggregated verdict of the two theories than T_1 (as can be seen in Table 6.3, where the difference in risk-weighted expected value between O_1 and O_2 is thirty

[45] This case is inspired in part by a similar case in Nissan-Rozen (2015).

	Credence	O_1 (risky)	O_2 (risk-free)
S_1	.9	0	100
S_2	.1	2500	100
T_1 (risk neutral)	.3	250	100
T_2 (risk averse)	.7	50	100

Table 6.4: George's Gamble, Normalized

times greater according to T_1 than T_2). Let's overcome this problem by stipulating the following: T_1 and T_2 are both egoistic ethical theories, and the values they assign to states of affairs represent the agent's lifetime utilities. T_2 is risk averse because it embraces a notion of *adequacy* according to which there is great value in approaching or achieving an adequate standard of wellbeing, but additional units of wellbeing become increasingly superfluous once one has already achieved an adequate level of lifetime wellbeing. The balancing point at which one's level of wellbeing is neither inadequate nor superfluous is 100—that is, for a life with 100 units of wellbeing, T_2's concern for adequacy is inert. Thus, T_1 and T_2 agree on the value of an option, like O_2, that guarantees 100 lifetime units of wellbeing. Table 6.4 re-represents George's situation according to this normalization.

So far so good, then: T_1 and T_2 are both ratio-scale theories, so we don't need to worry about structural diversity, and we have found a way of normalizing their value scales. But now a new problem presents itself: If we simply calculate risk-neutral expectations over T_1 and T_2, we find that $\text{EV}(O_1) = (.3 \times 250) + (.7 \times 50) = 75 + 35 = 110$ while $\text{EV}(O_2) = (.3 \times 100) + (.7 \times 100) = 30 + 70 = 100$, and we conclude that George rationally ought to choose O_1. *But*, if we instead calculate *risk-weighed* expectation over T_1 and T_2, using T_2's preferred square root function, we find something different: $\text{EV}_r(O_1) = (.3 \times \sqrt{250}) + (.7 \times \sqrt{50}) \approx 4.74 + 4.95 = 9.69$ while $\text{EV}_r(O_2) = (.3 \times \sqrt{100}) + (.7 \times \sqrt{100}) = 3 + 7 = 10$, and we conclude that George rationally ought to choose O_2.

In the Easy Case, I argued that risk-neutral expectational aggregation was appropriate on the grounds that both theories in which Alice had positive credence endorse a risk-neutral expectational principle of decision-making under uncertainty. (See §B.1 for details.) But what should we say when the first-order theories we are attempting to aggregate seem to disagree about the appropriate aggregation principle?

One way to think of the difficulty is to recall the idea that second-order principles of decision-making under moral uncertainty might "beg the question" against the first-order theories they are meant to help us handle our uncertainty over: If George is uncertain whether he ought to be risk-neutral or risk-averse, then in aggregating the risk-neutral and risk-averse theories by means of a risk-neutral decision procedure, he would seem to have forgotten that this is the very thing he's uncertain about.

Another feature of the difficulty is that either choice of aggregation procedure, in a case like George's, seems to give one of his first-order theories undue weight. The risk-neutral theory insists that small chances of very large positive payoffs should be taken as seriously as their expectations suggest, while the risk-averse theory insists that one should generally not sacrifice the surety of a good thing for a small chance of an *extremely* good thing, even when the expected value of doing so is positive. Adopting a risk-neutral second-order aggregation procedure will allow this sort of long-shot reasoning to dominate decision-making even when the agent's credence in the risk-neutral first-order theory is relatively low, while adopting a risk-averse second-order aggregation procedure may result in almost complete insensitivity to high-value, low-probability prospects, even when the agent's credence in the risk-neutral first-order theory is relatively high. The more the decision procedures diverge (e.g. if instead of the square-root risk averse theory we consider a cube-root theory, or a theory that simply rounds off "rationally negligible probabilities" to zero), the more extreme the effects of committing to the corresponding second-order aggregation principle will become.

There are a great many things that might be said about this problem (which we might call the "problem of decision-theoretic uncertainty"), but as in previous sections, I will focus on a few general lines of response.

1. **Least Common Denominator** The aggregate value assignment of comparability class $\{T_1, T_2, ...T_n\}$ is defined only insofar as the decision procedures endorsed by each of T_{1-n} are in agreement. Where there is disagreement, the result will be an imprecise assignment of cardinal values and/or an incomplete ordinal ranking of options.

2. **Open-Ended Internalism** What procedure should be used to aggregate a comparability class $\{T_1, T_2, ...T_n\}$ depends entirely on the agent's doxastic state, but may depend on features of her doxastic state other than the decision procedures endorsed by T_{1-n}. For instance, an agent may be uncertain

whether risk-neutrality or risk-aversion is correct at the first-order level but be justifiably certain, e.g., that risk-neutrality is correct at the second-order level, or that she should adopt a second-order risk attitude that averages (in some particular way) the risk functions of her various first-order theories, or that the only valid second-order norm is MFO or statewise dominance or... In any of these cases, she rationally ought to employ the second-order aggregation procedure that she judges to be correct, whether or not it is intuitively "compatible" with the first-order theories being aggregated.

3. **Externalism** Some decision-theoretic principle(s) have belief-independent rational requiring force, such that any agent is rationally required to aggregate over her uncertainties by means of these principles regardless of her doxastic state, and in particular regardless of what decision-theoretic principles are endorsed by the first-order theories in which she has positive credence.

An evaluation of these options must be put off to the next chapter, where I will argue for my own preferred resolution to the problem of decision-theoretic uncertainty. But to sketch the argument in advance: As I have already suggested, I agree with Weatherson that the regress problem forces us to treat at least *some* normative principle or principles as having external normative force. Hence, I will argue in the next chapter, when an agent is uncertain about all higher-order as well as first-order normative principles, open-ended internalism gets stuck in an infinite regress and the least common denominator approach on its own yields the vacuous result that everything is rationally permissible. So we must adopt at least a weak form of externalism. However, if the principles to which we attribute external force are *sufficiently* weak (e.g. statewise or stochastic dominance rather than maximizing expected choiceworthiness), then the least common denominator approach may still have a role to play, providing us with *stronger* aggregation procedures for certain comparability classes of theories.

6.5 Rational Choice under Normative Uncertainty: A Tentative General Account

In this chapter I have identified three aspects of the problem of intertheoretic value aggregation, each of which presents a distinct challenge for a complete theory of rational decision-making under moral uncertainty. I have described what I take to be the most plausible responses to each of these challenges, and argued for my own preferred responses (though as yet in only a promisory way with respect to the problem of decision-theoretic uncertainty). Taken in combination, these responses constitute the outline of a general theory of decision-making under moral uncertainty. This theory can be summarized as follows:

General View *multi-stage aggregation* ∧ (*universal content-based normalization* ∨ *top-down normalization*) ∧ (*externalism* + *least common denominator*)

To put this into slightly better English: A morally uncertain agent should decide what to do by aggregating the value assignments of the various moral theories in which she has positive credence, beginning with minimal comparability classes (united by shared structure, precise content-based normalization, and if necessary by a shared decision theory), whose aggregate value assignments are taken as inputs to broader comparability classes, and so on. The normalization and aggregation procedures within comparability classes should be grounded in the agent's own doxastic state as much as possible, but if content-based normalization turns out to be insufficient then top-down normalization principles may have a role to play as well, and some (perhaps quite weak) aggregation principle must have normative force that is independent of the agent's doxastic state.

At several points, I have defended various components of this view by comparing costs, i.e., conceding that my own preferred approach fails or might fail to satisfy some genuine desideratum while arguing that alternative views fail to satisfy more important desiderata. It would be more persuasive, certainly, if I could convincingly claim that my view had no substantial downsides. But as far as I can tell, there is simply no such view on offer in this domain. Irrelevantism, MFT, MFO, and every extant hedging theory all carry substantial costs, and if these costs cannot be avoided, then they must simply be compared.

The view I have put forward remains highly schematic, of course, and therefore its costs can be only imperfectly assessed. Among the details left to be filled in are (i) the precise order in which the multi-stage aggregation procedure unfolds (i.e., which comparability classes should be taken as input to which larger comparability classes); (ii) what kinds of content-based normalization are possible, especially between very disparate theories; (iii) whether there is a correct way to normalize theories when content-based normalization fails, and if so what top-down normalization principle gives the right account of these cases (e.g., by capturing the true purport of PES);

and (iv) what aggregation principle or principles have external normative force. (I will have something to say about this last question in the next chapter.)

Especially in light of these as-yet missing details, it is natural to ask: Isn't this all just too complicated to be the right account of things? Doesn't the comparative simplicity of MFT, MFO, or even irrelevantism look like a powerful reason to prefer one of these views over moral hedging, if this is the simplest and most cohesive view the advocate of hedging can provide? This concern is worth taking seriously. Nonetheless, the complexity of the view I have outlined proceeds from a simple idea: that a rational agent should try to do the right thing, relative to her own imperfect state of belief about her objective reasons. The complexity of the deliberative process that this entails for an ideal agent (and that non-ideal agents can only try to approximate) results from the complexity of the agent's doxastic state, i.e., from her distribution of belief over moral theories with complex and diverse structure, metaethical/explanatory content, and decision-theoretic commitments. Any theory of decision-making under moral uncertainty that takes seriously the idea of *responding rationally to one's actual belief state* will have to respect this complexity, not attempt to eliminate it (e.g., by tacitly replacing all the theories in which the agent has positive credence with simple cardinal surrogates). Nonetheless, the complexity of the resulting decision-procedure has a unifying motivation.

One aim of the next chapter will be to give a more precise account of this underlying unity: in particular, to suggest that a single norm, the enkratic principle correctly understood, can explain both the internalist and externalist elements of rational decision-making under moral uncertainty.

Chapter 7: The Regress Problem

The central question so far as been the following: What are the correct second-order rational norms for an agent who is uncertain about first-order moral norms—i.e., what is the correct deliberative procedure for such an agent to follow? We have not yet considered the natural follow-up question: How should an agent deliberate when she is uncertain not only about first-order moral norms but also about second-order rational norms? If, for instance, in additional to having positive credence in various first-order moral norms that offer conflicting guidance in the choice situation before her, she also has some positive credence in various competing second-order principles like MFT, MFO, and a variety of hedging principles involving different approaches to the problem of structural diversity, the normalization problem, and/or the problem of decision-theoretic uncertainty, then should she simply be guided by whichever of these second-order principles is most plausible (the higher-order equivalent of MFT), or choose the option supported by the greatest probability-weighed sum of second-order principles (the higher-order equivalent of MFO), or find some way to weigh the advice of these competing principles against each other (the higher-order equivalent of hedging)? And what if she is uncertain in turn about which of these *third*-order approaches to her second-order uncertainty is correct?

These questions introduce a fundamentally new difficulty, and suggest a final line of objection with which any complete defense of moral hedging must contend.

In the next section, I will consider the regress problem as an argument for normative externalism. I will then examine two possible lines of reply that avoid conceding the existence of any external norms, and argue that these replies are unsatisfactory. Instead, I'll then argue that we should admit the existence of one external norm: the enkratic principle, correctly formulated. I will consider some variants of the enkratic principle that have different implications for decision-making under normative uncertainty, without attempting to settle the potential debates between them.

This "enkratic externalist" position satisfies the demands of the regress argument for externalism. But it is not obvious that this is enough to solve the regress problem. In §7.2, I consider two aspects of the regress of higher-order norms that might be thought to pose a problem, even given the concession of a limited form of externalism. One concerns how an ideally rational agent should act when she is uncertain about norms at every level of an infinite hierarchy. The other concerns how a *non*-ideal agent should approximate the requirements of ideal rationality, when she is only capable of contemplating finitely many levels of such a hierarchy.

I will argue, however, that enkratic externalism does in fact solve both of these problems: If rationality requires full commitment to the enkratic principle, then higher-order uncertainties take the form, not of uncertainty about competing final principles, but of uncertainty about competing *heuristics* for predicting the demands of the enkratic principle. For this reason, I argue in §7.3 that the problem of ideal rationality simply doesn't arise, given enkratic externalism: Any ideally rational agent, capable of contemplating an infinite hierarchy of norms, would not need to rely on heuristics to determine the probabilities of objective value propositions, but could reason directly from her evidence. There are also at least two ways in which a regress of *ideal*, non-heuristic principles might still be thought to arise, despite the commitment to enkratic externalism, but I argue that both can be avoided.

But this does not resolve the problem of non-ideal rationality: Given that I must rely on heuristics, for instance, to approximate the expected choiceworthiness of my options, and that I must inevitably be at least somewhat uncertain about what heuristic norms to follow, I and other non-ideal agents still seem condemned to infinite regress. Taking up this problem in §7.4, I describe a model of non-ideal practical deliberation on which agents like ourselves are not rationally required to ascend an infinite hierarchy of heuristic norms, because at some point in the process of deliberation—before we have ascended more than a very short way up the hierarchy—we will justifiably conclude that the marginal cost of any further deliberation exceeds the marginal benefit.

7.1 The Regress Argument for Externalism

We first encountered the regress problem as an argument for normative externalism in Chapter 3. But we should now take the time to set out this argument in more detail. Its most compelling formulation, to my mind, is as follows.

The Regress Argument

1. If the rational requiring force of any normative principle N for agent A (i.e., whether A is rationally required to do as N advises) depends on A's beliefs about N, then in any choice situation where A is uncertain of the correctness of N_1, which advises her to choose option O, and assigns some positive degree of belief to rival norms $N_2, N_3, ...N_m$, some of which advise against choosing option O, A is not rationally permitted to choose option O straightaway on the basis of N_1, but is rationally required to deliberate in a way that accounts for the possible correctness of N_{2-m}.
2. In order to deliberate in a way that accounts for her uncertainty over N_{1-m}, A must make deliberative use of at least one higher-order normative principle N_1'.
3. A rational agent will always be at least somewhat uncertain of any normative principle or disjunction of normative principles, apart (perhaps) from unhelpful trivialities like "It's not the case that O is both rationally permissible and not rationally permissible."
4. If (3), then for any substantive normative principle N, whether first-order or higher-order, and in any choice situation, any rational agent will have non-zero credence in some normative principle of the same order that conflicts with N (i.e., that regards a disjoint set of options as permissible).
5. Therefore, if the rational requiring force of any normative principle N for agent A depends on A's beliefs about N, then any rational agent in any choice

situation will be rationally required to make deliberative use of an infinite series of higher-order normative principles. [from (1)-(4)]

6. Agents are not rationally required to make deliberative use of an infinite series of higher-order normative principles.

C. For at least some normative principle N, the rational requiring force of N for an agent A does not depend on A's beliefs about N. [from (5), (6)]

Weatherson, I think, has something like this argument for externalism in mind (Weatherson, 2014, pp. 155-7). I will argue shortly that the argument is sound, but that contra Weatherson, it provides most support not for externalism about first-order moral principles but rather for externalism about second-order rational principles (specifically, the enkratic principle). But first we should consider potential objections to the argument for externalism. Many such objections are possible. Premise (1) may seem implausibly demanding, and Sepielli's solution to the regress problem, which we will consider in §7.1.2, can be interpreted as an attempt to reject it. Premise (2) is plausible but non-obvious. Premise (3) depends on a commitment to Bayesian regularity, which might be rejected for various reasons. I will focus, however, on two general lines of response, which might be understood as attacking the above argument at various points, but are best understood as strategies for rejecting either (1), (4), or (6).

7.1.1 Fixed Point Solutions: Convergence or Contraction

The simplest way of responding to the regress problem is to claim that the regress will stop after some finite number of iterations, because a rational agent will in one way or another achieve certainty either about the relevant higher-order norms or about what she subjectively ought to do. More precisely, there are two ways in which the regress of higher-order norms might turn out to be less than maximally vicious.

1. **Convergence** For any rational agent A in any choice situation, there is some finite n such that all nth- and higher-order norms in which A has positive credence will agree on which of her options are rationally permissible.

2. **Contraction** For any rational agent A in any choice situation, for any level n in the hierarchy of higher-order norms, the set of options that some $n+1$st order norm in which A has positive credence regards as rationally permissible will be a subset of the set of options that some nth order norm regards as rationally permissible—that is, as A ascends the hierarchy of norms, the set of options that *might* be permissible will monotonically contract. Assuming that A has only finitely many options, then there must be some level m such that for any level $o \geq m$, the set of possibly permissible options according to oth-order norms in which A has positive credence will be identical to the set of possibly permissible options according to mth-order norms in which A has positive credence—that is, the set of *possibly permissible* options will stabilize,

> remaining the same for all orders of norm greater than or equal to m. Then, one might think, there will be some simple procedure for choosing among this stable set of possibly-permissible options.

Convergence would offer the simplest and most desirable solution to the regress problem, and would escape the argument for externalism by denying premise (4). But I see no reason to think that it is correct. Surely in any choice situation, for any level n in the hierarchy of norms, there is at least one pair of *logically* possible nth-order norms that offer conflicting guidance (i.e., disagree about which options are rationally permissible). And even if we grant, contra regularity, that it is sometimes rationally permissible to have zero credence in a logical possibility, what reason is there to think that we will *always* be able to eliminate as epistemic possibilities all nth-order norms, for some n, except those that agree on a particular set of options as permissible?

The more modest hypothesis of Contraction is suggested by Sepielli (2014b).[1] It is not obvious to what extent Contraction, if true, would help us solve the regress problem. One trivial way in which it could be true, for instance, is if we accept the apparent implication of regularity, that for any option in any choice situation, a

[1] He writes: "I hope to eventually show that the further we 'step back' from states of uncertainty [...] the fewer prospective actions will be left in the disjunctive set amenable to intentional explanation[...] Suppose, for example, that I can do any of mutually exclusive actions A...Z. Perhaps my uncertainty regarding objective normativity will intentionally explain my doing any of A...R rather than S...Z [...] Now suppose that I consider what to do relative to the [minimal probabilities] expressed in that uncertainty, and am uncertain as to the answer. Perhaps this new uncertainty will intentionally explain my doing any of A...G, rather than H...Z [...] My hope is to show that, as a general matter, potential actions will be hived off with each stepping-back, and that previously hived off actions will never reappear among the set amenable to intentional explanation. I have reasons to suspect that such a showing is possible, but I'll have to make good on that suspicion elsewhere." (Sepielli, 2014b, p. 539)

rational agent should have some positive credence, at every level of the hierarchy of norms, in one or more norms that regard that option as rationally permissible. In this case, the set of possibly permissible options would never expand as we ascend the hierarchy (satisfying Contraction), but it would never contract either. And this presumably does not constitute a satisfactory solution to the regress problem.

Contraction might be helpful if it were guaranteed that, as we ascend the hierarchy of norms, the set of possibly permissible options would contract in some significant way (without necessarily reaching the point at which all theories agree on which options are permissible), and never expand. Then it might be plausible to hold that, whenever an agent realizes that, say, for every level of the hierarchy $n \geq 5$, the nth-order norms in which she has positive credence will designate exactly the options $O_1, O_2, ...O_m$ as possibly permissible (i.e., at least one nth-order norm will designate each of O_{1-m} permissible, and no nth-order norm will designate any other option as permissible), she can then invoke some simple "ω-order" norm (e.g., that permits her to choose any of O_{1-m}).[2] This would amount to denying premise (6) in the argument for externalism, and holding that although rationality does require agents to consider the full infinite hierarchy of norms, the stabilization of the set of possibly permissible options above some finite level renders this requirement feasible for finite agents.

There is an obvious reason why this solution should seem unpromising: namely, if we have followed the regress argument this far, why should we not suppose that an agent can be rationally uncertain about ω-order norms just as much as she can be

[2] MacAskill appears to endorse this view (MacAskill, 2014, pp. 218-219).

about norms at finite levels of the hierarchy? For instance, along with the norm that permits any of O_{1-m}, she might consider a norm that instructs her to choose the option *most likely* to be permissible, perhaps according to her fifth-order norms (the first level of the hierarchy at which the set of possibly permissible options reaches its stable minimum), or perhaps according to the norms at the highest finite level of the hierarchy she is capable of considering, or perhaps somehow averaging the probability of an option being permissible over the levels of norms ≥ 5 (e.g, for each option O, taking the Cesàro sum of the probabilities that it is rationally permissible given by the 5th-order norms, 6th-order norms, 7th-order norms, ...). Thus, so long as the norms continue to disagree about which options are permissible at every finite level (or more precisely, so long as for every level, there is some higher level at which some norms disagree about which options are permissible), Contraction alone seems unpromising, even if Sepielli's monotonicity conjecture turns out to be correct.

7.1.2 Sepielli's View

In other work, however, Sepielli has suggested a different approach to the regress problem. This approach begins by distinguishing between *dispositional* uncertainty and *conscious* uncertainty. A rational agent, Sepielli claims, may be dispositionally uncertain without being consciously uncertain, and may rationally act on a norm N without considering alternative norms, even if she is dispositionally uncertain about the correctness of N, so long as she is not consciously uncertain (Sepielli, 2014a). This view avoids the argument for externalism by denying either premise (1) or premise (3)—premise (1) if we interpret the references to "uncertainty" in these premises as referring to dispositional uncertainty, premise (3) if we interpret them as referring to conscious uncertainty. Since I will claim that dispositional uncertainty is the more relevant condition, let's adopt the first interpretation of the argument, and understand Sepielli as challenging premise (1).

One immediate worry is that Sepielli's view seems too permissive: We don't want to concede that an agent is rationally permitted to act on any norm, however much her evidence calls that norm into question, so long as she doesn't *feel* uncertain about it. Sepielli offers a response to this worry, arguing that a rational agent will always experience some *pro tanto* impetus to attend consciously to her dispositional uncertainties, and that this will sometimes but not always lead her to experience conscious uncertainty about a normative principle N about which she was previously only dispositionally uncertainty, and therefore to seek some higher-order norm to guide her action in the face of this newfound conscious uncertainty. She may become consciously uncertain of this higher-order norm as well, and be forced to ascend the hierarchy another level, but she also may not, and in any event will at some point reach a level where her impetus to convert dispositional uncertainty into conscious uncertainty runs out, ending the regress.

Nevertheless, Sepielli's view is still implausibly lax. Consider where the line between dispositional and conscious uncertainty must lie, for this view to succeed in escaping the regress problem. We might draw a distinction between the merely-dispositional uncertainty of an agent who doesn't realize she has any reason to doubt proposition P, but were she to consider it, would conclude that she does, and the conscious uncertainty of an agent who realizes that P is less than certain. But understood in this way, the distinction would not resolve the regress problem, for a philosophically astute agent will realize that *any* (or nearly any) non-trivial normative proposition is less than certain, and hence will in this sense be consciously uncertain about any norm she considers guiding her action by.

Sepielli's distinction between conscious and dispositional uncertainty is different. Here is his explanation.

> I think we need to distinguish between two types of uncertainty. The first is dispositional, not necessarily conscious, the sort of attitude I have towards any claim I wouldn't bet my life on. The second is conscious, *involving a feeling of directionlessness, the kind that appears when I deliberate, and disappears when I'm "in the zone"* [emphasis added]. I am uncertain in only the first sense about what the strings on a guitar are; I am uncertain in both senses about what the strings on a banjo are. That is why I can simply play an A7 on a guitar, but can play an A7 on a banjo only *by trying.* (Sepielli, 2014a, p. 91)

As Sepielli himself admits, however, it is unclear why the absence of *this* sort of conscious uncertainty ("a feeling of directionlessness") should make it permissible for me to act straightaway on a norm N, without considering the possibility that N might be mistaken. He concedes that "the waning of conscious uncertainty is only a solution to the *psychological* problem of how we can act without [taking unguided leaps of faith]. It's not a solution to the *normative* problem of how we can manage moral risks non-recklessly" (pp. 91-2).

Sepielli's final conclusion with respect to the regress problem, then, is moderately pessimistic.

> I think the right thing to say is that meta-rules offer us a normative advantage by *forestalling* moral recklessness, rather than by eliminating it entirely. More precisely, there is a sense in which it is better to leap [i.e., "take a leap of faith" by acting on a norm N the truth of which is uncertain] in the face of uncertainty about meta-rules than to leap in the face of uncertainty about ordinary moral rules, better still to leap in the face of uncertainty about meta-meta-rules, and so on. (Sepielli, 2014a, p. 92)

On one interpretation, this view amounts to a concession that agents can never fully satisfy the requirements of rationality. On another, Sepielli is still denying premise (1), and holding that at least in some circumstances when an agent is dispositionally uncertain but not consciously uncertain about a norm N, she is rationally permitted to act on N straightaway without considering the possibility that N might be wrong. The latter conclusion would seem unwarranted, however, absent a reason why the *feeling* of certainty relieves an agent of her rational responsibility to consider the possibility of error. The former conclusion would be disappointing, and is worth avoiding if we can. In the next section, therefore, I will suggest that we should concede a weak form of externalism, one that preserves the requirement of moral hedging while avoiding an infinite regress of higher-order norms.

7.1.3 Enkratic Externalism

My proposal is that there is exactly one normative principle that has external, belief-independent normative force, namely, the enkratic principle, rightly formulated. Let's call this view *enkratic externalism.*[3] The view, so stated, is intentionally underspecified, for as we saw in §2.2, many formulations of the enkratic principle as possible. We will shortly consider, without attempting to definitely resolve, which of these formulations is right. But first let's make an initial assessment of the plausibility of enkratic externalism in its underspecified condition.

By conceding the conclusion of the regress argument given in §7.1, enkratic externalism allows us to accept the premises of that argument without paradox. Whether this is enough to resolve the regress problem is another matter, which we will take up in §§7.2-7.4. But on face it seems promising, for if an agent is rationally permitted—indeed, rationally required—to do as EP advises, even when she is uncertain of EP, and if EP can play the role of a second-order principle for decision-making under first-order moral uncertainty, then the regress simply stops at the second level: An agent may act on a second-order principle (EP) without considering alternatives, and therefore without needing to make deliberative use of any third-order principle(s).

[3] The idea that the principles of rationality that govern choice under first-order moral uncertainty have belief-independent force has been suggested by Krister Bykvist, though his conception of rationality seems to be Humean rather than enkratic: "Being rational, in the sense I have in mind, has to do with the coherence between your preferences about simple outcomes, what you fundamentally care about, and your preferences about prospects constructed out of these simple outcomes...[M]y tentative conclusion is that in cases of uncertainty of rational matters there is an answer to the question of what it is rational to prefer which is not sensitive to your own views about rationality." (Bykvist, 2013, p. 133)

But even if it can solve the regress problem, enkratic externalism faces challenges from two directions: from internalists like Sepielli, and from first-order moral externalists like Weatherson. From the internalist direction, the basic question is: How can an agent be rationally required to act on a principle regardless of her beliefs or evidence with respect to that principle? To begin with, following premise (1) of the argument for externalism, we might wonder why it is even rationally permissible to act (straightaway) on a principle of which one is uncertain. But enkratic externalism endorses a stronger claim than this: namely, that an agent is *rationally required* to act on EP, even if she justifiably regards it as more likely (perhaps *much* more likely) false than true. This seems to gratuitously offend the idea that favored moral hedging in the first place, that agents should be guided by their normative beliefs. Indeed, by this token, it seems paradoxically to offend the essential spirit of the enkratic principle itself.

It seems to me, however, that there is no paradox in viewing EP as an exception from its own internalist motivations. Enkratic externalism claims that EP is a *constitutive* or *definitional* requirement of rationality, i.e., that to be practically rational *just is* to be enkratic. To see the plausibility of this claim, consider the limiting case of EP in which an agent A is certain that she has most objective reason to choose option O. If, to put it roughly, being rational *just means* that you respond in a coherent way to your reasons as you understand them, it seems that rationality must require A to choose O—for how could choosing anything other than O constitute a coherent response to A's reasons in this circumstance, as she understands them? Even if A, as a good Bayesian, has *some* positive credence in the principle

EP^{-1} according to which she is rationally *prohibited* from choosing any option that she is certain she has most objective reason to choose, there is plausibly nothing irrational in ignoring that credence. By contrast, first-order moral principles like *maximize aggregate pleasure minus pain* or *don't intentionally deceive other agents* are not plausibly constitutive requirements of rationality (*pace* certain Kantians)—rather, they are theories of an agent's objective reasons, which an agent might fully or partially reject for purposes of practical deliberation without compromising her rationality.[4]

This explains why enkratic externalism should be acceptable to proponents of moral hedging, and also suggests a reason to prefer enkratic externalism to Weatherson's first-order moral externalism. Are there any countervailing reasons to prefer

[4] The regress problem for normative uncertainty has a rough analogue in the regress problem faced by conciliationist views of peer disagreement (how to handle "peer disagreement about peer disagreement"), which has received somewhat more attention in the recent literature. Weatherson (2007, 2013) has also taken the pessimistic view in this debate, arguing that conciliationist views are incoherent since they recommend moderately one's credence in conciliationism in the face of disagreement while also adopting degrees of belief with respect to other propositions that can only be justified by full credence in conciliationism. In response, Elga (2010) defends a version of conciliationism that is *self-exempting*, arguing that any basic principle of inductive reasoning "must be dogmatic with respect to its own correctness" (Elga, 2010, p. 185). That is, Elga claims that an epistemically rational agent should be steadfast in the face of peer disagreement about conciliationism and other basic inductive norms, believing with probability 1 that those norms are correct, while adjusting her credences on other questions in response to disagreement as conciliationism demands. This is not dissimilar from the enkratic externalist claim that a rational agent should be fully committed to EP in deliberation, while hedging for her uncertainty about other normative principles on the basis of EP. The important difference is that I am claiming, not that a rational agent should be dogmatic about the *truth* of EP, no matter what her evidence, but that she should *comply with* EP even when she is uncertain of its correctness. In the peer disagreement context as well, this externalist line strikes me as the most plausible conciliationist response to the threat of regress. That is, rather than claiming that an agent should always have credence 1 in conciliationism, whatever her evidence to the contrary, the conciliationist should instead claim that some basic norms of epistemic rationality, which include or imply conciliationism, have *belief-independent force*, such that agents are rationally required to comply with these norms even while rationally doubting their correctness. Like the enkratic externalist view I am defending, this claim will be plausible to the extent that these epistemic norms can be plausibly characterized as constitutive or definitional requirements of epistemic rationality, i.e., "just what it is" to be epistemically rational.

first-order moral externalism, apart from those we considered in Chapter 3? Weatherson does not say much on this question. The relevant passage (worth quoting at length in part for its particularly nice statement of the regress problem) reads as follows.

> There is a worry that externalism is not sufficiently action guiding, and can't be a norm that agents can live by. But any philosophical theory whatsoever is going to have to say something about how to judge agents who ascribe some credence to a rival theory. That's true whether the theory is the first-order theory that Jeremy Bentham offers, or the second-order theory that Andrew Sepielli offers. Once you're in the business of theorising at all, you're going to impose an external standard on an agent, one that an agent may, in good faith and something like good conscience, sincerely reject. *The externalist says that it's better to have that standard be one concerned with what is genuinely valuable in the world, rather than a technical standard about resolving moral uncertainty* [emphasis added]. But every theorist has to be a little bit externalist; the objector who searches for a thoroughly subjective standard is going to end up like Ponce de Leon. (Weatherson, 2014, pp. 156-7)

As the italicized sentence suggests, Weatherson's major motivation for preferring first-order moral externalism to any form of externalism that would allow for hedging is the fetishism argument, to which I have already replied. There is also a suggestion, perhaps, that it would be *arbitrary* for some "technical standard" like

Lockhart's PEMT to have external normative force. But enkratic externalism avoids this worry insofar as it is plausible that enkrasia is uniquely constitutive of practical rationality.[5]

7.1.4 Which Enkratic Principle?

As we've seen, many formulations of the enkratic principle are possible. The standard formulation of EP says that if A believes that she ought to choose option O, then she is rationally required to choose option O. Following Wedgwood, I pointed out in Chapter 2 that this principle either fails to account for uncertainty (since an agent might *believe* that she ought to choose O, and yet have positive credence that she has very strong reasons to choose some other option instead, such that it is on balance not rational for her to choose O), or else is almost never applicable to actual agents (if *belief* is understood to require the absence of any uncertainty). Wedgwood's formulation of the enkratic principle as a requirement to maximize expected choiceworthiness offers a more general principle of rationality that accounts for uncertainty about objective oughts.[6]

One version of enkratic externalism, then, holds that agents are rationally required, whatever their beliefs, to choose options with maximal expected choiceworthiness. Let's call this view EE^{MEC}. We might reasonably wonder whether this

[5] One might worry that fully escaping the regress problem requires us to grant belief-independent status to some normalization principle like PEMT, or other technical features of a theory of decision-making under moral uncertainty that do not seem like constitutive requirements of rationality. I address this worry in §7.3 below.

[6] As we've also seen, Sepielli, MacAskill and others endorse maximizing expected value/choiceworthiness as a general principle of rationality, though to my knowledge Wedgwood is the first to have proposed it as a formulation of the enkratic principle.

view is too strong, though. Is it really *constitutive of rationality* that I maximize expected value, with the very particular risk attitude (viz., neutrality) that this implies? Many theorists have argued that that non-neutral attitudes toward risk are rationally permissible (prominent examples include Keynes (1921) and Buchak (2013)), and it seems implausible that an agent who justifiably believes this to be the case and, rather than maximizing expected choiceworthiness *simpliciter*, maximizes some form of *risk-weighted* expected choiceworthiness (perhaps averaging over risk-weighed and risk-neutral theories of rational choice under moral uncertainty) thereby acts irrationally.

We might therefore prefer a weaker version of enkratic externalism. One salient option is to replace Wedgwood's expectational principle with a *stochastic dominance* principle. A practical option O stochastically dominates an alternative P for agent A iff, relative to A's subjective probabilities

1. For any degree of choiceworthiness d, the probability that O is (or will turn out to be) at least as choiceworthy as d is equal to or greater than the probability that P is ("") at least as choiceworthy as d, and

2. For some degree of choiceworthiness d, the probability that O is ("") at least as choiceworthy as d is strictly greater than the probability that P is ("") at least as choiceworthy as d.[7]

A stochastic dominance version of the enkratic principle would assert that, for any

[7] More precisely, these conditions define *first-order* stochastic dominance—I omit this qualification in the main text for simplicity. More a more detailed explanation of stochastic dominance reasoning, including its application to moral uncertainty over ordinally and lexically structured theories, see Appendix A.

agent A and option O, if there is some alternative option P such that A's probability distribution over P's possible degrees of objective choiceworthiness stochastically dominates A's probability distribution over O's possible degrees of objective choiceworthiness, then A is rationally required not to choose O. Let's call the version of enkratic externalism based on this formulation of the enkratic principle EE^{SD}.

Stochastic dominance is, in effect, expectation-maximizing minus any constraints on an agent's risk attitudes. Unlike expectational reasoning, which is controversial because of the constraint it imposes on risk attitudes, the principle that agents ought to reject stochastically dominated options is largely uncontroversial in decision theory. In formal terms, an agent will reject stochastically dominated options so long as her utility function is monotonically increasing, i.e., so long as she prefers a given chance of a higher-value outcome to the same chance of a lower-value outcome, all else being equal (Hadar and Russell, 1969, p. 28). This is much more plausibly a constitutive requirement of rationality than the risk-neutrality required by MEC.

A further advantage of a stochastic dominance formulation of the enkratic principle over Wedgwood's MEC formulation is its ability to handle moral uncertainty over ordinal and lexical theories. MEC is simply unable to handle ordinal theories (since ordinal rankings can't be multiplied by probabilities) and yields the implausibly fanatical conclusion that any non-zero probability of a gain or loss on the lexically prioritized value dimension according to a class of lexical theories takes absolute precedence over any gain or loss on the lexically deprioritized dimension. Stochastic dominance, while it is silent in a great many cases, can accommodate and yield plausible conclusions with respect to both ordinal and lexical theories (see Appendix A).

But there are drawbacks to EE^{SD} as well. For one thing, as just alluded to, stochastic dominance offers a significantly incomplete ranking of uncertain prospects—i.e., for most pairs of options, neither will stochastically dominate the other. Thus, relying *purely* on stochastic dominance reasoning would leave us with a very weak principle of decision-making under moral uncertainty, forcing us to concede that in the typical case, many of an agent's options will be rationally permissible. An alternative is to adopt the hybrid approach suggested in the last chapter, treating stochastic dominance as the only principle of rational requirement that has *external* normative force while holding that an agent may be subject to *internal* (belief-based) rational requirements to employ stronger principles, e.g., to aggregate expectational theories expectationally. We will explore this issue at greater length in §7.3.

Additionally, while expectational reasoning requires only *unit*-normalizability between moral theories, stochastic dominance reasoning requires *level*-normalizability, i.e., it requires that we be able to say whether the value of option O according to T_1 is greater than, less than, or equal to the value of option P according to T_2.[8]

In light of the apparent limitations on the scope of both MEC and SD, we might adopt the view that the right formulation of EP is a hybrid, e.g., requiring only that agents reject stochastically dominated options at stages of the aggregation process where level comparisons are possible, while requiring that they choose options with

[8] In general, any decision principle that requires or permits non-neutral risk attitudes will require level-normalization, since risk-seeking and risk-aversion are impossible without the ability to make such absolute level comparisons. MEC avoids the need for intertheoretic level comparisons by requiring risk neutrality.

maximal expected choiceworthiness at stages where unit comparisons but not level comparisons are possible.[9]

I have no strong view on which precise formulation of EP has external rational requiring force, except that it is either MEC, stochastic dominance, or some combination thereof. In the next three sections I will try to show that enkratic externalism can resolve the regress problem, regardless of the precise formulation of EP on which it is based.

7.2 Two Regress Problems

Although I initially framed the regress problem as an argument for normative externalism, that does not mean that conceding any form of externalism is enough to make the problem go away. As we will see in the coming sections, significant difficulties remain. To address these difficulties, it will be helpful to first divide them into two categories: those that concern *ideal* rationality, and those that concern *non-ideal* rationality.

Intuitively, the distinction is this: First, suppose that an ideally rational agent will have beliefs about normative principles at every level of an infinite hierarchy, and that at every level she will have positive credence in principles that disagree about which of her options in a given choice situation are permissible. Given full

[9] It is more plausible that risk neutrality is a requirement of rationality in such contexts since the absence of level comparability simply precludes non-neutral risk attitudes. On this line of thought, perhaps an agent who distributes belief over a comparability class of cardinal theories that are unit-comparable but not level-comparable relative to one another should compute expectations internal to each theory by that theory's own preferred decision procedure (which might involve non-neutral risk attitudes) but then aggregate the comparability class according to (risk-neutral) MEC.

cognitive access to her own credences and unlimited cognitive resources, how should such an agent decide what to do? This is the *ideal regress problem*.

But second, even if we had a fully satisfactory solution to the ideal problem, we still face the following difficulty: Any finite, non-ideal agent like ourselves will in general not be able to contemplate an infinite hierarchy of normative principles every time she has to decide what to do. Unless her credal sate with respect to higher-order norms is extremely simple, she either will not *have* well-defined credences about higher-order norms or, at least, will not have immediate cognitive access to those credences. Moreover, even if she did have such credences and could access them in order to decide what to do, she certainly lacks the cognitive resources to carefully consider the various competing norms at each level and apportion her credence correctly to the evidence. So given any account of how an ideally rational agent should act when she is uncertain over an infinite hierarchy of norms, there is a further question how if at all a non-ideal agent can approximate the requirements of ideal rationality. This is the *non-ideal regress problem*.

As we will see, both problems remain puzzling even if we accept enkratic externalism. Nevertheless, I will argue this commitment ultimately enables us to resolve both problems.

7.3 The Problem of Ideal Rationality

In §7.1.3 I suggested that, if the enkratic principle could be taken as a second-order principle of choice under moral uncertainty on which an agent is rationally permitted to act straightaway, despite her uncertainties, then there would be simply no need for third-order norms of decision-making under second-order uncertainty, and hence no regress problem to contend with. But does EP alone obviate the need for any second-order normative principles about which an agent might be rationally required to account for her uncertainties? In other words, is (the correct formulation of) EP a *complete* principle of rational decision-making under first-order moral uncertainty, or does it simply *constrain the field* of second-order principles? If the latter, then the regress problem is unresolved: for even if an agent may rationally disregard EP-incompatible second order norms in which she has positive credence (e.g. MFT, MFO, or the norm that tells her to minimize expected choiceworthiness), she must still account for her uncertainty about EP-compatible norms. And then, it seems, she will be stuck contemplating the same infinite hierarchy of competing higher-order norms, made somewhat skinnier but no less tall by the constraints of EP.

There are two different worries about the completeness of EP as a second-order norm that deserve separate attention.

7.3.1 Gappiness

Both MEC and SD have the potential to leave agents with *rational options*, i.e., require of an agent in a particular choice situation not that she choose one particular option but that she choose one of some disjunction of options. MEC allows for rational options in two circumstances: (i) when the expected value of one or more options is undefined and (ii) when two or more options are tied for maximal expected value and hence are judged equally rational. SD allows for rational options whenever there is more than one option in a given choice situation that is not stochastically dominated by any other option. Since rational options are a more general phenomenon under SD than under MEC, let's focus on SD in this section.

Two interpretations of the stochastic dominance formulation of EP are possible. On one interpretation, EP tells me that I am rationally required not to choose stochastically dominated options, and is simply silent with respect to options that are not stochastically dominated. On another interpretation, EP tells me both that I am rationally required not to choose stochastically dominated options, and that I am rationally permitted to choose any option that is not stochastically dominated. Let's call these principles Incomplete SD and Complete SD respectively.[10]

Incomplete SD seems to leave the regress problem unresolved. Consider an agent A facing options O_1, O_2, O_3, O_4. Suppose that O_1 stochastically dominates O_3 and O_4, but neither O_1 nor O_2 is stochastically dominated by any other option. Enkratic externalism, with EP understood as Incomplete SD, says that A is permitted (indeed, required) to rule out O_3 and O_4 straightaway on the basis of EP. But it does *not* say that A is permitted to choose O_1 or O_2 straightaway on the basis of EP, since EP (understood as Incomplete SD) *doesn't say* that O_1 is permissible or that O_2 is permissible—it simply *refrains* from saying that they're impermissible,

[10] Likewise, there is an incomplete interpretation of MEC that simply prohibits me from choosing any option for which there is an alternative option with greater expected choiceworthiness, while remaining silent on all other options, and a complete interpretation of MEC, according to which I am rationally permitted to choose all and only those options for which there is no option with greater expected choiceworthiness.

and indeed, from saying anything about them at all.

Now, suppose that A has positive credence in (at least) the following two second-order principles: N_1 "Under first-order moral uncertainty, always choose the option that minimizes the probability of the worst possible outcome," and N_2 "Under first-order moral uncertainty, always choose the option that maximizes the probability of the best possible outcome." Both of these norms are compatible with Incomplete SD, so the external force of this principle does not allow A to rule out N_1 or N_2. Suppose that according to N_1, A is rationally required to choose O_1 and according to N_2, A is rationally required to choose O_2. Then it seems that A has credence in two conflicting second-order principles, that the version of externalism we have adopted does not permit her to simply ignore this uncertainty, and therefore that she must decide what to do in light of this second-order uncertainty in light of third-order principles. Since she will presumably be uncertain in similar ways about third order principles, we're off to the races and the regress problem has been reinstated.

What this indicates is that the version of EP to which enkratic externalism ascribes external normative force must be *complete*, i.e., it must not simply prohibit certain options but rather classify *every* option in a given choice situation as either permissible or impermissible.[11] This allows formulations of EP like Complete SD, Complete MEC (according to which any option than which there there is no alternative with greater expected choiceworthiness is permissible), or a hybrid principle

[11] Arguably it would be sufficient for EP to always designate *at least one* option in any choice situation as permissible, even if it was silent with respect to other options, since the agent would then have an option that she was rationally permitted to choose straightaway despite her uncertainties. But I'll set this possibility aside for the sake of simplicity.

that applies stochastic dominance reasoning at some stages of the aggregation process and expectational reasoning at others, Complete (SD + MEC). It also allows the sort of hybrid view discussed in the last chapter, e.g. stochastic dominance + least common denominator, so long as the "least common denominator" principle is understood as having external normative force. Such a hybrid principle might be expressed as follows, where the subjective reason to choose an option O relative to a given comparability class of theories is determined by aggregating the objective reasons to choose O according to each theory, based on the strongest decision procedure common to all theories in the comparability class.

Hybrid EP An agent is rationally permitted to choose option O if and only if there is no alternative option P such that, for any degree of subjective reason d, the total probability of all comparability classes of theories according to which there is subjective reason to choose P greater than or equal to d is greater than or equal to the total probability of all comparability classes of theories according to which there is subjective reason to choose O greater than or equal to d, and for some d, the total probability of comparability classes according to which there is subjective reason to choose P greater than or equal to d is strictly greater than the corresponding probability for O.

So long as we attribute external normative force to a complete principle like Complete SD, Complete MEC, or Hybrid EP, therefore, we don't need to worry about the regress problem reappearing via incompleteness, since any of these principles will permit an agent to conclude straightaway that certain options are rationally permissible, despite her second-order uncertainties.

7.3.2 Normalization Uncertainty

We might worry, however, that even a complete version of the enkratic principle will not do away with second-order uncertainty and the possibility of regress. For both expectational and stochastic dominance reasoning in the context of intertheoretic uncertainty require procedures for intertheoretic normalization, about which agents may well be uncertain. And while it is plausible that rejecting stochastically dominated options, and perhaps even maximizing expected value, is constitutive of rationality and hence a belief-independent rational requirement, it is much less plausible that a particular intertheoretic normalization procedure is among the constitutive requirements of rationality as well.[12]

It seems to me, however, that uncertainty about normalization methods is not enough to generate a vicious regress of higher-order norms. Although, as in the last section, the problem of normalization uncertainty arises for both MEC and SD versions of the enkratic principle in roughly the same manner, it will simplify the exposition in this context to focus on MEC, where the role of normalization methods is more familiar. Consider by analogy a kind of normalization uncertainty that might afflict expectational reasoning under merely empirical uncertainty. Suppose that Gerald is a hedonistic utilitarian and is concerned with the hedonic welfare interests of non-human animals as well as humans. Suppose, however, that he is

[12] Here we see the force of Weatherson's worry that a "technical standard about resolving moral uncertainty" should not be among the normative standards to which we hold agents independent of their beliefs.

uncertain to what extent animals with relatively simple nervous systems experience pleasure and pain. Thus when he considers grasshoppers, for instance, he has some suspicion that the phenomenal intensity of the pain quale a grasshopper experiences when one of its C fibers fires is less than the phenomenal intensity of a pain quale a human being experiences when one of her C fibers fires, but he is unsure whether this is the case, and if so, how great the difference in phenomenal intensity of suffering per C fiber firing is. Thus although he cares both about preventing human C fibers from firing and about preventing grasshopper C fibers from firing, he is unsure how to compare these two values—i.e., unsure how to normalize the scale of human C fiber firings with the scale of grasshopper C fiber firings.[13]

Presumably this uncertainty about normalization methods will not involve Gerald in a vicious regress. Rather, given his commitment to maximizing expected hedonic utility, he will simply average over the various possible normalizations in order to decide how to weigh human pain against grasshopper pain. Likewise, a *morally* uncertain agent who is committed to maximizing expected choiceworthiness but is unsure how to normalize the various moral theories in which she has positive credence should, in light of her commitment to MEC, simply take a probability-weighted average over possible normalizations. This does not involve the introduction of some extra principle beyond MEC—rather, this sort of averaging is just what expectational reasoning amounts to in the first place.[14]

[13] This question might have practical significance if, say, Gerald is a researcher who must decide whether to devote his time and energy to developing a more effective antidepressant for humans or a more humane pesticide for grasshoppers.

[14] Likewise if the formulation of EP that has external normative force is a stochastic dominance principle, computing probabilities that a given option has a degree of choiceworthiness greater than or equal to d *involves* accounting for one's uncertainties about how to normalize the value scales of

7.4 The Problem of Non-Ideal Rationality

It may seem at this point that we've successfully avoided the threat of regress. But there is another sort of regress problem that confronts non-ideal reasoners under moral uncertainty, and that even a complete form of enkratic externalism alone does not obviously resolve.

7.4.1 Consequentialism and the "Heuristic Regress"

To see this non-ideal regress problem, consider again the analogy with empirical uncertainty. Suppose I'm certain that hedonistic utilitarianism is true and that, when I am uncertain about the consequences of my actions, I ought to maximize expected hedonic utility. In principle, this requires that for each option O, I consider every possible world that has non-zero probability conditional on my choosing O, assign conditional probabilities to each of those worlds based on my evidence, and calculate the total quantity of hedonic utility (pleasure minus pain) in each of those worlds.

Of course, for any agent with anything like human cognitive capacities and limitations, this is impossible. So what I must do instead is come up with some *heuristic* for approximating the expected value of each option, relative to my evidence. I might, for instance, choose to only focus on certain kinds of consequences (e.g. QALYs saved or lost). I might choose to restrict the scope of the consequences

a particular normative theory with the scale of objective choiceworthiness *simpliciter* over which the stochastic dominance principle is defined.

I consider, ignoring as improbable or unknowable the effects on my actions on the distant future or regions of the world outside some immediate radius of effect. I might choose to only consider certain evidence, e.g. a small number of relevant peer-reviewed studies. Or I might simply defer to certain experts, e.g. taking my estimates of the effects of some philanthropic activity from the estimates of an organization like GiveWell.

The problem is that there are many possible heuristic procedures for estimating the hedonic consequences of my actions, some better than others. Some will require me to invest more time in deliberation, others less. Some will ignore potentially important consequences, others will divert my deliberative resources toward consideration of trivial consequences. Just like my moral theory requires me, as best I can, to choose the *actions* with greatest expected hedonic utility, so it requires me to adopt the *deliberative procedures* with greatest expected hedonic utility. And just as I cannot compute the expected utility of my actions directly, I cannot compute the expected utility of adopting a given deliberative heuristic directly. So it seems that, in order to decide what deliberative heuristic to use, I must employ some *meta*-heuristic. For instance, I might attempt to mimic the deliberative processes of another agent whom I regard as a particularly good deliberator, or I might try to remember the consequences that ensued on previous occasions when I deliberated in a particular manner and choose the deliberative heuristic with the best track record, or... But just as I am uncertain which first-order heuristic is optimific in expectation, I will be uncertain which of these second-order heuristics is optimific in expectation. And so, it appears, I am doomed to regress.[15]

Just as a full commitment to maximizing expected hedonic utility still apparently leaves me vulnerable to a regress of higher-order heuristics, so a full commitment to maximizing expected choiceworthiness leaves me vulnerable to a similar regress.[16] As a non-ideal agent, I lack the cognitive resources to compute the expected choiceworthiness of my options directly, and so must reply on heuristics. I must try to choose the heuristic with greatest expected choiceworthiness (i.e., presumably, the heuristic that will yield first-order choices with greatest expected choiceworthiness). But there are many possible heuristics for estimating expected choiceworthiness, and I am unsure which is best. And so on.

An ideal agent who has perfect access to her own credal states and unlimited cognitive resources is not vulnerable to this regress, since she doesn't need to rely on heuristics in the first place—instead, she can compute the expected utility or expected choiceworthiness of her options directly on the basis of her credences (and, presumably, apportion those credences directly to her evidence without the need for epistemic heuristics).

Structurally, this heuristic regress problem resembles the original regress of higher-order norms that enkratic externalism was meant to avoid. The difference is simply that the norms at each level are not final principles of rational requirement which the agent believes to be *true* with certain probabilities, but rather heuristic principles for approximating the expected value of options, given certain lower-order

[15] This problem was brought to my attention by Ben Holguín, who to the best of my knowledge is the first to have noticed it.

[16] Once again, it is convenient to focus on EE^{MEC} rather than EE^{SD}, although the heuristic regress problem arises on either view.

inputs, which the agent believes to be *optimal* with certain probabilities.

Luckily, I will argue, there is a solution to this new regress problem, uniquely available to non-ideal agents. Describing this solution will allow us to address a question on which I have so far been silent: namely, how non-ideal reasoners can apply (or rather, approximate) complex decision procedures for choice under moral uncertainty like those described in the last two chapters.

7.4.2 Non-Ideal Deliberation under Moral Uncertainty: The Power of Priors

I propose that we adopt the following model of non-ideal practical deliberation.[17]

1. A non-ideal (= cognitively limited) but rational (= enkratic) agent A deliberating between options $O_1, O_2, ...O_n$ enters deliberation with access to a number of *standing beliefs*: in particular, (i) a belief about which of her options have maximal expected choiceworthiness, (ii) a belief about the expected marginal benefit of further deliberation to update her beliefs about the expected choiceworthiness of her options in light of her evidence, and (iii) a belief about the expected marginal cost of such deliberation.[18] Call the set of options that an

[17] Again, for ease of presentation, I assume EE^{MEC} throughout the following exposition, i.e., I assume that a rational agent always chooses options that have maximal expected choiceworthiness, whatever her beliefs about MEC. The exposition can be easily adapted to other versions of enkratic externalism, e.g. by replacing all references to "maximal expected choiceworthiness" with "not stochastically dominated."

[18] These beliefs can be understood dispositionally. Thus, an enkratic agent's standing belief about which of her options is most choiceworthy in expectation can be identified with the answer to the question, "If you had to choose *right now*, which of your options would you choose?" Likewise, insofar as she is rational, her beliefs about the expected marginal benefit and cost of further

agent believes, at a given moment of deliberation, to have maximal expected choiceworthiness her *enkratic set*.

2. Deliberation, for a non-ideal but rational agent, consists in updating her beliefs about the choiceworthiness of her options in light of her evidence. While an ideal agent can update costlessly (instantaneously and effortlessly), a non-ideal agent cannot. Rather, she must invest time in considering the relevance of the available evidence to her various empirical and normative beliefs. This might consist, among other things, of applying heuristics, developing meta-heuristics by which to assess those heuristics, applying the meta-heuristics, and so forth. But each of these steps requires scarce resources of time and effort.

3. As an agent deliberates, her standing beliefs will change, sometimes consciously and sometimes unconsciously. In particular, she may update the content of her enkratic set (i.e., change her beliefs about which of her options have maximal expected choiceworthiness), and she will also change her beliefs about the expected marginal benefit and cost of further deliberation. (Strictly speaking, this latter process may not involve any *belief change*, since her beliefs about the expected marginal value of deliberation are time-indexed, i.e., the value of deliberating for one minute longer at time t is distinct from the value of deliberating for one minute longer at time $t + 1$.)

4. In nearly all circumstances (apart from cases of genuine decision-theoretic paradox), the expected marginal cost of further deliberation will come to ex-

deliberation can be inferred from her dispositions to continue deliberating or to stop deliberating and act straightaway.

ceed the expected marginal value after a finite period of deliberation. This happens for at least two reasons: (a) Deliberation will in general have diminishing marginal benefit. A rational agent will first examine the strongest and most important evidence, consider the most important and most uncertain features of her various options, etc. The longer she deliberates, the more of the low-hanging fruit of deliberation she will exhaust, and the less she can expect further deliberation to improve her ultimate choice, per unit of time invested. (b) Nearly all choices are to some extent time-sensitive—that is, the quality of the agent's options will start to decay if she deliberates too long. While it can sometimes be rational to delay certain choices even for many years in order to acquire new evidence and improve my beliefs (Wise, 2013), it is never or almost never rational to delay forever, and typically not rational to delay for very long at all. If I wait until I am 45 years old to choose a career path, for instance, the deliberation this delay enables will be more than offset by the decreased quality of my career options.

5. Whenever a non-ideal agent comes to believe that the expected marginal cost of further deliberation would exceed the expected marginal benefit, she is rationally required to stop deliberating and choose based on her standing beliefs—that is, choose one of the options presently belonging to her enkratic set.[19]

Figure 7.1 offers a graphical illustration of the model. An agent A enters

[19] Note that this model has both empirical and normative components. The empirical component is the assumption that agents begin with certain standing beliefs that change in the course of deliberation. The normative component is the claim that an agent ought to deliberate so long as the expected marginal benefit of deliberation exceeds the expected marginal cost, and ought to cease deliberating and choose straightaway as soon as this inequality is reversed.

into deliberation believing that some subset of her options have maximal expected choiceworthiness, but also believing that the expected marginal benefit of further deliberation exceeds the expected marginal cost, and hence that she should deliberate further before making her choice. As she deliberates, her beliefs about which of her options have maximal expected choiceworthiness change, as do her beliefs about the costs and benefits of further deliberation. After some finite period of deliberation, the lines representing expected marginal cost and benefit of further deliberation cross (at what I have labeled the "decision point"). At that point, A rationally ought to choose one of the options in her enkratic set, i.e. one of the options that she then regards as having maximal expected choiceworthiness.

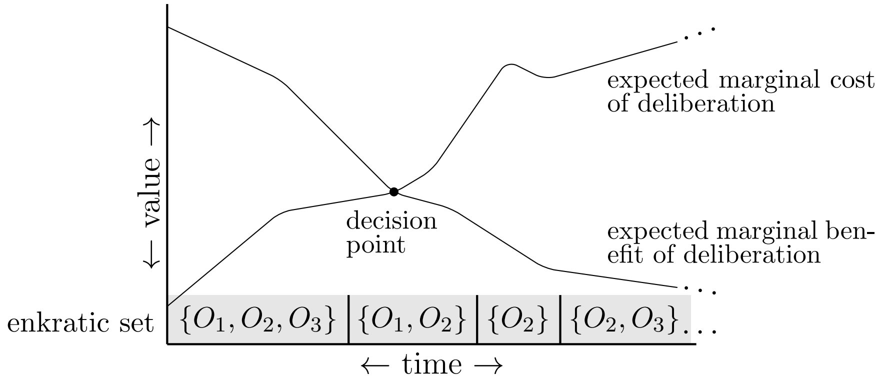

Figure 7.1: Non-Ideal Practical Deliberation

This model of deliberation strikes me as psychologically realistic. Consider a simple case of non-moral deliberation: I walk into an ice cream shop intending to buy a cone. I am immediately confronted with a large set of practical options. I begin with some standing belief about which of my options is most choiceworthy in expectation (choiceworthiness in this circumstance being, presumably, tightly correlated with my own hedonic utility): If I had to choose straightaway then I would choose, say, mint chocolate chip. But rather than choosing straightaway, I stand a few steps away from the counter, deliberating, since I'm uncertain which option will bring me most hedonic utility and it seems to me that the potential improvement of my choice to be had by a brief period of deliberation outweighs the cost of delayed ice cream. So I deliberate, imagining what each of my options would be like, calling up memories of the last time I ate various flavors of ice cream, perhaps recalling my slight disappointment the last time I chose mint chocolate chip. After thirty second of deliberation my standing beliefs about my options change, and I come to believe that chocolate raspberry is most choiceworthy in expectation. After another ten seconds, I conclude that further delay is unlikely to significantly improve my choice, and I choose chocolate raspberry straightaway. This seems to me a realistic description of the sort of deliberation that actual, non-ideal agents engage in all the time, even though most of us do not explicitly invoke terms like "expected choiceworthiness" and "marginal value" when we deliberate about ice cream.

How does all of this solve the heuristic regress problem, though? Simply thus: A non-ideal agent is *not* rationally required to continue ascending the hierarchy of higher-order heuristics until achieving certainty. It *may* turn out that what she rationally ought to do, in some circumstances, is compare the advice of various heuristic decision procedures, evaluate those heuristics in light of higher-order heuristics, etc. But at some point the rational thing to do is to cease deliberating and act, rather than engage in *any* form of further deliberation.

To put it in terms of the argument for externalism with which we began this chapter, we need not accept externalism with respect to heuristic norms for estimating expected choiceworthiness, because we can reject premise (1) of the argument for externalism when considering non-ideal, cognitively limited agents. Such an agent need not always account for her uncertainty about the optimality of a given heuristic norm, because it may be positively irrational for her to invest the time and effort needed to do so.

This doesn't defeat the regress argument for externalism, because the premises of that argument are still true with respect to *ideal* agents for whom the sort of deliberation I have just described (updating beliefs about the expected choiceworthiness of options based on one's credal state and evidence) is costless. The argument therefore succeeds in showing that pure internalism has unacceptable consequences with respect to the requirements of ideal rationality. But since the *heuristic* regress problem only arises for non-ideal agents, it can be solved by the observation that, with respect to such non-ideal agents, premise (1) is false.

In summary: The regress problem does force us to accept that at least one normative principle has external, belief-independent rational requiring force. I have argued that this one principle is the enkratic principle, rightly formulated, and considered some candidates for that right formulation, without attempting to give any final resolution of the question. Enkratic externalism, I have argued, solves the regress problem for ideal agents by providing a second-order principle of decision-making under first-order moral uncertainty that obviates the need for any hierarchy of higher-order normative principles. A different regress problem seems to confront non-ideal agents who must use heuristic norms to approximate ideal principles like

MEC. But, I have argued, a plausible model of non-ideal practical deliberation shows that such agents are not rationally required to keep ascending the infinite hierarchy of heuristic norms, resolving this problem as well.

7.5 Conclusion

I have had two main objectives in this dissertation. The first objective has been to defend the rational relevance of moral uncertainty, and more specifically the rationality of moral hedging (at least for many agents in many circumstances). My positive case for these claims has centered on the *enkratic conception of rationality*, according to which the requirements of practical rationality are grounded in an agent's beliefs about objective value or choiceworthiness. But I have also attempted to defend hedging against a variety of worries and objections from the recent literature.

My second objective has been to develop, in outline, a theory of rational decision-making under moral uncertainty. Here the central idea has been that of *content-based aggregation.* I have argued that, if how an agent should aggregate the value assignments of her competing theories depends on non-surface features of those theories' content, then the picture of rational decision-making that is likely to emerge is one of multi-stage aggregation of a nested structure of comparability classes bound together in various ways and to varying degrees by shared theoretical propositions.

The philosophical investigation of rational choice under moral uncertainty is still in its infancy, and even if the central claims I have advanced in this dissertation are correct, many more questions remain to be answered. Nevertheless, I hope to have shown that there is good reason to consider these questions and that we can have some reasonable hope of arriving at satisfactory answers.

Appendix A: Moral Uncertainty for Deontologists

[Note: The following is the penultimate draft of "Moral Uncertainty for Deontologists," forthcoming in *Ethical Theory and Moral Practice* (Tarsney, forthcoming b). In it I argue that stochastic dominance provides a natural aggregation procedure for deontological theories, supporting a threshold principle on which it is rationally impermissible to choose options that carry a risk greater than .5 of violating a deontological constraint when there are less morally risky alternatives available. Though the paper focuses on a representation of deontological value assignments suggested by Colyvan et al. (2010) that has the structure of the extended real number line $[-\infty, +\infty]$, the arguments are equally applicable to merely-ordinal or multidimensional, lexical representations of deontological theories.]

A.1 Introduction

We are often uncertain about what we morally ought to do. Such uncertainty can arise in two importantly distinct ways. On the one hand, some moral uncertainty is grounded in empirical uncertainty: for instance, is this substance that I am about to put in my friend's coffee sweetener, or is it arsenic (Weatherson, 2014)? On the other hand, there is what we might call *purely* moral uncertainty, uncertainty about the basic principles of morality, which no empirical information could directly or easily resolve: for instance, is it permissible to tell my friend a white lie about his new haircut?

Moral consequentialists have typically had the most to say about uncertainties of both kinds. With respect to empirically based moral uncertainty, consequentialists typically claim that agents should choose practical options that have maximal *expected value* (the expected value of an option O being found by summing, over all possible outcomes of that action, the value of that outcome times its probability conditional on one's choosing O). This expectational formula can be adjusted in a variety of ways, e.g. to allow for risk aversion or risk seeking.

The study of *purely* moral uncertainty is still in its infancy, but to date the dominant positive approaches in this literature have shared an almost exclusively consequentialist, expectational flavor: Lockhart (2000), Ross (2006), Sepielli (2009), and MacAskill (2014), for instance, all defend expected value approaches to decision-making under purely moral uncertainty. The critics of this approach, on the other hand, have predominantly been those who deny the need for a theory of rational choice under purely moral uncertainty, because they deny that what an agent ought to do depends on her purely moral beliefs or degrees of belief (e.g. Weatherson (2014), Harman (2015), Hedden (2016)).[1]

Expectational approaches to uncertainty, however, seem ill-suited for if not

[1] It is standard in the literature to reserve the term "moral uncertainty" for what I am calling *purely* moral uncertainty. I use the term more broadly in this paper both to avoid repeated inelegant references to "morally relevant empirical uncertainty" and because one of my aims will be to emphasize the continuity between empirically-based and purely moral uncertainty, so it is helpful to have an umbrella term that refers to both at once.

actively incompatible with non-consequentialist moral theories like Kantian deontology, most obviously because (unlike, say, classical utilitarianism) there is no natural way of interpreting these non-consequentialist theories as assigning degrees of rightness and wrongness that are amenable to being multiplied by probabilities and summed to yield quantitative expectations. Even setting aside any thought that theories of choice under uncertainty must take an expectational form, we will see in the next section that there are substantial reasons to doubt whether Kantians et al can provide any plausible answer to certain inescapable questions about moral choice under uncertainty.

The purpose of this paper, however, is to suggest one way in which deontologists *can* say something definite and precise about decision-making under moral uncertainty, of both the empirically based and purely moral varieties. I will propose that *stochastic dominance reasoning* provides powerful motivation for a *threshold principle* of choice under uncertainty, which implies *inter alia* that when an agent A faces a choice between practical options O and P, where P is known for certain to be morally permissible and the status of O is uncertain, A subjectively ought not choose O if her credence that O is objectively impermissible is greater than or equal to .5.

In the next section I describe in greater detail the challenge to deontology posed by both kinds of moral uncertainty, which some in the recent literature (e.g. Jackson and Smith (2006, 2016) and Huemer (2010)) have characterized as a decisive objection at least to absolutist forms of deontology. In §A.3 I introduce the idea of stochastic dominance reasoning. In §A.4 I show that stochastic dominance reasoning yields substantive and plausible conclusions for how an agent who accepts the general framework of deontological, agent-centered constraints (whether absolute or defeasible) ought to act under both kinds of moral uncertainty, which in at least the simplest cases will take the form of a straightforward threshold principle. Finally, §A.5 addresses a powerful objection to threshold principles given by Jackson and Smith, that such principles seem to violate "ought" agglomeration, since a pair of actions each of which is below the threshold for acceptable moral risk can, in combination, exceed that threshold. I argue that, in combination with the observation that deontological moral evaluation is relativized to the particular choice situations in which acts arise, the stochastic dominance approach can overcome this objection and preserve "ought" agglomeration.

A.2 Absolutist Deontology and Uncertainty

Following Portmore (2016, p. 7), a deontological ethical theory may be characterized as one that denies what all consequentialist theories affirm, namely, that the objective deontic status and degrees of objective rightness/wrongness of any act are determined by how the outcome or prospect of that act compares to the outcomes or prospects of its alternatives.[2] More precisely (to avoid accidentally counting as deontological views that deny that acts have deontic status or degrees of rightness), a deontological theory is one that at least sometimes attributes objective deontic status and/or degrees of objective rightness to acts for reasons that do not depend

[2] A prospect is an objective probability distribution over outcomes; hereafter I will use "outcome" to mean "outcome or prospect." I will likewise speak of "degrees of rightness" when I mean "degrees of rightness or wrongness."

on how the outcomes of those acts compare to the outcomes of alternatives. An imperfect litmus test for whether a theory is deontological is whether it endorses the existence of *agent-centered constraints*, act types that are treated as *prima facie* objectively wrong or prohibited and such that a token of that act type is at least sometimes objectively wrong or prohibited even when its performance would prevent multiple future tokens of the same act type that are similar in all ethically relevant respects to the prohibited token.[3] Thus a deontological theory might hold, for instance, that it is wrong for me to lie now even though my lie will prevent three future acts of lying that are similar to mine in all ethically relevant respects (except that they will not themselves prevent future instances of lying).[4]

An *absolutist* deontological theory holds that, for some act types characterized non-relationally (i.e., whether an act is an instance of the act type does not depend on any fact about how it relates or compares to its available alternatives), not only is their deontic status not fully determined by how their outcomes compare to those of alternatives, but all acts of that act type have the same deontic status regardless of *any* facts about their outcomes or those of alternatives. For instance, an absolutist theory might hold that all acts that have the non-relational property of *being a lie* (asserting something that the agent believes to be false with the intent that another

[3] The litmus test is imperfect because ethical theories that incorporate these *prima facie* deontological features can be "consequentialized" (Portmore, 2011). As we will see below, a desideratum for deontological approaches to uncertainty is to avoid a drift in the direction of expectational reasoning about the risk of constraint violations, which, if fully embraced, amounts to evaluating actions by a comparison of their prospects and hence to consequentialism, despite the presence of agent-centered constraints.

[4] Agent-centered constraints may also be positive, e.g., I am required to keep a promise even if my doing so will result in several fewer future instances of promise keeping, or more instances of promise violation.

agent believe it, in a circumstance where the second agent has not consented to being so deceived) is wrong even when telling the lie would prevent the deaths of a thousand innocent people who will be killed if the agent does anything other than lie.

Deontological theories in general and absolutist theories in particular face the following challenge: How should an agent decide what to do when she is uncertain whether a given act is an instance of a deontologically prohibited (or required) act type? For instance, consider the following case.

Possible Promise I have unexpectedly found myself in possession of tickets to the Big Game this afternoon. But as I am celebrating my good fortune, it occurs to me that I may have promised my friend Petunia that I would help her repaint her house later today, at just the time when the game is being played. I have a vague recollection of making such a promise but can't remember with any confidence. I have no immediate way of contacting my friend and must decide, here and now, whether to go to the game at the risk of breaking a promise, or go to her house, avoiding the moral risk but missing the game.

Possible Promise is an instance of empirically based moral uncertainty, but we can describe a closely related case of "pure" moral uncertainty.

Dubious Promise A week ago, Petunia sent me a text message asking if I would help paint her house today. I replied, saying that I would. Unbeknownst to Petunia, however, I was in the hospital at the time, recovering from a minor operation and under the influence of a fairly strong narcotic painkiller. By the

> time the influence of the painkiller subsided, I had completely forgotten my conversation with Petunia, and only just remembered it a moment ago, while planning my trip to the Big Game.

I take it that, under these circumstances, I may reasonably be uncertain whether I am morally required to skip the game and help my friend, even if I am certain of a background deontological conception of morality on which an ordinary, fully capacitated promise would have been morally binding.[5] Because my moral uncertainty does not trace back to any empirical uncertainty (e.g. about what I told Petunia, my mental state at the time, or her expectations of my future behavior), it is pure moral uncertainty, uncertainty about the content of a basic moral principle.

In either version of the case, the question is what I should do given my uncertainty about what morality requires of me. To frame this problem rightly, we must introduce a familiar distinction between *objective* and *subjective* normative properties. The objective normative properties of an act depend on facts about the world, but not (in general) on facts about an agent's beliefs or evidence. The subjective normative properties of an act, conversely, depend on facts about the agent's beliefs and/or evidence but not on facts about the world independent of the agent's beliefs/evidence. For instance (borrowing a helpful illustration from Portmore (2016)), suppose you are tasked with defusing a bomb and must decide whether to cut the green wire or the red wire. You believe, and your evidence overwhelmingly indicates, that cutting the green wire will defuse the bomb while cutting the red wire

[5] Adjust the strength of the imagined painkillers as needed until this case strikes you as one of substantial moral uncertainty.

will cause the bomb to explode. But in fact, cutting the red wire will defuse the bomb, while cutting the green wire will cause the bomb to explode. In this case, it is objectively right but subjectively wrong to cut the red wire, and objectively wrong but subjectively right to cut the green wire.[6]

This distinction in hand, let's return to the case of the Possible Promise. There are various things a deontologist can say about this case. However, it is generally agreed that she should *not* say any of the following: (1) It is subjectively wrong for me to go to the game iff I have *any* positive credence that I promised to help Petunia paint her house and hence that it would be *objectively* wrong for me to go to the game. (2) It is subjectively wrong for me to go to the game iff I am *certain* (i.e., credence 1) that I promised to help Petunia paint her house and hence that it would be objectively wrong for me to go to the game. (3) It is subjectively wrong for me to go to the game iff I *did in fact* promise to help Petunia paint her house and hence it *would in fact* be objectively wrong for me to go to the game.

Answer (1) is unsatisfactory because I can rarely if ever be *certain* that a

[6] Among the possible objective normative properties of an act are being objectively right or wrong; being objectively obligatory, permissible, or forbidden; being what the agent objectively ought or ought not do; and being required, permitted, or forbidden in virtue of an agent's all-things-considered objective reasons. For each of these objective properties, we may define a subjective counterpart. In general, whether an action has a given subjective normative property will depend on the agent's beliefs and/or evidence regarding the corresponding objective property. For our purposes, the distinctions between the various normative properties in each category will not matter very much, so I will sometimes switch between talking e.g. of objective wrongness, objective prohibition, and objective "ought not" as convenience dictates. Our central focus, however, will be on the interaction between an agent's beliefs about the objective properties of being (deontologically) *morally required* or *morally prohibited* and the subjective property of being *rationally required* or *rationally prohibited*, that is, being required or prohibited in virtue of an agent's all-things-considered subjective reasons. These concepts are related by way of objective reasons: a rational agent who believes, e.g., that she is deontologically morally required to choose some practical option O will believe that she has decisive or at least very strong objective reason to choose O, and (at least in general) it is an agent's beliefs about her objective reasons that give rise to her subjective reasons.

course of action does not violate a promise or some other source of deontological obligation. Hence, the principle behind (1) implies that all or nearly all acts are subjectively wrong. Answer (2) is unsatisfactory for a symmetrical reason: I can rarely if ever be certain that an act *would* violate a deontological obligation, and hence if those obligations are only ever relevant to the subjective normative properties of my actions under conditions of certainty, they are effectively inert. Answer (3) is unsatisfactory because, by collapsing subjective deontic status into objective deontic status, it fails to serve the purpose for which subjective normative concepts are introduced, namely to provide epistemically imperfect agents with useful action guidance in the face of uncertainty.

Absent a better answer than these, deontology is in trouble. It will seem either to paralyze action entirely (as per (1)), to collapse into consequentialism insofar as its agent-centered constraints are never practically relevant to epistemically imperfect agents (as per (2)), or to be simply useless as guides to action for such agents (as per (3)). Thus Jackson and Smith (2006, 2016) and Huemer (2010), *inter alia*, have suggested that the problem of uncertainty poses a fatal objection at least to absolutist forms of deontology.

There may be many lines of response more promising than (1)-(3) that the deontologist can pursue in the face of this challenge. But the major focus of the recent literature has been on variants of the *threshold view*, according to which, for any objectively prohibited act type W there is some threshold t, $0 < t < 1$, such that it is always subjectively wrong for me to choose a practical option O if my credence that O would be an instance of W is greater than or equal to t.[7]

The threshold view faces an important and widely acknowledged technical objection, which we will address in §A.5. But it also faces much simpler worries about *motivation* that have yet to be seriously addressed. First, why should there be any credal threshold at which an act abruptly switches status, from subjectively permissible to subjectively impermissible? What plausibility can there be, say, in the idea that if I have .27 credence that going to the game would break a promise to my friend, there is nothing at all wrong with going, but if I have .28 credence, I am strictly prohibited from going? Second, and closely related, what could serve to make any *particular* credal threshold the right one? Why should the threshold be set *here* rather than *there*? And how can we ever hope to *know* where it has been set?[8]

In the next two sections, I will propose an answer to these questions: namely, that *stochastic dominance reasoning* both explains why there should be sharp thresholds for subjective permissibility and fixes the value of those thresholds in a perfectly non-arbitrary manner, thereby vindicating a version of the threshold view.

[7] Jackson and Smith (2006) treat the threshold view as the most natural response the absolutist deontologist can give to the problem of uncertainty, though as we will see in §A.5 they argue that it is subject to a conclusive objection. Hawley (2008) and Aboodi et al. (2008) both defend versions of the threshold view against this objection. Isaacs (2014) suggests that it is subjectively permissible for an agent to take a morally risky action only if she *knows* that that action would not be an instance of a deontologically prohibited type, which is a close cousin of the threshold view provided one accepts that belief above some credal threshold is a necessary condition for knowledge.

[8] This objection from arbitrariness is pressed by both (Portmore, 2016, pp. 10-11) and (Jackson and Smith, 2016, p. 284).

A.3 Stochastic Dominance

The idea of dominance reasoning is familiar to philosophers from game-theoretic contexts like the prisoner's dilemma. If I am uncertain about the state of the world, but certain that, *given any possible state of the world*, option O is more choiceworthy than option P, then O is said to *strictly dominate* P. If I am certain (i) that, given any possible state of the world, O is *at least as choiceworthy* as P, and (ii) that given some state(s) of the world O is more choiceworthy, then O is said to *weakly dominate* P.

Stochastic dominance extends these familiar ideas as follows. A practical option O stochastically dominates an alternative P iff, relative to the agent's doxastic or epistemic state

1. For any degree of choiceworthiness d, the probability that O is (or will turn out to be) at least as choiceworthy as d is equal to or greater than the probability that P is ("") at least as choiceworthy as d, and

2. For some degree of choiceworthiness d, the probability that O is ("") at least as choiceworthy as d is strictly greater than the probability that P is ("") at least as choiceworthy as d.

An illustration: Suppose that I am going to flip a fair coin, and I offer you a choice of two tickets. The Heads ticket will pay you \$1 for heads and nothing for tails, while the Tails ticket will pay you \$2 for tails and nothing for heads. The Tails ticket neither strictly nor weakly dominates the Heads ticket because, if the coin lands

Heads, the Heads ticket will yield a more desirable outcome. But the Tails ticket *does* stochastically dominate the Heads ticket. There are three possible outcomes of the game, which in ascending order of desirability are: winning \$0, winning \$1, and winning \$2. The two tickets offer the same probability of an outcome *at least as good as* \$0, namely 100%. Likewise, they offer the same probability of an outcome at least as good as \$1, namely 50%. But the Tails ticket offers a better chance of an outcome at least as good as \$2, namely 50%, versus the 0% probability offered by the Heads ticket.

The principle that it is never rational to choose a stochastically dominated option is extremely compelling. First, note that unlike the stronger principle that a rational agent should always maximize expected utility or expected value, the stochastic dominance principle places no unwelcome constraints on an agent's attitude toward risk. For instance, if I am offered a choice between a ticket that pays \$1 if a fair coin lands heads on a single flip and a ticket that pays \$4 if a fair coin lands tails on both of two flips, it may be rational for me to take the risk averse option and select the first ticket even though the expected payoff of the second ticket is twice as large. Unlike an expectational principle, stochastic dominance is silent in this sort of case: the second ticket offers a better chance of a payoff at least as good as \$4 (namely, 25% rather than 0%), but the first ticket offers a better chance of a payoff at least as good as \$1 (namely, 50% rather than 25%), so neither option is stochastically dominated.[9] Relatedly (as we will see at greater length in

[9] Of course, expected utility theory will only requires that you choose the riskier ticket if the payoffs are given in utiles or some other unit of pure value, rather than in dollars. But the point can be made just as easily in these terms, since it is commonly believed that it is sometimes rationally permissible to be risk averse with respect to utility.

the next section), stochastic dominance reasoning does not entail the implausible results of expectational reasoning with respect to very small chances of very large or infinite payoffs (as in Pascal's Wager), since paying some definite cost for a very small chance of a very large reward will not stochastically dominate the null option of forgoing both the cost and the chance of reward.

Given these observations, it is hard to see how one can ever make the case for selecting a stochastically dominated option. It is, in general, *possible* that the stochastically dominated option will yield a more desirable payoff than the dominant option, but whatever level of payoff one is most concerned with getting, the dominant option offers an equally good or better chance of a payoff at least that desirable. The *way* in which that payout comes about is, by definition, a matter of indifference: If, say, I preferred winning \$0 with a Heads ticket to winning \$0 with a Tails ticket in the first case of stochastic dominance reasoning given above, then the desirability of the possible outcomes would depend on more than the monetary payout, and the Tails ticket would no longer stochastically dominate the Heads ticket. If I do *not* have such a preference, then the mere fact that the Heads ticket *might* turn out to yield a more desirable outcome is not a sufficient reason for choosing it over the Tails ticket.[10]

In the next section, I will show that the principle of eliminating stochastically

[10] In more formal terms, a stochastically dominant option will be preferred to a stochastically dominated alternative by any agent whose utility function is monotonically increasing, i.e., for whom a given chance of a higher-value outcome is preferred to the same chance of a lower-value outcome, all else being equal (Hadar and Russell, 1969, p. 28). The notion of stochastic dominance defined above, and which I employ throughout the paper, is sometimes called *first-order* stochastic dominance, distinguishing it from higher-order stochastic dominance principles that place increasingly stringent constraints on an agent's risk attitudes.

dominated options can justify a threshold principle for choice under deontological moral uncertainty.

A.4 Stochastic Dominance and Deontological Uncertainty

Let's return to the pair of cases given in §A.2, in which I am unsure whether I am required to skip the Big Game and help my friend paint her house, either because I am empirically uncertain whether I told her I would help or because I am purely morally uncertain whether some past act constituted a morally binding promise.

Let's provisionally suppose the following about these cases: (1) The background deontological conception of morality that I accept is absolutist, and therefore treats the objective requirement to keep one's promises as lexically stronger than any non-moral reason (like the reasons stemming from my desire to see the game). (2) This means that (a) an act of promise keeping is *better* or *more choiceworthy* than any act that violates a moral obligation or is morally neutral (at least so long as the agent is appropriately motivated, e.g. by considerations of duty), and (b) an act of promise breaking is *worse* or *less choiceworthy* than any act that fulfills a moral obligation or is morally indifferent (at least so long as no special exculpatory condition applies).[11] (We will later consider what happens when we relax these

[11] It may seem that we are some risk of describing deontological appraisal of actions in evaluative rather than normative terms, when part of what distinguishes deontological approaches to ethics is that they take normative notions to be primary, where consequentialists give primacy to the evaluative. But nothing turns on the choice of evaluative or normative language—we can, for instance, take the description of one action as "better" than another from the standpoint of a given agent as mere shorthand for the normative claim that it is more strongly supported by the agent's reasons, without raising any new difficulties for the argument given in this section.

assumptions.)

Accepting these assumptions allow a decision-theoretic representation of deontological moral considerations, along lines suggested by Colyvan, Cox, and Steele (2010). In their model, absolutist deontology is characterized by the following axiomatic extension of standard utility theory:

D1* If [outcome] O_{ij} is the result of an (absolutely) prohibited act, then any admissible utility function u must be such that $u(O_{ij}) = -\infty$.

D2* If [outcome] O_{ij} is the result of an (absolutely) obligatory act, then any admissible utility function u must be such that $u(O_{ij}) = +\infty$. (Colyvan et al., 2010, p. 512)

On this model, while actions from duty have "infinite" positive value and actions against duty have "infinite" negative value, actions that neither violate nor fulfill duties have finite value determined, presumably, by prudential or desire-based reasons and perhaps by other sorts of moral reason (e.g. consequence-based reasons of benevolence) that do not generate absolute obligations or prohibitions.[12]

The mentions of infinite positive and negative value in Colyvan et al's axioms need not be taken too literally. As they point out (pp. 521ff), decision-theoretic models of non-consequentialist ethical theories may be descriptively adequate without being explanatory: A Kantian does not avoid lying, for instance, *because* she regards acts of lying as having infinite disvalue. Nevertheless, the fact that, for a

[12] Talk of the "value" or "utility" of actions should be understood as simply a less cumbersome stand-in for talk of *choiceworthiness*—i.e., it should be understood as denoting a normative rather than an evaluative property of actions.

Kantian, moral obligations and prohibitions are lexically stronger than prudential reasons can be accurately *represented* by treating the value of an action from duty as the upper bound on the scale of reason strength, and the disvalue of an action against duty as the lower bound.

This representation in hand, we can employ stochastic dominance reasoning to draw conclusions about how a committed deontologist ought to respond to moral uncertainty. Consider the case of the Possible Promise. Suppose that the prudential value of seeing the Big Game is +20, while the prudential value of helping Petunia paint is +5 (and that I know these prudential facts with certainty). It follows that the option of helping Petunia will stochastically dominate the option of going to the game if and only if the probability that I promised to help Petunia is greater than or equal to .5.

Suppose, first, that the probability is exactly .5, that is, I regard it as equally likely that I did as that I did not make a promise to Petunia. Then I may reason as follows. My decision has four possible outcomes (Table A.1), which in order from best to worst are: (i) I did make a promise to Petunia, but I go to the game, violating that promise ($-\infty$). (ii) I did not make a promise to Petunia, but skip the game to help her paint anyway (+5). (iii) I did not make a promise to Petunia, and go to the game, violating no obligation (+20). (iv) I did make a promise to Petunia, and I skip the game to help her paint, thereby fulfilling that obligation ($+\infty$).

Given my credences, helping Petunia paint (option P) stochastically dominates going to the game (option G): P and G have the same chance of producing an outcome at least as good as $-\infty$ (100%); P has a better chance of producing an outcome at least as good as +5 (100% vs. 50%); P and G have the same chance of producing an outcome at least as good as +20 (50%); and P has a better chance of producing an outcome at least as good as $+\infty$ (50% vs. 0%) (Table A.2). Thus, if I accept the principle that I should never choose a stochastically dominated option, I am compelled to choose P rather than G.[13]

	Promised (.5)	**Didn't Promise (.5)**
Painting	$+\infty$	+5
Game	$-\infty$	+20

Table A.1: Possible Promise v1, payoff matrix

	$-\infty$	**+5**	**+20**	$+\infty$
Painting	1	1	.5	.5
Game	1	.5	.5	0

Table A.2: Possible Promise v1, probability of payoff $\geq x$

On the other hand, suppose that my credence that I promised to help Petunia paint is only .49. In that case, option G has a better chance than option P of yielding an outcome at least as good as +20 (51% vs. 49%), so P does not stochastically dominate G (Tables A.3-A.4).

	Promised (.49)	**Didn't Promise (.51)**
Painting	$+\infty$	+5
Game	$-\infty$	+20

Table A.3: Possible Promise v2, payoff matrix

[13] The talk of "outcomes" should not be seen as illicitly consequentialist. Violating or fulfilling a deontological obligation is an "outcome" only in the formal sense of being a distinct act-state combination valued differently than other act-state combinations. Calling these "outcomes" does not imply, for instance, that the wrongness of breaking a promise has anything to do with its causal consequences.

	$-\infty$	**+5**	**+20**	$+\infty$
Painting	1	1	.49	.49
Game	1	.51	.51	0

Table A.4: Possible Promise v2, probability of payoff $\geq x$

Notice that this argument applies to Dubious Promise, the case of pure moral uncertainty, just as it applies to Possible Promise, the case of empirically based moral uncertainty. In either case, the option of helping my friend will stochastically dominate the option of going to the game iff my credence that I am objectively morally obligated to help my friend is greater than or equal to .5. Thus, it seems, stochastic dominance reasoning gives the deontologist something definite, precise, and well-motivated to say *both* about choices made under morally relevant empirical uncertainty *and* about choices made under pure moral uncertainty.[14]

An immediate worry: We have so far assumed that the background deontological conception is absolutist. But many philosophers who identify as deontologists and endorse characteristically deontological moral phenomena like agent-centered constraints do *not* think of deontological constraints as absolute.[15] Fortunately,

[14] Interestingly, the view that it is permissible to take a morally risky act only when the probability that the act is not objectively wrong is greater than .5 has a history in the Catholic moral theology literature, under the name "probabiliorism" (Sepielli, 2010, pp. 51-2). Stochastic dominance, in these terms, implies that a deontological absolutist must adopt a position *at least* as rigorous as probabiliorism in the cases of "asymmetrical" moral risk with which this literature is chiefly concerned.

[15] The literature on deontological moral uncertainty is divided in focus between absolute and non-absolute versions of deontology. Jackson and Smith (2006) direct their criticisms at absolutist theories. Aboodi et al. (2008) suggest that this is a mistake, since "hardly any (secular) contemporary deontologist is an absolutist" (p. 261, n5). But Huemer (2010, p. 348, n3) marshals a credible array of apparent (contemporary or near-contemporary) deontological absolutists. Note that the difficulties of absolutism arise for anyone who takes one class of reasons to be lexically stronger than another: assuming that there is always, in any choice situation, a non-zero risk that the lexically stronger reasons are in play (e.g. that a given course of action *might* cause some innocent person torturous suffering), one may worry that the lexically weaker reasons (e.g. headache prevention) will always be preempted by that risk and hence never rise to the level of practical relevance.

But in any case, the challenge posed by uncertainty is no less acute for non-absolute deontologists. The non-absolutist must still answer the question of how an agent should decide what to do when however, the stochastic dominance argument does not turn on the assumption of absolutism. Following Colyvan et al, who conclude that the absolutist version of deontology is implausible, we may represent the rightness of fulfilling an obligation not as $+\infty$ but as merely a very large finite positive number, i.e., as a reason of finite strength though much stronger than ordinary prudential reasons. And likewise, we may represent the wrongness of violating an obligation not as $-\infty$ but as merely a very large finite negative number.[16] Suppose that keeping a promise to my friend has a value of $+9001$ while breaking that a promise has a value of -9001. It is easy to see that the stochastic dominance argument for helping Petunia paint will go through *mutatis mutandis.*

Note, however, that a reliance on stochastic dominance rather than expectational reasoning lets us avoid one of the chief pitfalls for absolutist approaches to uncertainty, namely, the risk of "Pascalian paralysis": If violating a moral obliga-

she is unsure whether a given course of action would violate a deontological constraint. As with the absolutist, the most natural answer she can give is some version of the threshold view, but to defend this view she will have to overcome all the same obstacles as the absolutist (Jackson and Smith, 2016, pp. 287-8). The only new option that the non-absolutist position seems to open up is an expectational view on which the subjective reason-giving force of a deontological constraint for an uncertain agent is the product of the inherent stringency of that constraint (which is finite, since the constraint is non-absolute) times the probability that a given course of action would violate the constraint. But if she goes this expectational route, it is no longer clear that her view should be counted as a form of deontology, rather than simply a version of consequentialism that incorporates agent-centered constraints. (Cf. discussion in Portmore (2016). Portmore allows that a deontologist may give an expectational account of subjective rightness without becoming a consequentialist, so long as the underlying account of *objective* rightness remains deontological. But he argues convincingly that the challenges of uncertainty arise not just in cases where we are ignorant of facts (e.g., what the consequences of some action will be or would have been) but also in cases where there is *no fact of the matter* about whether some course of action would have constituted a constraint violation (e.g., due to physical indeterminism or counterfactual underdetermination), and that therefore the deontologist who wishes to go the expectational route must give a partially expectational account of *objective* as well as subjective rightness, which he takes it *would* amount to a form of consequentialism. For a different line of argument against expectational approaches to deontological moral uncertainty, see Tenenbaum (2017).)

[16] See Colyvan et al (2010, pp. 515-8) for an axiomatic characterization of such a non-absolutist deontological theory.

tion is treated as an outcome with the value of $-\infty$, then actions that are almost certainly morally permissible but carry even a vanishingly small risk of violating a moral obligation will carry an expected value of $-\infty$—the same expected value, in fact, as acts that are *certain* to violate a moral obligation.[17] Worse still, if (as seems plausible to me) *every* possible action has some non-zero chance of fulfilling a moral obligation and some non-zero chance of violating a moral obligation, then the expected value of every possible action is undefined ($\infty + (-\infty)$).

One way to avoid Pascalian difficulties is to hold that a rational agent may never have non-zero credence in any outcome with infinite positive or negative value. But this seems implausible and has only *ad hoc* motivation. The better response is to modify our decision theory, either weakening or amending expected utility theory in a way that allows an agent with modest credence in infinite values and disvalues to nevertheless remain responsive to finitary considerations. Stochastic dominance is one such weakening of expected utility theory: As far as stochastic dominance principles are concerned, it is rationally permissible for me to go to the game, despite the risk of infinite moral turpitude, so long as that risk has a probability less than .5.[18]

The .5 credal threshold for permissibility follows from stochastic dominance

[17] In the literature on "pure" moral uncertainty, this is often described as the worry that "fanatical" moral theories will hijack expectational reasoning—i.e., moral theories that attribute infinite value and disvalue to options, perhaps in very strange or counterintuitive ways, will take precedence over all finitary moral theories, so long as one assigns them even the most vanishingly small degree of positive credence (Ross, 2006, pp. 765-7).

[18] Of course, stochastic dominance need not be understood as the *only* or the *strongest* principle of rational requirement under uncertainty (cf. the last three paragraphs of this section). Various principles are possible that occupy an intermediate position between stochastic dominance and expected utility maximization and that might impose a more rigorous form of moral caution on agents acting under uncertainty while still avoiding the problem of Pascalian paralysis.

reasoning when an agent is certain of all her non-moral reasons. But when she is uncertain which option her non-moral reasons favor, stochastic dominance may become more demanding. Suppose, for instance, that in the Possible Promise case I am uncertain whether I would have a better time at the game or painting with my friend. Perhaps I believe that there is a one-in-three chance that my team will lose, and I know that while seeing my team win would give me a utility of $+20$, seeing them lose would give me a utility of -10. If I simply ignore the game and help my friend paint, on the other hand, I am guaranteed a utility of $+5$. In this case, the degree of belief that I promised my afternoon to Petunia at which stochastic dominance will require that I do so is reduced—specifically, to .4 rather than .5.[19] But in simple cases of deontological moral-prudential conflict, where I am certain that one option O is morally permissible but prudentially worse than another option P, and certain that P is prudentially better than O but uncertain whether it is morally permissible, we may say that stochastic dominance requires me to choose option O iff the probability that P is objectively impermissible is greater than or equal to .5.[20]

[19] Suppose my credence that I promised to help Petunia paint is exactly .4. There are five possible outcomes, with values of $-\infty$, -10, $+5$, $+20$, and $+\infty$. Options P and G have the same chance of producing an outcome at least as good as $-\infty$ (100%); P has a better chance of producing at outcome at least as good as -10 (100% vs. 60%); P has a better chance of producing an outcome at least as good as $+5$ (100% vs. 40%); P and G have the same chance of producing an outcome at least as good as $+20$ (40% — the chance that G produces an outcome at least this good is a product of the .6 chance that I do not violate an obligation in going to the game and the $.\overline{6}$ chance that my team wins the game); and P has a better chance than G of producing an outcome at least as good as $+\infty$ (40% vs. 0%) (Tables A.5-A.6). If, however, my credence that I made the promise were any lower, then G would have a better chance than P of producing an outcome at least as good as $+20$ and P would no longer stochastically dominate G.

[20] Cases of this sort have been a major focus of the recent literature on "pure" moral uncertainty, in particular the cases of abortion and vegetarianism, both of which seem to present agents (at least in many cases) with a conflict of prudential reasons on one side and uncertain moral reasons on the other. See for instance Guerrero (2007), Moller (2011), and Weatherson (2014) for discussion of these cases.

	Prms (.4)	**¬Prms & Win (.4)**	**¬Prms & Lose (.2)**
Painting	$+\infty$	$+5$	$+5$
Game	$-\infty$	$+20$	-10

Table A.5: Possible Promise v3, payoff matrix

	$-\infty$	-10	$+5$	$+20$	$+\infty$
Painting	1	1	1	.4	.4
Game	1	.6	.4	.4	0

Table A.6: Possible Promise v3, probability of payoff $\geq x$

Of course, even allowing for the possibility of a higher threshold in cases of prudential uncertainty, the requirements of stochastic dominance will not rise to the level of practical stringency that we intuitively expect at least some deontological constraints to possess. The kind of reasoning I have described makes no distinction between "weaker" constraints (like the constraint against telling white lies) and "stronger" constraints (like the constraint against killing the innocent).[21] And while a threshold of .5 may seem plausible for the weaker constraints, it seems much less plausible when we think about the stronger constraints—surely a deontologist should not conclude, for instance, that a judge need not worry about the constraint against punishing the innocent so long as the probability that her sentencing decision violates that constraint is a mere .49 rather than .51.

Importantly, however, stochastic dominance is only a *sufficient* condition for rational requirement, not a *necessary* condition. Everything I have said so far leaves it open to the deontologist to argue for a higher threshold, at least for some cate-

[21] Except, on a non-absolutist view, when there is some chance that the constraint in question is outweighed by other kinds of considerations—since the chance of a stronger constraint being outweighed will presumably be smaller, *ceteris paribus*, the threshold of moral safety at which the option that risks violating that constraint ceases to be stochastically dominated will be higher.

gories of constraint, on grounds other than stochastic dominance. And even if one takes the laxity of the stochastic dominance threshold to be intuitively unacceptable with respect to any sort of deontological constraint, progress has still been made insofar as the deontologist can now say with confidence that you must *at least* believe it to be more likely than not that your morally risky action violates no constraint, even a relatively weak constraint, before non-deontological (e.g. prudential or consequentialist) reasoning may take over.

Nevertheless, it seems to me that this objection from the variable stringency of constraints represents a residual difficulty that it will be hard for deontologists to fully overcome. It is intuitive to hold that the practical force of deontological constraints must be sensitive to both (i) the *probability* of a constraint being violated by a given course of action and also (ii) the *seriousness*, *importance*, or *stringency* of the particular constraint in question. But taken together, these intuitions push us in the direction of an expectational view that, while it may still incorporate agent-centered normative considerations, has lost the rest of its distinctively deontological character (see note 15 *supra*). The deontologist may resist this pressure, but to do so she must be prepared at some point to deny at least one of the above intuitions.

A.5 Option Individuation and Ought Agglomeration

So far I have argued that stochastic dominance reasoning offers a foundation for a threshold view of deontological moral choice under both empirically based and pure moral uncertainty. But such threshold views face another important difficulty, which we have yet to confront.

Jackson and Smith (2006) argue that threshold views violate the principle of "ought" agglomeration, that $ought(A)$ & $ought(B) \vdash ought(A$ & $B)$, in cases where an agent is faced with the possibility of performing two acts, each of which individually is below the threshold for permissible moral risk, but which in combination exceed that threshold. They illustrate this difficulty by way of the following thought experiment.

Two Skiers Two skiers are headed down a mountain slope, along different paths. Each, if allowed to continue, will trigger an avalanche that will kill ten innocent people. (If both are allowed to continue, they will trigger two such avalanches and kill both groups of ten.) The only way to save each group is to shoot the corresponding skier dead with your sniper rifle. You can shoot either skier individually or, being an extraordinarily crack shot, you can shoot both with a single bullet. The moral theory you accept (with certainty) tells you that you ought to kill culpable aggressors in other-defense, but are absolutely prohibited from killing innocent threats. Unfortunately, you are uncertain of the intentions of the skiers, assigning them each equal and independent probabilities of acting obliviously, and thus as innocent threats, rather than as ill-intentioned aggressors.

Suppose you subscribe to a threshold view on which you ought to kill a potential aggressor in other-defense iff your credence that he is acting innocently is less than t. And suppose that, with respect to each of the two skiers, your credence that he is innocent is *just* less than t. Thus, it seems that you ought to shoot Skier 1 and you ought to shoot Skier 2. But, if you shoot *both* Skier 1 and Skier 2, the chance that *one* of them is innocent and hence that you will have violated a deontological constraint is *greater* than t and hence, it seems, you ought not perform the compound action: *shoot Skier 1 and shoot Skier 2.*

It is immediately tempting to point out that *shooting Skier 1* and *shooting Skier 2* are two separate actions, and that all that should matter about the case from a moral standpoint is whether the chance of either action individually violating a deontological constraint is greater than t. And I will shortly argue that something very much like this is the right response for the deontologist to give. But Jackson and Smith seem to have headed off this response: While shooting each skier separately looks like two separate actions, shooting both skiers *with the same bullet*, as they have imagined you have the option of doing, is a single action, with a chance greater than t of violating a deontological constraint. But surely it is implausible that you are permitted (indeed, required) to shoot both skiers with separate bullets, but prohibited from shooting them both with the same bullet.

A plausible solution to the agglomeration problem must therefore find some appropriate criterion for "individuating" the shooting of Skier 1 and the shooting of Skier 2 that allows us to distinguish the two sources of moral risk, even given the option of shooting both skiers with a single bullet. Aboodi, Borer, and Enoch (2008) offer one such proposal. They suggest that, if all deontological constraints can be thought of as *rights* possessed by particular rightsholders, a deontological agent acting under uncertainty ought to ensure that she does not run a risk greater than the threshold t of violating the risks of any *particular* rightsholder, though she may permissibly run a risk greater than t of violating the rights of *some* rightsholder. They suggest this might be justified, *inter alia*, by the broadly contractualist thought that what determines the permissibility of my conduct is whether anyone can reasonably or legitimately object to it: A rightsholder cannot reasonably object to my conduct merely because it creates *some* risk that her rights will be violated, nor because it creates a large aggregate risk that *someone's* rights will be violated. But she can reasonably object if I run an excessive risk of violating *her rights in particular*. Thus, Aboodi et al conclude, in the Two Skiers case you ought to shoot both skiers, because in so doing there is no rightsholder that you run a risk greater than t of wronging. (Skier 1 and Skier 2 are each rightsholders, but there is no compound entity, Skier 1 + Skier 2, that is a rightsholder or can lodge objections against your conduct.)

This patient-based approach, however, has seriously counterintuitive consequences, as Huemer (2010) points out. Suppose, as I have argued holds true in the simplest sort of case, that the threshold for permissible risk of violating a deontological constraint is .5, and consider the following case (a slightly simplified retelling of Huemer's "War Options").

The Weapon Minerva, a military officer, faces a situation in which she knows that, if she does not act, a powerful weapon will fall into the hands of a ruthless enemy, who will use it to kill 100,000 innocent people. Fortunately, there are two surefire ways to prevent the weapon from falling into enemy hands: (1) Minerva

> has received intelligence indicating that a certain scientist will shortly reveal the plans for constructing the weapon to the enemy. It is unclear whether the scientist is willingly helping the enemy or acting under exculpating duress (say, threats on the life of his family). Minerva thinks it slightly more likely than not (.55) that he is acting under duress and that it would therefore be objectively wrong to kill him even to save a greater number. (2) In addition to the weapon plans, the enemy needs certain materials to construct the weapon, which are being kept in a small town of 25,000 people. Minerva knows that if she orders the carpet bombing of this town, the weapon materials will be destroyed, but 45% of the town's population will be killed. Since she has no specific information about which members of the town's population will be at risk, the epistemic probability that any individual townsperson will be killed is .45.

On Aboodi et al's patient-centered approach, it looks as though Minerva is permitted to carpet bomb the town (since there is no rightsholder whom she runs a risk greater than .5 of violating) but is not permitted to assassinate the scientist (since in doing so she would run a risk greater than .5 of wronging him).[22]

Fortunately, there is another way of resolving the agglomeration problem that both avoids these counterintuitive consequences and follows naturally from a central feature of the deontological approach to ethics: Because deontological moral assessment of options is relativized to the particular *choice situations* in which those

[22] Of course, the particular value of the threshold plays no role in this argument—the case can be easily modified to accommodate any threshold greater than or less than .5.

options arise, it seems to me, deontological moral theories simply cannot and do not assess *compound* options, i.e., combinations of options that arise in different choice situations. Thus, while *shoot Skier 1* and *shoot Skier 2* are each individually subject to deontological moral assessment, the compound option *shoot Skier 1 & shoot Skier 2* is not. But since it is a theory's moral assessment of options that provides the basis for stochastic dominance reasoning, which in turn grounds the threshold principle for deontological moral risk, this means that the threshold principle is simply inapplicable to compound options, and hence does not generate a prohibition against shooting both skiers. Rather, we are left free to make the natural inference from the fact that you ought to shoot Skier 1 and ought to shoot Skier 2 to the conclusion that you ought to shoot both skiers, preserving "ought" agglomeration.

Let's spell out this line of reasoning in a bit more detail. Deontological moral assessment is relativized to choice situations in the following sense: Just as deontological theories (in contrast, for instance, to agent-neutral consequentialist theories like classical utilitarianism) direct me to distinguish between the rights and wrongs of my own actions and those of others, and hence, for instance, not to commit one rights violation to prevent someone else from committing five, so likewise they direct me to distinguish between the rights and wrongs *of the choice immediately before me* and the rights and wrongs of my other choices, past, present, or future. Thus for instance a deontological theory does not direct me to lie now even if it is the only way of extricating myself from a situation in which I can predict with certainty that I will tell five lies. To put this in terms of the quantitative representations we employed in the last section, a deontological theory will assign the option of lying now the large negative value (either finite or infinite) associated with violating a constraint, but will not assign this same degree of disvalue to the option of not lying, even though this option carries with it the certainty of future constraint-violating lies, which will themselves be assigned the disvalue of constraint violations in the choice situations where they arise.[23]

Colyvan et al. (2010) put this point by suggesting that deontological moral duties are both agent-relative and *time*-relative (p. 513), but this is not quite right, since one can make multiple choices at the same time, and in such cases the relativization of deontological considerations remains to the individual choice situation, rather than to the time at which both choices are made.[24] For instance, consider the following case:

The Buttons of Wrongness A mad ethicist has rigged a contraption that will test your judgments about two ethical dilemmas at once. You may press one of four buttons, colored red, green, yellow, and blue. If you press the red button, a message will be sent from your phone to one of your friends, telling him a white lie about his recent haircut. If you press the green button, $1000 will be stolen from the accounts of a large corporation and donated to GiveDirectly. If you press the yellow button, *both* these things will happen. If you press the blue button, *neither* will happen. Pressing any of the buttons will deactivate the rest, and you will have no other opportunity to reverse the effects of your

[23] A deontological theory should likely assign *some* disvalue to options that I know will lead me to future constraint violations, but the point is that it does not assign the same kind or degree of disvalue that it assigns to options that directly violate a constraint.

[24] In fairness, Colyvan et al may not have meant to assert otherwise, since later on the same page they put the point in terms of choice situations rather than times.

selection once it has been made.

There is a sense, clearly, in which you make only one choice in this scenario, namely *which button to press.* But in a more ethically relevant sense, you make two choices, namely *whether to tell a white lie* and *whether to steal money from the large corporation for GiveDirectly*—even though you will put both of these choices into effect by means of a single action, namely, pressing a button. And to the deontologist, what you will do with respect to one of these choices simply has no bearing on what you are permitted to do with respect to the other—not because they are potential violations of two different rightsholders (it would make no difference if your friend with the new haircut was also the majority shareholder of the large corporation) but simply because they are *two different choices.*[25]

What does this mean for Jackson and Smith's Two Skiers case and the agglomeration objection to threshold views? Simply this: Because deontological moral assessment is choice situation relative, deontological moral theories are just not in the business of assessing compound options like *shoot Skier 1 & shoot Skier 2.* Just like the options of telling your friend a white lie and stealing for GiveDirectly in the

[25] This is not to say that deontologists are committed to metaethical possibilism, the view that what I ought to do with respect to a particular choice situation can *never* depend on what I believe I will do with respect to other choice situations. For instance, the deontologist might still allow that in the classic test case for actualism vs. possibilism (Jackson and Pargetter, 1986, p. 235), Professor Procrastinate should decline the invitation to review a recently published book, even though he knows himself to be the person best qualified to review it, because he also knows that given his inveterate tendency to procrastinate, he will not finish the review on time. The point is rather the narrower one that *the strength of a deontological moral prohibition or requirement* is not altered by one's expectations about how one will behave in other choice situations. (Thus, what allows the deontologist to take the actualist line with respect to Professor Procrastinate is that reviewing the book is, one assumes, morally optional. If it were the case, for whatever reason, that Professor Procrastinate were morally obligated to agree to review the book, then the deontologist should agree with the possibilist that he ought so to agree, whatever his expectations about any of his other choices or actions.)

Buttons of Wrongness case, these options inhabit different choice situations, *even if* they are (or can be) effectuated by means of a single action (viz., shooting both skiers with the same bullet). For this reason, we cannot say for instance that the option of shooting neither skier stochastically dominates the option of shooting both skiers: Stochastic dominance reasoning makes reference to a probabilistic assignment of degrees of objective value or choiceworthiness to options, but since deontological theories don't assign degrees of value or choiceworthiness to compound options, they can be neither stochastically dominant nor dominated. Since the threshold principle of choice under deontological moral uncertainty is grounded in stochastic dominance reasoning, this principle in turn is only applicable to simple options. Thus in the Two Skiers case, the threshold principle does not imply that you ought not shoot both skiers. Rather, we can and should reason, by the principle of agglomeration, from the fact that you ought to shoot Skier 1 and ought to shoot Skier 2 to the conclusion that you ought to shoot both skiers.

A.6 Conclusion

Moral uncertainty, of both the empirically based and purely moral varieties, presents a challenge for deontologists. They must give an account of how agents should deliberate and act in the face of such uncertainties, ideally an account that is as precise and intuitively well-motivated as the expectation-maximizing account available to consequentialists. I have suggested that a version of the threshold account grounded in stochastic dominance reasoning may meet this need. If we treat morally right/obligatory actions as having very large positive moral value (whether finite or infinite) and morally wrong/prohibited actions as having very large negative moral value (“”), then it will often turn out that morally risky options are stochastically dominated. In the simplest sort of case, where only one of two options carries moral risk, and that option is certain to be preferable *but for* the risk of violating a moral requirement, it turns out that that option is stochastically dominated if and only if the probability that it is objectively prohibited is greater than or equal to .5, though this threshold may be less than .5 in more complex cases. The principle that one rationally ought not choose stochastically dominated options is extremely compelling, representing a weakening of expectational reasoning that both permits a variety of risk attitudes (risk aversion and risk seeking as well as risk neutrality) and avoids worries about moral fanaticism or Pascalian paralysis arising from tiny probabilities of infinite value and disvalue. It therefore provides an appealing basis for the threshold approach to deontological moral uncertainty, one that moreover (when combined with the right understanding of deontological moral evaluation) can avoid the most powerful objection to that approach. It seems, then, that deontologists are better equipped to respond to the challenges of moral uncertainty than many philosophers have thought.

Appendix B: What Grounds Intertheoretic Comparisons of Moral Value

The purpose of this appendix is to explain in greater detail what grounds the correctness of content-based value aggregation methods of the sort described in Chapters 5-6. More specifically, I aim to show (i) that it is *rational* for agents in certain doxastic states to aggregate the value assignments of the moral theories in which they have positive credence in particular ways, based on the content of the theories being aggregated, and (ii) that the propositions belief in which makes it *rational* for agents to aggregate in these ways are plausibly also *true*, making the content-based aggregation procedures they justify not only rational but *correct*.

The account I will give relies heavily on the Bayesian commitments taken on in Chapter 1. Specifically, I will assume that an ideally rational agent has credences, satisfying the probability calculus, defined over all propositions that can be constructed from her conceptual resources, and that those conceptual resources are not unduly impoverished. This allows us to describe such an agent as distributing belief over complete theories of the world, understood as maximal consistent sets of propositions ("maximal" relative to her conceptual resources), with her credence in all such theories summing to 1. For reasons that will become apparent, it will be useful to think of morally uncertain agents as dividing their beliefs not simply among moral theories (i.e., maximal consistent sets of *moral* propositions) but among theories of the world *simpliciter*, containing empirical propositions, moral propositions, logical/mathematical propositions, higher-order normative propositions about the requirements of rationality, etc. As I have done through most of the dissertation, I set aside questions of non-ideal or bounded rationality (except briefly in §B.5), and content myself with arguing that *ideally* rational agents would respond in particular ways to particular states of moral uncertainty.

For simplicity, I will take as my central example the Easy Case from Chapter 5, in which Alice divides her beliefs between hedonistic utilitarianism and a pluralistic theory that values both hedonic and aesthetic goods. We will introduce other cases as necessary to illustrate potential complications absent from the Easy Case.

B.1 The Easy Case: Decision-Theoretic Compatibility

Let's begin by giving a more thorough description of the Easy Case. Expanding on and precisifying the description of the case from Chapter 5, we may say that Alice has beliefs, *inter alia*, about the following propositions.

P_1 The choiceworthiness of an option is equal to the value of the world that results from it, where the value of a world is the cardinal sum of non-derivative value that supervenes on value-bearing empirical phenomena in that world.[1]

P_2 Hedonic experiences (i.e., experiences with positive or negative hedonic valence)

[1] "Value" being used in this context in a way not synonymous with "choiceworthiness," since only practical options and not worlds are direct objects of choice.

are non-derivative value-bearers.

P_3 The explanation/grounding for the non-derivative value of hedonic experience is φ.

P_4 Aesthetic goods (beauty and the like) are non-derivative value-bearers.

P_5 The explanation/grounding for the non-derivative value of aesthetic goods is ψ.

P_6 The value of one hedon is equal to the value of one aestheton.

P_7 Nothing apart from hedonic experiences and aesthetic goods is a non-derivative value-bearer.

In the simplest version of the case, Alice then divides her beliefs between two theories.

T_1 (hedonism): $\{P_1, P_2, P_3, \neg P_4, \neg P_5, \neg P_6, P_7, ...\}$

T_2 (pluralism): $\{P_1, P_2, P_3, P_4, P_5, P_6, P_7, ...\}$

When we introduced the Easy Case, we also stipulated that T_1 and T_2 share an expectational decision procedure. This fact might be represented in at least two importantly distinct ways. First, it might be that T_1 includes a proposition like

P_8 Every agent is always rationally required (subjectively ought) to maximize expected hedonic value.

while T_2 includes a proposition like

P_9 Every agents is always rationally required (subjectively ought) to maximize expected hedonic + aesthetic value.

But either of these propositions imply that what an agent rationally ought to do is entirely insensitive to her moral beliefs, which I have argued is implausible (Chapters 2-3).

More plausible versions of T_1 and T_2 might include the following propositions respectively.

P_{10} Every agent who has perfectly true moral beliefs (i.e., assigns every true basic moral proposition credence 1 and every false basic moral proposition credence 0) is always rationally required to maximize expected hedonic value.

P_{11} Every agent who has perfectly true moral beliefs (i.e., assigns every true basic moral proposition credence 1 and every false basic moral proposition credence 0) is always rationally required to maximize expected hedonic + aesthetic value.

Let's refer to the versions of T_1 and T_2 that include P_{10} and P_{11} as T_{1a} and T_{2a} respectively.

Neither P_{10} nor P_{11} say anything about how an agent who lacks perfectly true moral beliefs is rationally required to act. So a theory that includes either of these propositions *could* consistently include any number of further propositions that speak to this question. That is, a theory that includes P_{10} or P_{11} might endorse irrelevantism, MFT, MFO, or hedging, *inter alia*.

When I introduced the hedonism/pluralism case in Chapter 5, I claimed in effect that it would be most natural for an agent who divides her moral beliefs between T_{1a} and T_{2a} to accept a generalization of these principles like

P_{12} Every agent is always rationally required (subjectively ought) to maximize expected choiceworthiness.

and more generally, that it would be most natural for an agent who has positive credence in T_{1a} and T_{2a} to accept a principle like P_{12} conditional on the disjunction $T_{1a} \vee T_{2a}$.

If we wished, we could simplify our task by simply stipulating that Alice divides her belief between versions of T_1 and T_2 that both include P_{12}. But in any case, for T_1 and T_2 to be *maximal* consistent sets of propositions, they must include *some* proposition or propositions specifying the rational requirements that apply to morally uncertain agents. And this illustrates an important and often-overlooked point. Several critics of hedging have put forward some variant of the claim that "It's not the business of a moral theory to tell you what to do if you don't believe that theory" (e.g. (Hudson, 1989, p. 224), (Gracely, 1996, p. 331)). But this neglects the fact that a moral theory is part of a complete theory of the world (though, for a given agent, the same moral theory may figure in many complete theories of the world, given her uncertainties about non-moral matters), and that components of an agent's moral theory may derive from principles that have implications outside the domain of first-order morality. For instance, one way of being a classical utilitarian is to accept the following two propositions.

P_{13} The choiceworthiness of an action is wholly determined by the net hedonic utility of its consequences.

P_{12} Every agent is always rationally required (subjectively ought) to maximize expected choiceworthiness.

Considered as the logical closure of these two propositions, this version of utilitarianism *does* tell an agent what to do when she is morally uncertain.[2]

Finally: If it is true, as I argued in Chapter 7, that a principle like P_{12} has external, belief-independent normative force, then it is simply irrelevant to the question of what Alice is rationally required to do whether the theories over which she distributes her belief contain P_{12} or other, conflicting principles of rational requirement (like P_8 or P_9). She will be rationally required to maximize expected choiceworthiness regardless—at least, so long as it is possible for her to make the comparisons necessary to form an expectation.

I will therefore take it for granted that, insofar as it is possible for her to do so, Alice ought to aggregate the reasons proffered by T_1 and T_2 by computing the expected value of her options given the disjunction of the two theories. To summarize the preceding paragraphs, there are three possible routes to this conclusion: Given a normative internalist view on which the appropriate aggregation procedure for a class of theories is the strongest procedure on which those theories can agree,

[2] One might quibble, of course, over whether P_{12} is really a component of an agent's *moral* theory, and claim that moral theories must contain only propositions like P_{10} that are conditioned on an agent's having perfectly true moral beliefs. But I cannot see why this sort of semantic division between moral theories and the more general theories of the world in which they are embedded should make any difference for our purposes, i.e., with respect to the requirements of rationality for an agent in a particular epistemic/doxastic state.

we might (i) simply stipulate that the versions of T_1 and T_2 in which Alice has credence both include P_{12} (the principle that Alice ought to maximize expected choiceworthiness), or we might (ii) stipulate that the versions of T_1 and T_2 in which Alice has credence include P_{10} and P_{11} respectively, and then argue that an agent who accepts P_{10} conditional on the other propositions we have stipulated as elements of T_1 and accepts P_{11} conditional on the other propositions we have stipulated as elements of T_2 ought to accept P_{12}, as a natural generalization of P_{10} and P_{11}, conditional on $T_1 \vee T_2$. Finally, we might (iii) take the externalist line defended in Chapter 7 and hold that the rational requirement stated by P_{12} is incumbent on agents regardless of their beliefs (about P_{12} or anything else). To simplify the following discussion, however, I will take option (i) and simply stipulate that Alice's beliefs are such that $P_{12} \in T_1$ and $P_{12} \in T_2$.

The principle of maximizing expected choiceworthiness stated by P_{12}, however, is non-vacuous only so long as the expected value of options is defined. The crucial question, therefore, is whether T_1 and T_2 meet the preconditions for defining expectations over them, namely structural compatibility and scale normalizability. Let's now turn to this question.

B.2 The Easy Case: Structural Compatibility and Scale Normalizability

In Alice's case, structural compatibility is trivial. Since $P_1 \in T_1, T_2$, both theories have one-dimensional cardinal structure.[3] The real issue, then, is scale normalizability. The crucial claim in Chapter 5 that allowed the Easy Case case to go forward as an example of comparability was that the value of a hedon according to T_1 is equal to the value of a hedon according to T_2. So the task now is to explain this claim—partly to defend it, partly to generalize our conclusions to cases where it or its analogues fail.

My strategy for defending scale normalizability has two parts. First, I will claim that, conditional on $T_1 \vee T_2$, Alice ought to accept

P_{14} The degree of value borne by a given unit of hedonic experience is independent of whether aesthetic goods are non-derivative value-bearers.

But second, I will argue that, even if Alice does not accept P_{14}, she ought to accept some other proposition (or distribute her belief over some set of propositions) that establish a different normalization between T_1 and T_2, which still allows her to compute and maximize expected value.

To make the first claim more precise, I claim that P_{14} is very plausible given $\{P_1, P_2, P_3, P_7\}$ (i.e., given that hedons have non-derivative value, a complete and

[3] More specifically P_1 seems to guarantee ratio structure, since there is a meaningful zero point corresponding to any world that contains no value-bearers. But since an expectational decision procedure requires only interval structure, the issue of interval vs. ratio structure is unimportant.

fixed story about what grounds or explains that value, and that nothing besides hedons and aesthetons has non-derivative value), and therefore that, *ceteris paribus*, an agent who accepts $\{P_1, P_2, P_3, P_7\}$ ought to accept P_{14}. The *ceteris paribus* clause means that other features of an agent's epistemic/doxastic state, which we have so far left unspecified, might make it rational for her to reject P_{14}, but that in the default or normal case (leaving this intentionally vague), it will be rational to accept it. This is to say, then, that in the default description of the Easy Case, $P_{14} \in T_1, T_2$ (just as I suggested in the last section that, in the default description of the case where Alice accepts T_{1a} and T_{2a}, $P_{12} \in T_{1a}, T_{2a}$).[4]

Why is this so? One reason is P_3, or more precisely, the fact that $P_3 \in T_1, T_2$. It is a stipulation of the Easy Case that T_1 and T_2 are in full agreement, not just that hedonic experience has non-derivative value, but on everything to do with hedonic value, and in particular with respect to the underlying moral metaphysics that explains or grounds the value of hedonic experience. Given that, at least as far as anything we have said up to this point is concerned, T_1 and T_2 do not disagree about anything that has any direct connection to the issue of hedonic value, it is at least *prima facie* plausible that they assign the same *degree* of value to a given unit

[4] Of course, given the rational requirement of regularity, Alice's credence $\mathrm{Cr}(P_{14}|\{P_1, P_2, P_3, P_7\})$ cannot be 1—if rational agents ought to be at least a little uncertain about everything, then Alice ought to be at least a little uncertain whether the value of a hedon is independent of the value of aesthetons, even conditional on (the remaining propositions of) T_1 and T_2. Nevertheless, I will set this detail aside for the moment, and suppose for simplicity that, if P_{14} coheres with the remaining propositions of T_1 and T_2 better than any alternative, then Alice may simply distribute the entirety of her belief over versions of T_1 and T_2 both of which include P_{14}. Of course, the idea that a rational agent could distribute her belief over just two complete theories of the world is entirely inconsistent with the assumption of regularity, or indeed with much weaker assumptions about the limits of rational certainty. But it is a convenient simplification for the moment. We will later examine what happens when this assumption is relaxed, and consider a version of the Easy Case in which Alice is uncertain not just about propositions like P_4 and P_5 (i.e., about the value of aesthetons) but also about propositions like P_{14}.

of hedonic experience.

To plausibilify this point, consider an analogous case of empirical uncertainty: Suppose Bob is certain that hedonic, maximizing, expectational utilitarianism is true, and indeed is certain of *everything* except for the empirical question of what will happen if he presses the red button. More precisely, he divides the entirety of his belief between two complete theories of the world: T_3, according to which utilitarianism is true and pressing the red button will save five innocent people, and T_4, according to which utilitarianism is true and pressing the red button will kill ten innocent people. Call these two competing propositions about the effects of pressing the red button R_s and R_k. In order to compute the expected value of pressing the red button, Bob must make *some* assumption about how the value of a hedon conditional on R_s compares with the value of a hedon conditional on R_k. And of course, unless for some quite unusual reason his beliefs about the value of hedonic experience are linked to his beliefs about the consequences of pressing the red button, he will assume that the value of a hedon is independent of whether R_s or R_k is true—that is, he will accept the analogue of P_{14}.

Just as the value of a hedon is, *prima facie* or *ceteris paribus*, independent of what would happen if Bob presses the red button, so it strikes me as natural to think that the value of a hedon is, *prima facie* or *ceteris paribus*, independent of whether aesthetons are value-bearers as well. Though questions about the value of hedons and the value of aesthetons "belong to the same domain," in a way that questions about the value of hedons and the consequences of pressing the red button do not, in both cases the questions are logically orthogonal. There is no obvious reason why settling the question of whether aesthetons have non-derivative value should affect Alice's beliefs about the value of hedonic experience one way or the other. The burden is on the opponent of intertheoretic comparability, then, to explain why Alice should not accept P_{14}—for if she accepts P_{14}, and is rationally required to maximize expected choiceworthiness (either in virtue of accepting P_{12} or in virtue of P_{12} being an external rational requirement), then she will be rationally required to make intertheoretic comparisons and engage in moral hedging (since accepting P_{14} enables her to compute the expected choiceworthiness of options notwithstanding her uncertainty between T_1 and T_2). (This argument will be set out in greater detail in the next section.)

Note that, since we have assumed that T_1 and T_2 are *maximal* consistent sets of propositions, they cannot simply be silent on the question of how the value of a hedon given P_4 (that aesthetons have non-derivative value) compares with the value of a hedon given $\neg P_4$. Rather than P_{14}, these theories might contain propositions like the following.

P_{15} The value of a hedon if P_4 is true is exactly half the value of a hedon if P_4 is false.

P_{16} The value of a hedon if P_4 is true is determinately at least half, and no more than twice, the value of a hedon if P_4 is false, but indeterminate with respect to any intermediate ratio.

P_{17} The value of a hedon if P_4 is true is incomparable with the value of a hedon if P_4 is false.

A set of propositions that contained no proposition of this sort would not be maximal, since any of these propositions could be added to it while preserving consistency.[5]

A proposition like P_{15}, of course, would establish comparability just as much as P_{14}—it just normalizes the value scales of T_1 and T_2 differently.[6] A proposition like P_{16} would establish rough comparability. Opponents of any kind of intertheoretic comparability, on the other hand, must hold that Alice should accept P_{17}. And to do this, they must show a disanalogy with the case of Bob, who should clearly accept the analogue of P_{14} and not the analogue of P_{17}.

The only argument I am aware of that might support such a conclusion is the following: A moral theory's value scale is simply a representation of its ordinal preferences over options in conditions of empirical uncertainty—derivable from those preferences so long as they satisfy certain plausible constraints (e.g. the von Neumann-Morgenstern axioms), as shown by the representation theorems of standard expected utility theory (von Neumann and Morgenstern, 1947). But these representations are unique only up to positive affine transformation—that is, given

[5] One might deny this claim by denying that P_{14-17} are in fact coherent propositions, i.e., that it is meaningful to talk about how the value of a hedon conditional on P compares with the value of a hedon conditional on $\neg P$. But the case of Bob shows that this is false: If such propositions were generally incoherent, then expectational reasoning would be impossible. One might also claim that such propositions are incoherent only when the P on which we are conditioning the value of a hedon is a necessary rather than a contingent proposition. I will address this objection in §B.5.

[6] Rebecca Stangl has suggested to me in conversation that, on certain stories about the nature of value, discovering new sources of value in the world should lead us to regard other sources of value as less valuable than we previously thought they were. For instance, if the explanation for the value of hedonic experience (i.e., what fills in the variable φ in P_3) is that it contributes to living a good life, then learning that aesthetic experience *also* contributes to living a good life might lead Alice to conclude that hedonic experience is less important than she would otherwise have thought, since it is only *one* contributor to the good life rather than the *sole* contributor. In this case, Alice should accept something like P_{15} rather than P_{14}. But this is no obstacle to my thesis: T_1 and T_2 are still scale-normalizable, only with a different normalization (that gives comparatively more weight to T_1 and less to T_2).

any value assignment that represents a theory's preferences over options, any transformation of that value assignment of the form $mx + b$, for (positive) real-valued constant m and (positive or negative) real-valued constant b, will result in another value scale that represents the theory's preferences equally well. So, the claim goes, just as interpersonal utility comparisons are impossible in standard expected utility theory, where a person's utility function is just a representation of her preferences over risky prospects, intertheoretic value comparisons are likewise impossible—and just as the value of a hedon for person A is incomparable with the value of a hedon for person B, on standard expected utility theory, so the value of a hedon according to T_1 must be incomparable with the value of a hedon according to T_2.[7]

This argument strikes me as extraordinarily unconvincing. To begin with, it is simply wrong about the structure of many moral theories. Classical, hedonistic utilitarianism, for instance, does *not* derive its value scale from preferences that it enjoins on agents under empirical uncertainty. Rather, classical utilitarianism *starts from* a claim like P_1, that states of affairs have value in proportion to the value they contain (more specifically, the balance of pleasure minus pain), and *derives from that fact* the conclusion that agents in particular epistemic states ought to prefer some options to others. That is to say, for theories like classical utilitarianism, the axiological facts ground the normative facts, not the other way around. The representation theorems therefore play no role in these theories.[8]

[7] This argument is suggested by Ittay Nissan-Rozen, who writes that "I cannot think of any plausible alternative to the 'representation of moral rankings' explication of the term 'moral value', and under this explication intertheoretic comparisons of moral value are...meaningless." (Nissan-Rozen, 2015, p. 18)

[8] This strikes me as the correct way of understanding the utilitarianism of Bentham and Mill, but of course exegetical claims are beside the point. The point is that such a moral theory is

Moreover, the analogy to the problem of interpersonal utility comparisons points out the simple implausibility of this argument. In the interpersonal context, the implausible conclusion of standard expected utility theory that A's slow and agonizing death is not a greater harm to A than B's mild itch is to B shows us that individual utility or wellbeing is *not* simply a representation of preferences over risky prospects. Likewise in the intertheoretic context, the implausible conclusion that, for instance, the value difference between universal/perpetual happiness and universal/perpetual agony according to T_1 is not greater than the value difference between the actual world and the actual world plus one mild itch according to T_2 might be taken to show us that moral value is not simply a representation of a moral theory's ranking of risky options.

But finally, Nissan-Rozen's argument is simply question-begging as an argument against intertheoretic comparability or moral hedging. For he takes for granted that an agent may have rationally justified credence in a subjective first-order moral theory that ranks options under empirical uncertainty in a way that allows the construction of a value scale for that theory via the representation theorems, while ignoring the parallel possibility that an agent may have rationally justified credence in a *second-order* theory that ranks options under moral uncertainty in a way that allows the construction of an intertheoretic value scale. That is, if Alice can reasonably arrive at the ranking of options relative to her state of moral uncertainty that the intuitive normalization of T_1 and T_2 would suggest, then P_{14} will simply emerge from that ranking via the representation theorems, just as its analogue will emerge

possible, and furnishes a counterexample to Nissan-Rozen's argument.

from Bob's ranking of options in the red button case.

I conclude the following, then: (i) It seems *prima facie* rational for Alice, given what else we have said about her epistemic state, to accept P_{14} (i.e., to divide her belief between versions of T_1 and T_2 each of which includes P_{14}). (ii) The most likely scenario in which this is not the case is one in which Alice has reason to accept propositions like P_{15} or P_{16}, which still establish intertheoretic comparability and the basis for hedging, rather than P_{17}, which does not.

B.3 A Simple Argument for Hedging

We have now shown that Alice has the ability to make rationally well-grounded intertheoretic comparisons. We can now suppose, as per the arguments of the preceding sections, that Alice divides her belief exclusively between final versions of T_1 and T_2,

T_{1f} (hedonism): $\{P_1, P_2, P_3, \neg P_4, \neg P_5, \neg P_6, P_7, P_{12}, P_{14}, ...\}$

T_{2f} (pluralism): $\{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_{12}, P_{14}, ...\}$

such that her credence $\mathrm{Cr}(T_{1f}) = p$ and $\mathrm{Cr}(T_{2f}) = 1 - p$.

Since $P_{14} \in T_{1f}, T_{2f}$, both theories affirm that the value of a hedon according to T_{1f} is equal to the value of a hedon according to T_{2f}. And since all it is for the value of a hedon *according to the two theories* to be equal is for the theories to affirm that equality, we may simply say that the value of a hedon according to T_{1f} is equal to the value of a hedon according to T_{2f}. Further, since $P_6 \in T_{1f}$, the value of an aestheton according to T_{2f} is equal to the value of a hedon according to T_{2f}. So by transitivity of equality, the value of a hedon according to T_{1f} is equal to the value of an aestheton according to T_{2f}.

The following argument, then, appears to be valid.

1. Alice has credences $\mathrm{Cr}(T_{1f}) = p$ and $\mathrm{Cr}(T_{2f}) = 1 - p$.

C. The expected choiceworthiness of an option O, for Alice, is equal to $hedons(O) + (1 - p)(aesthetons(O))$.

This establishes, *a fortiori*, the possibility of intertheoretic comparison and aggregation for Alice.

The following argument is valid as well.

1. Alice is rationally required to maximize expected choiceworthiness.
2. The expected choiceworthiness of an option O, for Alice, is equal to $hedons(O) + (1 - p)(aesthetons(O))$.

C. Alice is rationally required to choose options that maximize $hedons(O) + (1 - p)(aesthetons(O))$.

This establishes, it seems to me, that the problem of intertheoretic value comparisons does not constitute a general objection to moral hedging: If it is possible for an agent to be under a rational requirement to maximize expected choiceworthiness (either in virtue of accepting a normative principle to that effect, or in virtue of such a principle having external, belief-independent rational requiring force), then in at least some states of moral uncertainty, those expectations will be well-defined and she will be rationally required to hedge. Someone who takes PIVA to constitute a general objection to hedging, therefore, must first claim that the expected choiceworthiness of options is *sometimes* undefined (i.e., there are some states of uncertainty that preclude well-defined expectations) and then argue on this basis that agents cannot *ever* be rationally required to maximize expected choiceworthiness. But it is not clear how such an argument might go, and to my knowledge no such argument has been put forward by the opponents of moral hedging.[9]

Notice that the arguments of the previous sections (to the effect that, given Alice's acceptance of $\{P_1, P_2, P_3, P_7\}$, she ought also to accept propositions like P_{12} and P_{14}) were strictly inessential to the argument of the present section that it is at least sometimes possible to make intertheoretic value comparisons and rational to engage in moral hedging: We could, if we like, simply *stipulate* that Alice accepts P_{12} and P_{14} along with P_{1-7}. The purpose of giving arguments rather than stipulations is to show, not just that there are *some* pairs of theories such that an agent who

[9] Note that the fact that expected choiceworthiness is sometimes undefined does not obviously problematize even the claim that rational agents are *always* rationally required to maximize expected choiceworthiness: In cases where expectations are undefined, this requirement is simply inert—so long as it is understood, as it must be, to say that an agent is always required to choose an option with maximal expected choiceworthiness *whenever such an option is available*, or alternatively, that an agent is always rationally prohibited from choosing an option O when there is some alternative option P such that (i) P has greater expected choiceworthiness than O and (ii) there is no option Q that has greater expected choiceworthiness than P. (Such qualifications are necessary not just for the case of undefined expectations but also for cases in which an agent faces infinitely many options with expected values of of 1, 2, 3, 4, ... or $1/2$, $2/3$, $3/4$, $4/5$, ... (Arntzenius et al., 2004; Landesman, 1995).) If we conclude that expected choiceworthiness is sometimes undefined, we might then claim that agents are always required to choose options with maximal expected choiceworthiness whenever such an option is available and that other requirements of rationality govern the cases where such an option is not available.

divides her beliefs between those theories ought rationally to hedge for her moral uncertainties and has rational grounds for the intertheoretic comparisons necessary to do so, but also that these theories are *reasonable* and internally coherent, rather than mere artificial combinations of jointly implausible propositions. This makes our conclusions more relevant to real agents who, we hope, divide their beliefs mainly among reasonable and internally coherent normative theories.

B.4 Generalizing (or Failing to Generalize) from the Easy Case

Nevertheless, the Easy Case is of course highly idealized, and we would like to have *some* indication that its conclusions generalize, particularly to agents who are more radically uncertain than Alice. The complications involved even in this simplest of simple cases indicate, I think, that this cannot be done in a single stroke—establishing content-based aggregation methods between theories takes work, in particular a careful specification of the content of those theories that will, unless we proceed by stipulatory fiat, involve substantive questions about the most plausible way of explicating an initially under-described theory.

Nevertheless, in this section I will briefly address a few particularly salient dimensions along which we might hope to generalize the idea of content-based intertheoretic aggregation. The general claim I hope to make plausible is that in many cases it is rational for an agent to accept an "independence assumption" analogous to P_{14} that permits scale normalization between theories, or otherwise to distribute her belief over propositions like P_{14}, P_{15}, and P_{16} that can be aggregated to yield at least rough normalization and comparability. But I will also attempt to show, at the end of the section, that this approach to intertheoretic comparisons leaves open the possibility of *in*comparability between theories.

B.4.1 Indirect Comparability

A noteworthy feature of the Easy Case is the presence of *moral certainty*, namely, Alice's certainty with respect to the existence and basic nature of hedonic value.[10] If this were a necessary feature of the case—that is, a necessary condition for content-based aggregation—that would be a discouraging result, since it seems that an epistemically rational agent will be at least a little uncertain about any feature of morality.

Fortunately, it is not. Consider the following, slightly less simple case: Carmen, like Alice, is a committed consequentialist, who is uncertain what things in the world are non-derivative value-bearers. The three candidates she takes as possible value-bearers are hedonic experience, aesthetic goods, and knowledge. She divides her beliefs between three theories: T_5 values hedonic experience and aesthetic goods, but not knowledge; T_6 values hedonic experience and knowledge, but not aesthetic goods; and T_7 values aesthetic goods and knowledge, but not hedonic experience.[11]

Representing things as we did in the Easy Case, we can say that Carmen has beliefs about the following propositions, *inter alia*.

[10] If Alice accepts P_{14} then she is also certain, we might say, about the *degree* of hedonic value. But even if she accepts P_{15} instead, or distributes her belief over propositions like P_{14-17}, she is still certain not only that hedons have value (P_2) but also of the explanation/grounding for that value (P_3).

[11] Thanks to Andrew Sepielli for this case and for pressing me to consider the challenge it represents.

P_1 The choiceworthiness of an option is equal to the value of the world that results from it, where the value of a world is the cardinal sum of non-derivative value that supervenes on value-bearing empirical phenomena in that world.

P_2 Hedonic experiences (i.e., experiences with positive or negative hedonic valence) are non-derivative value-bearers.

P_3 The explanation/grounding for the non-derivative value of hedonic experience is φ.

P_4 Aesthetic goods (beauty and the like) are non-derivative value-bearers.

P_5 The explanation/grounding for the non-derivative value of aesthetic goods is ψ.

P_{18} Knowledge is a non-derivative value bearer.

P_{19} The explanation/grounding for the non-derivative value of knowledge is υ.

P_6 The value of one hedon is equal to the value of one aestheton.

P_{20} The value of one hedon is equal to the value of one epistemon.

P_{21} The value of one aestheton is equal to the value of one epistemon.

P_{22} Nothing apart from hedonic experiences, aesthetic goods, and knowledge is a non-derivative value-bearer.

Carmen then divides her beliefs between the following theories:

T_5 (hedons + aesthetons): $\{P_1, P_2, P_3, P_4, P_5, P_6, \neg P_{18}, \neg P_{19}, \neg P_{20}, \neg P_{21}, P_{22}, ...\}$

T_6 (hedons + epistemons): $\{P_1, P_2, P_3, \neg P_4, \neg P_5, \neg P_6, P_{18}, P_{19}, P_{20}, \neg P_{21}, P_{22}, ...\}$

T_7 (aesthetons + epistemons): $\{P_1, \neg P_2, \neg P_3, P_4, P_5, \neg P_6, P_{18}, P_{19}, \neg P_{20}, P_{21}, P_{22}, ...\}$

Are T_{5-7} intercomparable? As in the Easy Case, this will depend on what Carmen believes about whether and how the value of potential value-bearers (hedons, aesthetons, and epistemons) varies between theories that endorse them as value-bearers. Suppose that Carmen believes the following propositions with certainty:

P_{23} Given that hedonic experience is a non-derivative value-bearer, the degree of value borne by a given unit of hedonic experience is independent of whether aesthetic goods/knowledge are non-derivative value-bearers.

P_{24} Given that aesthetic goods are non-derivative value-bearers, the degree of value borne by a given unit of aesthetic goods is independent of whether hedonic experience/knowledge are non-derivative value-bearers.

P_{25} Given that knowledge is a non-derivative value-bearer, the degree of value borne by a given unit of knowledge is independent of whether hedonic experience/aesthetic goods are non-derivative value-bearers.

Carmen's acceptance of these independence assumptions allows her to normalize and compare T_{5-7}: P_{23} says that the value of a hedon according to T_5 is equal to the value of a hedon according to T_6, establishing a normalization of those theories. Likewise, P_{24} says that the value of an aestheton according to T_5 is equal to the value of an aestheton according to T_7, and P_{25} says that the value of an epistemon according to T_6 is equal to the value of an epistemon according to T_7.

Note also that the presence of these independence assumptions—i.e., Carmen's belief that *how much* value is borne by one kind of value bearer does not depend on what other kinds of value-bearers exist—imposes consistency constraints on the theory-internal "exchange rates" between value-bearers. That is, given the independence assumptions P_{23-25} and a knowledge of the theory-internal exchange rates for any two theories, we can deduce what the theory-internal exchange rate of the third theory must be. For instance, given the independence assumptions P_{23-25} plus the facts that $P_6 \in T_5$ and $P_{20} \in T_6$, we can deduce that $P_{21} \in T_7$.

1. $V^{T_7}(\text{aestheton}) = V^{T_5}(\text{aestheton})$ [from P_{24}]
2. $V^{T_5}(\text{aestheton}) = V^{T_5}(\text{hedon})$ [from $P_6 \in T_5$]
3. $V^{T_5}(\text{hedon}) = V^{T_6}(\text{hedon})$ [from P_{23}]
4. $V^{T_6}(\text{hedon}) = V^{T_6}(\text{epistemon})$ [from $P_{20} \in T_6$]
5. $V^{T_6}(\text{epistemon}) = V^{T_7}(\text{epistemon})$ [from P_{25}]
6. $V^{T_7}(\text{aestheton}) = V^{T_7}(\text{epistemon})$ [from 1-5]

C. $P_{21} \in T_7$ [from 6]

This sort of reasoning generalizes to cases in which independence assumptions are replaced by well-defined *dependence* assumptions (e.g., if rather than accepting P_{23}, Carmen believes that the value of a hedon according to T_5 is twice as great as the value of a hedon according to T_6). In either case, the independence/dependence assumptions and the theory-internal exchange rates are constrained by the requirement of mutual consistency.

The sort of reasoning described above also allows us to handle cases in which some pairs of theories share no overlapping value. For instance, consider an agent who divides her belief between three theories: one that values only hedonic goods, another that values only aesthetic goods, and a third that values both hedonic and aesthetic goods and that regards one hedon as equal in value to one aestheton. Given independence assumptions analogous to P_{23-25}, to the effect that the value of a hedon is the same according to the hedonistic theory as according to the pluralistic theory and likewise for the value of an aestheton, we can infer that the value of a hedon according to the hedonistic theory is equal to value of a hedon according to the pluralistic theory, which is equal to the value of an aestheton according to the pluralistic theory, which is equal to the value of an aestheton according to the aesthetic theory, and hence that the value of a hedon according to the hedonistic theory is equal to the value of an aestheton according to the aesthetic theory.[12]

B.4.2 Uncertainty about Value Independence

So far I have assumed that agents who are uncertain what sorts of things have value are nevertheless certain either that the degree of value borne by a particular value bearer is independent of what other value-bearers exist or that it is dependent in some particular way (e.g., that the existence of a second value-bearer would

[12] This reasoning by chains of equalities is what vindicates the idea of pluralistic theories as "bridges" between monistic theories introduced in Chapter 6: The transitivity of equality lets an agent establish intertheoretic comparisons even when she is not certain of any particular value claim that unites all the theories in which she has positive credence (as hedonic value does in the Easy Case) and even when there are some pairs of theories in which she has positive credence that have no value claim in common (like the hedonistic and aesthetic theories in the last example), so long as any two theories are linked by a *chain* of theory-pairs each of which has some value claim in common that allows for normalization.

reduce the value per unit of the first value-bearer by 50%). But what if an agent is uncertain about the dependence relationship between value-bearers?

Consider Daniel, who divides his belief between classical utilitarianism and a simple version of prioritarianism, according to which the interests of those above some threshold of absolute wellbeing count for only half as much as the interests of those below that threshold. Daniel wants to know how he should weigh these theories against one another when they conflict. To do this, he must consider the following question: Does prioritarianism value the interests of the less well off *more* than classical utilitarianism, or the interests of the better off *less*, or both? That is, what is the relationship between the value of, say, an increment of wellbeing to an already-well-off person according to classical utilitarianism and the value of that same increment of wellbeing according to prioritarianism? If Daniel were certain, say, that if prioritarianism is true, then the interests of the less well off count for twice as much as they would under classical utilitarianism, while the interests of the better off have exactly the same degree of moral weight whether utilitarianism or prioritarianism is correct, then he would have no difficulty in normalizing utilitarianism and prioritarianism. But what if he is uncertain?

To make Daniel's case precise, let's suppose that his moral beliefs center on the following propositions.

P_1 The choiceworthiness of an option is equal to the value of the world that results from it, where the value of a world is the cardinal sum of non-derivative value that supervenes on value-bearing empirical phenomena in that world.

P_2 Hedonic experiences (i.e., experiences with positive or negative hedonic valence) are non-derivative value-bearers.

P_{26} Nothing apart from hedonic experience is a non-derivative value-bearer.

P_{27} An additional unit of hedonic experience (i.e. a hedon) has twice as much value if the subject of that experience has a level of hedonic wellbeing below threshold t than it does if the subject of that experience has a level of wellbeing above threshold t.

P_{28} The value of an additional hedon for a subject below t is independent of whether P_{27} is true.

P_{29} The value of an additional hedon for a subject below t is twice as great if P_{27} is true than it is if P_{27} is false.

P_{30} The value of an additional hedon for a subject above t is independent of whether P_{27} is true.

P_{31} The value of an additional hedon for a subject above t is only half as great if P_{27} is true as it is if P_{27} is false.

Daniel's uncertainty about the dependence relations between utilitarianism and prioritarianism, then, amounts to an uncertainty whether $P_{28} \wedge P_{31}$ is true or $P_{29} \wedge P_{30}$ is true. Because of this uncertainty, Daniel divides his beliefs between *three* theories rather than two.

T_8 (utilitarianism): $\{P_1, P_2, P_{26}, \neg P_{27}...\}$

T_9 ("low prioritarianism"): $\{P_1, P_2, P_{26}, P_{27}, P_{28}, \neg P_{29}, \neg P_{30}, P_{31}, ...\}$

T_{10} ("high prioritarianism"): $\{P_1, P_2, P_{26}, P_{27}, \neg P_{28}, P_{29}, P_{30}, \neg P_{31}, ...\}$[13]

Given this description of the case, it is clear how Daniel ought to normalize his moral theories for purposes of moral hedging: The value of a hedon for someone below threshold t according to utilitarianism is equal to the value of a hedon for someone below threshold t according to low prioritarianism, but only half as great as the value of a hedon for someone below threshold t according to high prioritarianism. The value of a hedon for someone above threshold t according to utilitarianism is equal to the value of a hedon for someone above threshold t according to high prioritarianism, and twice as great as the value of a hedon for someone above threshold t according to low prioritarianism. This means, of course, that the value of a hedon for *anyone* (above or below threshold t) is twice as great according to high prioritarianism as it is according to low prioritarianism.

The more of his prioritarian credence Daniel assigns to high prioritarianism rather than low prioritarianism, therefore, the greater the total influence of his prioritarian credence on his decisions under moral uncertainty will be. And this is as it should be: High prioritarianism says that there is *more* at stake than utilitarianism recognizes, i.e., the interests of the less well off are *more urgent*. Low prioritarianism

[13] In fact, Daniel will presumably divide his beliefs between *four* theories, since one version of classical utilitarianism will include $P_{28} \wedge P_{31}$ while another will include $P_{29} \wedge P_{30}$. This complication is relevant if we suppose that Daniel is uncertain between a "high value" and a "low value" version of classical utilitarianism just as he is uncertain between a "high value" and a "low value" version of prioritarianism. For simplicity, I have assumed that Daniel is certain of the value of a hedon given classical utilitarianism (i.e., given $\neg P_{27}$), in which case the difference between the versions of utilitarianism that include $P_{28} \wedge P_{31}$ and $P_{29} \wedge P_{30}$ are interchangeable for all practical purposes. As far as I can see, uncertainty between high value and low value versions of utilitarianism would not raise any interestingly new difficulties.

says that there is *less* at stake, i.e., that the interests of the already-well-off just don't matter that much, contra utilitarianism. High prioritarianism deserves more weigh in a morally uncertain agent's moral calculus than does low prioritarianism.

Of course, like the other cases we have examined in this appendix, Daniel's case is artificially simplified. In a more realistic case, Daniel might distribute his prioritarian belief over a wide—perhaps infinite—range of prioritarian weighting functions (more sophisticated alternatives to the "hedons count for twice as much above threshold t" function used above), and for each of those functions, might have credence in many "higher" or "lower" theories that normalize that particular weighting function with classical utilitarianism differently. Any attempt at a realistic representation of such a belief state would yield a number of propositions and theories much too large to be explicitly enumerated. Nonetheless, the lesson from Daniel's case should generalize to more complex cases: Once we make explicit the agent's distribution of belief over the various possible states of dependence or independence that might be exhibited by a value-bearer (hedonic value, aesthetic value, hedonic-value-for-subjects-below-threshold-t) relative to some set of theories that assign it non-derivative value, the result will be a more fine-grained collection of theories, each of which is connected to the others by some definite relations of dependence or independence that allow for intertheoretic normalization and aggregation.

B.4.3 "Infectious Incomparability"

I have so far assumed that, even if an agent divides her beliefs among propositions like P_{14-16} that posit different comparisons between the same pair of theories, she does not assign positive credence to a proposition like P_{17} that posits *in*comparability. But this is of course unrealistic—however strong the case for intertheoretic comparability, Alice should have at least *some* positive credence in the view that T_1 and T_2 are incomparable. MacAskill (2013) has pointed out that, given an expectational decision rule, incomparability is "infectious": An agent who has any non-zero credence that incomparable values are at stake in the choice before her will find that the expected value of her options is undefined. Does this not mean that, so long as Alice has non-zero credence in P_{17}, this incomparabilist view will "swamp" whatever credence she has in comparison propositions like P_{14-16}?

If so, then this would be bad news not just for proponents of moral hedging, but for proponents of expectational reasoning more generally. For one thing, if the true moral theory turns out to involve incomparable values (say, special obligations to loved ones versus impersonal moral obligations, or a parent's special obligations to one of her children versus another), then any non-zero credence that a given choice involves competition between those incomparable values will result in expectations being undefined. But perhaps more importantly, even if the true ethical theory involves no incomparability, the same problem presents itself from a different direction for agents attempting to maximize expected value under empirical uncertainty. For MacAskill's "infectious incomparability" problem has a close analogue in the problem of “infectious infinities”: An agent deciding whether to perform a practical option O who has some positive credence that O will result in infinite positive value (say, by causing another agent to find the true religion and enjoy eternal salvation) and some positive credence that O will result in infinite negative value (say, by causing another agent to find a false religion and suffer eternal damnation) will discover that the expected value of O is undefined (since $\infty - \infty$ is undefined). If an epistemically rational agent should have positive credence in everything or nearly everything, then it appears as though straightforward expectational reasoning is simply impossible (or rather, entirely possible but disappointingly unhelpful in its results). It seems to me, therefore, that the problem of small credences generating undefined expectations is a general problem for decision theory, rather than a challenge to hedging under moral uncertainty specifically—that is, we have good reason to require that the true decision-theoretic account of choice under uncertainty not be *simply* expectational, but include some device for dealing with small probabilities of extreme or undefined value.[14]

B.4.4 Incomparability Simpliciter

My focus so far in this appendix has been on cases in which any two theories, even if they do not share any directly overlapping content (e.g. both attributing non-derivative value to hedonic experience), are nevertheless connected by a chain

[14] Possible decision-theoretic approaches to this problem include Nicholas Smith’s theory of “rationally negligible probabilities” (Smith, 2014, 2016), on which agents are simply permitted to ignore possibilities that fall below a certain probability threshold, and Lara Buchak’s “risk-weighted expected utility maximization” (Buchak, 2013), on which agents are rationally permitted to adopt any consistent set of risk attitudes, including attitudes that are less sensitive to extremely low probabilities of extremely large gains or losses than risk-neutral expected utility maximization.

of theories in which each directly linked pair *do* share some such overlapping content that allows for normalization. Thus, the approach I have described has limited scope: It only applies to agents whose moral beliefs satisfy this condition. This leaves open the possibility that, when two theories in which an agent has positive credence are not connected by any chain of overlapping, normalizable theories, then they are simply incomparable.

But there are two other ways in which incomparability can arise, even among theories appropriately connected by common content: namely, when an agent judges either that two value bearers are incomparable *simplicter* (that is, incomparable even on the hypothesis that they are both genuine, non-derivative value bearers) or when she judges that the dependency relation for a shared form of value between two theories is one of incomparability. Let's consider one final case to illustrate these points.

In this case, Ella divides her beliefs mainly between classical utilitarianism, according to which she ought to maximize global hedonic utility, and Nietzschean perfectionism, according to which she ought to maximize a certain kind of self-perfection. Considering the positive arguments for both of these views plausible, she also has some slight credence that *both* global hedonic utility and Nietzschean perfection have non-derivative moral value. Ella thus has beliefs about the following propositions, *inter alia*.

P_1 The choiceworthiness of an option is equal to the [possibly agent-relative] value of the world that results from it, where the value of a world is the cardinal sum of non-derivative value that supervenes on value-bearing empirical phenomena in that world.

P_2 Hedonic experiences (i.e., experiences with positive or negative hedonic valence) are non-derivative value-bearers.

P_{32} Traits that contribute to an agent's own self-perfection are non-derivative (agent-relative) value-bearers.

P_{33} Nothing besides hedonic experience and an agent's own self-perfection is a non-derivative value-bearer.

Ella then divides her belief between three theories:

T_{11} (utilitarianism): $\{P_1, P_2, \neg P_{32}, P_{33}...\}$

T_{12} (perfectionism): $\{P_1, P_2, P_{32}, \neg P_{33}...\}$

T_{13} (pluralism): $\{P_1, P_2, P_{32}, P_{33}...\}$

If, say, Ella's pluralistic theory T_{13} includes the proposition that the value of one hedon is equal to the value of one perfecton, and if she is certain that the value of a hedon (given that hedons have non-derivative value) is independent of whether perfectons have value and vice versa, then the case is simple and her three theories can be easily normalized.

But suppose instead that T_{13} includes the following proposition.

P_{34} The value of hedonic experience and the value of self-perfection are incomparable.

If $P_{34} \in T_{13}$, then the approach I have described offers no basis for normalizing T_{11-13}.

Alternatively, suppose that Ella believes the following with certainty.

P_{35} The value of a unit of self-perfection given $P_{32} \wedge P_2$ is incomparable with the value of a unit of self-perfection given $P_{32} \wedge \neg P_2$.

P_{35} asserts that the value of a perfecton according to T_{12} is incomparable with the value of a perfecton according to T_{13}. Unlike an assumption of independence (or any definite dependency, e.g. that perfectons have only half as much value according to T_{13} as they do according to T_{12}), this prevents any intuitive, content-based normalization between T_{12} and T_{13} (and by extension between T_{12} and T_{11}). So if Ella accepts P_{35} (i.e., if $P_{35} \in T_{11}, T_{12}, T_{13}$), then the approach to intertheoretic normalization I have been describing will be of no help to her. (The same is true, of course, if she accepts an analogous proposition to the effect that the value of a hedon according to T_{11} is incomparable with the value of a hedon according to T_{13}.)

In the last section, I claimed that an adequate resolution of the various decision-theoretic puzzles involving small probabilities of incomparable values, infinite values, etc, should let us avoid the problem of "infectious incomparability" for agents who have very low credence in intertheoretic incomparability. But this is only true if Ella's unconditional credence in P_{35} and her conditional credence in $(P_{34}|P_2 \wedge P_{32})$ are very close to zero. If these credences are substantial, even if they are less than 1, then presumably the decision-theoretic resolution of problems involving minuscule probabilities will not allow Ella to ignore them.

How much this restricts the scope of the account I have given depends, therefore, on how plausible one finds propositions like P_{34} and P_{35} as opposed to the competing assumptions of value comparability within pluralistic theories (contra P_{34}) and independence (or determinate dependence) of the value of one kind of value-bearer on the existence of other value-bearers (contra P_{35}). For my part, incomparabilist theories like the version of T_{13} that includes P_{34} seem fairly implausible, though they may have greater plausibility in the context of particular pluralistic theories, like T_{13}, that combine particularly irreconcilable sorts of value. And it is unclear what would ever motivate an agent to assign substantial probability to a proposition like P_{35}. (In the next section, I will attempt to further strengthen the case that agents ought to default to independence assumptions like P_{14} absent specific grounds for rejecting these assumptions, which will give us additional reason to be skeptical of propositions like P_{35}.) Nevertheless, the account I have given leaves room for intertheoretic incomparability even in cases where the moral theories in which an agent has positive credence are linked together by chains of overlapping content.

B.5 What Grounds Credence in Comparisons?

I have attempted to show that the rationality of making certain intertheoretic value comparisons, and of hedging for one's moral uncertainties on the basis of those comparisons, can be grounded in an agent's doxastic state. But it might appear that this only pushes the problem of intertheoretic value comparisons back a step. An interlocutor might argue: "Granted, it may be rational for an agent who *believes* certain intertheoretical comparison propositions—for instance, Alice who believes that her T_1 and T_2 each assign the same value to a given unit of hedonic experience—to act on the basis of those comparisons. But the deeper question is, what could ever make such a comparison *true*? And if we lack any satisfactory answer to this question, does it not undercut the claim that agents like Alice can be *rationally justified* in believing such propositions?" In other words, even if we have explained what grounds *the rationality of acting as if certain intertheoretic comparisons are correct*, for agents who accept those comparisons, we have not yet answered the question of *what grounds the intertheoretic comparisons themselves.* In this section I will attempt to respond to this objection—not by giving a definite answer to the question of what grounds or makes-true intertheoretic comparisons like P_{14}, but rather by drawing on and extending arguments from the preceding sections to argue that there *must be* an answer of one kind or another, and suggesting a few possibilities.

The argument is simply this: (1) In the context of various forms of non-normative uncertainty, it is intuitively clear that propositions analogous to P_{14} are not only reasonable to believe but also true (at least, insofar as any evaluative propositions are ever true). And (2) there is no reason to expect that whatever story we tell about the truth of these propositions will not extend straightforwardly to propositions like P_{14} in the context of normative uncertainty.

To motivate (1), consider again the case of Bob, who is certain of the truth of hedonistic utilitarianism but uncertain what will happen if he presses the red button (and thus divides his belief between two theories of the world, T_3 and T_4). As we noted, for Bob to calculate the expected value of pressing the red button, he must assume *something* about how the value of a hedon according to T_3 compares with the value of a hedon according to T_4—and presumably, he should assume that the value of a hedon *is the same* according to both theories, i.e., that the value of a hedon is *independent of what would happen if he presses the red button.* Furthermore, it seems clear that this belief is *true.*[15]

What grounds or makes true the proposition that the value of a hedon is independent of the consequences of pressing the red button? Various stories are possible. Perhaps it is a fact about possible worlds, i.e., that hedons are equally valuable in the possible worlds where pressing the red button will save five and the possible worlds where it will kill ten. Perhaps it is simply the fact of *logical* independence, i.e., that propositions about the consequences of pressing the red button have no logical implications for the value of hedons. Perhaps it is some weaker abstract truth to the effect that logically independent features of the world are *prima facie* independent, i.e., should be treated as independent in the absence of any facts particular to the case that would ground a dependency between them.[16]

The latter two hypotheses extend straightforwardly to the case of Alice and

[15] If hedons simply don't have value, then it will be trivially true. It could only fail to be true, as far as I can see, if concepts like *value* are so hopelessly confused that no proposition about value can *ever* be true (a là the version of moral nihilism considered in Ross (2006) and MacAskill (2014)).

[16] This would for instance accommodate the idea that, if there are more sources of value in the world, each source of value counts for less—even though the proposition that aesthetic goods are non-derivative value bearers does not (at least on its own) *logically* entail that hedonic experience is less valuable than it otherwise would be, there might be non-logical entailments that defeat the presumption of independence.

other similar cases of moral uncertainty: The value of hedons is logically independent of whether aesthetons have value, and this fact either on its own or coupled with the absence of defeaters to the presumption of independence might ground the truth of P_{14}.[17] But the first hypothesis, that of equality across possible worlds, points out a worry about the analogy between moral and empirical uncertainty: namely, that basic moral propositions are (plausibly) necessary rather than contingent truths, which presents an apparent obstacle to spelling out independence in terms of possible worlds.

But this should not worry us too much, for there are cases of *non*-normative uncertainty in which the objects of uncertainty are necessarily true/false propositions that nevertheless seem clearly analogous to the case of Bob. Consider Frederick, a precocious six-year-old utilitarian who finds himself faced with a trolley problem: On the main track, he knows, are exactly 65 people, all of whom will be killed if the trolley is allowed to continue. Frederick can divert the trolley onto a siding, but if he does so it will collide with a train car loaded with highly volatile explosives, causing them to detonate. On a field next to the siding, Frederick can see a marching band practicing, in a rectangular formation that (Frederick can quickly determine by counting) is nine people wide and eight people deep. Frederick must make a decision immediately, and has no time to count all the members of the marching

[17] Adding to the plausibility of these hypotheses, it is worth noting that the proponent of incomparability must explain what grounds the fact that the values of rival moral theories like T_1 and T_2 are *in*comparable, and in doing so will presumably appeal to some similar general principle, e.g. a *prima facie* principle that the values of rival theories are incomparable absent some special fact that grounds their comparability. But why should incomparability be the default state of affairs rather than independence (especially when independence seems to be the default in cases of non-normative uncertainty)?

band. But fortunately, he has overheard his older sister practicing her multiplication tables, and remembers that 9×8 is *either* 63 or 72. He can't remember which, and has no time to improve his epistemic position on this question. As the trolley speeds toward the junction, either answer seems equally plausible to him. Precocious child that he is, it occurs to him that he should just split the difference between 63 and 72. While he can't calculate it exactly, he realizes that whatever number is midway between 63 and 72 will be greater than 65.

It seems clear to me that, in his state of mathematical uncertainty, Frederick should not turn the trolley—even though doing so would in fact save more lives (since $9 \times 8 = 72$). But, just as in the case of Bob, expectational reasoning in the face of Frederick's uncertainty whether $9 \times 8 = 72$ or 63 requires an implicit assumption that the importance of the relevant values (hedons, lives saved, etc) is independent of the mathematical facts, i.e., that Frederick should not value saving a life any more or less on the hypothesis that $9 \times 8 = 72$ than on the hypothesis that $9 \times 8 = 63$.[18]

[18] If you think that there is some minimal threshold for rational agency that requires closure of belief under the relatively simple deductive inferences required to compute 9×8, then of course you can simply imagine an analogous case involving an adult reasoner who is relevantly uncertain about some more difficult mathematical proposition. For instance, suppose that Georgiana, a utilitarian and an accomplished logician whose beliefs meet any closure requirement that could be reasonably imposed on human agents, must decide whether to kill one trillion people in order to save a number of people equal to the billionth prime number. (Granted, summoning up intuitions about this case may require suspending a deal of disbelief about certain implications of utilitarianism.)

There is nothing untoward, as far as I can see, in appealing to the mathematical reasoning required to compute expectations under conditions of mathematical uncertainty. Granted, someone who accepts the arithmetical principles required to compute expectations but holds or entertains false arithmetical beliefs (e.g., that $9 \times 8 = 63$) is in some sense not fully coherent, having at least some credence in a contradiction. But to hold that ordinary principles of practical rationality do not apply to an agent who has non-zero credence in a contradiction is to catastrophically restrict the scope of those principles: For any finite reasoner, there will be some logical and mathematical truths too difficult to prove in a fixed period of time. If such a reasoner is forced to make a choice, within such a fixed period of time, that turns on the truth of such a proposition, she must form some belief about it, and presumably that belief ought to be of a strength less than certainty.

If we are tempted to ground the independence assumption in facts about possible worlds in the case of Bob, then, what should we say about the case of Frederick? Perhaps we can simply stay the course, appealing not to metaphysically or even logically possible worlds (since there are no metaphysically or logically possible worlds in which $9 \times 8 = 63$) but to *epistemically* possible worlds.[19] Or perhaps the possibility of expectational reasoning in the face of mathematical uncertainty should lead us to accept an alternative story about independence assumptions, in Bob's case as well as Alice's and Frederick's, that makes no appeal to possible worlds.

In any case, the question of what precisely grounds intertheoretic value comparisons—what makes it true in the Easy Case that the value of a hedon is independent of whether aesthetons have value (or that it is dependent in some particular way)—is one best left to the metaphysicians. But it seems clear that a morally uncertain agent *can reasonably accept* propositions like P_{14}, and therefore that the rationality of hedging and of *making* the intertheoretic value comparisons that hedging requires can be grounded in an agent's reasonable beliefs.

[19] Of course, it seems that *merely* epistemic "possible worlds," like that in which $9 \times 8 = 63$, must be in some sense ersatz. There are no Lewisian possible worlds, as real as our own, where $9 \times 8 = 63$. But since few believe that there are *any* fully non-ersatz, Lewisian possible worlds, this should not be seen as a great cost to most of those who are tempted to ground the truth of Bob's independence assumption in modal facts.

Bibliography

Aboodi, R., A. Borer, and D. Enoch (2008). Deontology, Individualism, and Uncertainty: A Reply to Jackson and Smith. *Journal of Philosophy 105*(5), 259–272.

Alpert, M. and H. Raiffa (1982). A Progress Report on the Training of Probability Assessors. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, pp. 294–305. Cambridge University Press.

Arntzenius, F. (2014). Utilitarianism, Decision Theory and Eternity. *Philosophical Perspectives 28*(1), 31–58.

Arntzenius, F., A. Elga, and J. Hawthorne (2004). Bayesianism, Infinite Decisions, and Binding. *Mind 113*(450), 251–283.

Arpaly, N. and T. Schroeder (1999). Praise, Blame and the Whole Self. *Philosophical Studies 93*(2), 161–188.

Bennett, J. (1974). The Conscience of Huckleberry Finn. *Philosophy 49*(188), 123–134.

Bennett, K. (2011). By Our Bootstraps. *Philosophical Perspectives 25*(1), 27–41.

Bernoulli, D. (1954 (1738)). Exposition of a New Theory on the Measurement of Risk. *Econometrica: Journal of the Econometric Society 22*(1), 23–36.

Bhikkhu, T. (2013, November). Sañña Sutta: Perceptions (AN 7.46), translated from the Pali by Thanissaro Bhikkhu.

Bostrom, N. (2009, January). Moral Uncertainty — Towards a Solution? http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html.

Bostrom, N. (2011). Infinite Ethics. *Analysis and Metaphysics 10*, 9–59.

Broome, J. (1997). Is Incommensurability Vagueness? In *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press.

Broome, J. (1999). Normative Requirements. *Ratio 12*(4), 398–419.

Broome, J. (2001). Normative Practical Reasoning. *Aristotelian Society Supplementary Volume 75*(1), 175–193.

Buchak, L. (2013). *Risk and Rationality*. Oxford University Press.

Bykvist, K. (2013). Evaluative Uncertainty, Environmental Ethics, and Consequentialism. In R. I. Hiller, Avram and L. Kahn (Eds.), *Consequentialism and Environmental Ethics*. Routledge.

Camerer, C., G. Loewenstein, and M. Weber (1989). The Curse of Knowledge in Economic Settings: An Experimental Analysis. *Journal of Political Economy 97*(5), 1232–1254.

Chang, R. (Ed.) (1997a). *Incommensurability, Incomparability and Practical Reason*. Harvard University Press.

Chang, R. (1997b). Introduction. In *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press.

Colyvan, M., D. Cox, and K. Steele (2010). Modelling the Moral Dimension of Decisions. *Noûs 44*(3), 503–529.

Crouch, W. (2010). Moral Uncertainty and Intertheoretic Comparisons of Value. Bphil thesis, University of Oxford.

Daniels, N. (1979). Wide Reflective Equilibrium and Theory Acceptance in Ethics. *Journal of Philosophy 76*(5), 256–282.

de Lazari-Radek, K. and P. Singer (2014). *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford University Press.

Diener, E., E. M. Suh, R. E. Lucas, and H. L. Smith (1999). Subjective Well-Being: Three Decades of Progress. *Psychological Bulletin 125*(2), 276–302.

Dreier, J. (2000). Dispositions and Fetishes: Externalist Models of Moral Motivation. *Philosophical and Phenomenological Research 61*(3), 619–638.

Elga, A. (2010). How to Disagree About How to Disagree. In T. Warfield and R. Feldman (Eds.), *Disagreement*. Oxford University Press.

Foot, P. (1972). Morality as a System of Hypothetical Imperatives. *The Philosophical Review 81*(3), 305–316.

Gauthier, D. P. (1986). *Morals by Agreement*. Oxford University Press.

Gert, J. (2003). Requiring and Justifying: Two Dimensions of Normative Strength. *Erkenntnis 59*(1), 5–36.

GiveWell (2013, November). Mass Distribution of Long-Lasting Insecticide-Treated Nets (LLINs). http://www.givewell.org/international/technical/programs/insecticide-treated-nets.

Gracely, E. J. (1996). On the Noncomparability of Judgments Made by Different Ethical Theories. *Metaphilosophy 27*(3), 327–332.

Greaves, H. and T. Ord (forthcoming). Moral Uncertainty about Population Ethics. *Journal of Ethics and Social Philosophy*.

Greene, J. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3*. MIT Press.

Greenwell, J. R. (1977). Abortion and Moral Safety. *Crítica: Revista Hispanoamericana de Filosofía 9*(27), 35–48.

Guerrero, A. A. (2007). Don't Know, Don't Kill: Moral Ignorance, Culpability, and Caution. *Philosophical Studies 136*(1), 59–97.

Gustafsson, J. E. and O. Torpman (2014). In Defence of My Favourite Theory. *Pacific Philosophical Quarterly 95*(2), 159–174.

Hadar, J. and W. R. Russell (1969). Rules for Ordering Uncertain Prospects. *The American Economic Review 59*(1), 25–34.

Hájek, A. (2003). Waging war on pascal's wager. *Philosophical Review 112*(1), 27–56.

Hájek, A. (2012). Interpretations of Probability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012 ed.).

Harman, E. (2015). The Irrelevance of Moral Uncertainty. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Volume 10. Oxford University Press.

Harman, G. (1975). Moral Relativism Defended. *Philosophical Review 84*(1), 3–22.

Hawley, P. (2008). Moral Absolutism Defended. *Journal of Philosophy 105*(5), 273–275.

Hedden, B. (2016). Does MITE Make Right? On Decision-Making under Normative Uncertainty. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Volume 11. Oxford University Press.

Hobbes, T. (2004 (1651)). *Leviathan*. Clarendon Press.

Horgan, T. and M. Timmons (2010). Untying a Knot From the Inside Out: Reflections on the "Paradox"of Supererogation. *Social Philosophy and Policy 27*(2), 29–63.

Horty, J. F. (2014). *Reasons as Defaults*. Oxford University Press Usa.

Hudson, J. L. (1989). Subjectivization in Ethics. *American Philosophical Quarterly 26*(3), 221–229.

Huemer, M. (2010). Lexical Priority and the Problem of Risk. *Pacific Philosophical Quarterly 91*(3), 332–351.

Isaacs, Y. (2014). Duty and Knowledge. *Philosophical Perspectives 28*(1), 95–110.

Jackson, F. (1991). Decision-Theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics 101*(3), 461–482.

Jackson, F. and R. Pargetter (1986). Oughts, Options, and Actualism. *Philosophical Review 95*(2), 233–255.

Jackson, F. and M. Smith (2006). Absolutist Moral Theories and Uncertainty. *Journal of Philosophy 103*(6), 267–283.

Jackson, F. and M. Smith (2016). The Implementation Problem for Deontology. In E. Lord and B. Maguire (Eds.), *Weighing Reasons*, pp. 279– 291. Oxford University Press.

Joyce, J. M. (2002). Levi on Causal Decision Theory and the Possibility of Predicting One's Own Actions. *Philosophical Studies 110*(1), 69–102.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

Kahneman, D. and A. Deaton (2010). High Income Improves Evaluation of Life but not Emotional Well-Being. *Proceedings of the National Academy of Sciences 107*(38), 16489–16493.

Kahneman, D., B. L. Fredrickson, C. A. Schreiber, and D. A. Redelmeier (1993). When More Pain is Preferred to Less: Adding a Better End. *Psychological science 4*(6), 401–405.

Kalai, E. and M. Smorodinsky (1975). Other Solutions to Nash's Bargaining Problem. *Econometrica: Journal of the Econometric Society 43*(3), 513–518.

Kant, I. (1960 (1793)). *Religion within the Limits of Reason Alone.* Harper and Row. tr. T. M. Greene and H. H. Hudson.

Kavka, G. S. (1983). The Toxin Puzzle. *Analysis 43*(1), 33–36.

Kerstein, S. J. (2002). *Kant's Search for the Supreme Principle of Morality.* Cambridge University Press.

Keynes, J. M. (1921). *A Treatise on Probability.* London: MacMillan and Co.

Kolodny, N. (2005). Why Be Rational? *Mind 114*(455), 509–563.

Korsgaard, C. M. (1996). *The Sources of Normativity.* Cambridge University Press.

Landesman, C. (1995). When to Terminate a Charitable Trust? *Analysis 55*(1), 12–13.

Lockhart, T. (1977). Another Moral Standard. *Mind 86*(344), 582–586.

Lockhart, T. (2000). *Moral Uncertainty and Its Consequences.* Oxford University Press.

MacAskill, W. (2013). The Infectiousness of Nihilism. *Ethics 123*(3), 508–520.

MacAskill, W. (2014). *Normative Uncertainty.* Ph. D. thesis, University of Oxford.

Marquis, D. (1989). Why Abortion is Immoral. *The Journal of Philosophy 86*(4), 183–202.

Moller, D. (2011). Abortion and Moral Risk. *Philosophy 86*(03), 425–443.

Nissan-Rozen, I. (2015). Against Moral Hedging. *Economics and Philosophy 31*(3), 1–21.

Norcross, A. (2006). The Scalar Approach to Utilitarianism. In H. West (Ed.), *The Blackwell Guide to Mill's Utilitarianism*, pp. 217–32. Wiley-Blackwell.

Nover, H. and A. Hájek (2004). Vexing Expectations. *Mind 113*(450), 237–249.

Oddie, G. (1994). Moral Uncertainty and Human Embryo Experimentation. *Medicine and Moral Reasoning 3*, 144.

Oddie, G. and P. Menzies (1992). An Objectivist's Guide to Subjective Value. *Ethics 102*(3), 512–533.

Parfit, D. (1971). Personal Identity. *Philosophical Review 80*(January), 3–27.

Parfit, D. (1973). Later Selves and Moral Principles. In A. Montefiore (Ed.), *Philosophy and Personal Relations.* Routledge and Kegan Paul.

Parfit, D. (1984). *Reasons and Persons.* Oxford University Press.

Parfit, D. (2011). *On What Matters.* Oxford University Press.

Pascal, B. (1852/1669). *Pensées.* Dezobry et E. Magdeleine.

Pascal, B. (1997 (1657)). *The Provincial Letters.* Wipf and Stock Publishers.

Pfeiffer, R. S. (1985). Abortion Policy and the Argument from Uncertainty. *Social Theory and Practice 11*(3), 371–386.

Portmore, D. (2008). Are Moral Reasons Morally Overriding? *Ethical Theory and Moral Practice 11*(4), 369–388.

Portmore, D. (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality.* Oxford University Press USA.

Portmore, D. (2016). Uncertainty, Indeterminacy, and Agent-Centered Constraints. *Australasian Journal of Philosophy*, 1–15.

Quinn, W. (1993). Putting Rationality in Its Place. In *Morality and Action*, pp. 228 – 255. Cambridge University Press Cambridge.

Railton, P. (1986). Moral Realism. *Philosophical Review 95*(2), 163–207.

Rawls, J. (1951). Outline of a Decision Procedure for Ethics. *Philosophical Review 60*(2), 177–197.

Raz, J. (1975). Permissions and Supererogation. *American Philosophical Quarterly 12*(2), 161–168.

Reisner, A. (2013). Is the Enkratic Principle a Requirement of Rationality? *Organon F 20*(4), 436–462.

Roese, N. J. and K. D. Vohs (2012). Hindsight Bias. *Perspectives on Psychological Science 7*(5), 411–426.

Rosen, G. (2002). Culpability and Ignorance. *Proceedings of the Aristotelian Society 103*(1), 61–84.

Rosen, G. (2004). Skepticism About Moral Responsibility. *Philosophical Perspectives 18*(1), 295–313.

Ross, J. (2006). Rejecting Ethical Deflationism. *Ethics 116*(4), 742–768.

Sepielli, A. (2006). Review: Moral Uncertainty and Its Consequences. *Ethics 116*(3), 601–604.

Sepielli, A. (2009). What to Do When You Don't Know What to Do. *Oxford Studies in Metaethics 4*, 5–28.

Sepielli, A. (2010). *'Along an Imperfectly-Lighted Path': Practical Rationality and Normative Uncertainty.* Ph. D. thesis, Rutgers University Graduate School - New Brunswick.

Sepielli, A. (2012). Normative Uncertainty for Non-Cognitivists. *Philosophical Studies 160*(2), 191–207.

Sepielli, A. (2013). Moral Uncertainty and the Principle of Equity among Moral Theories. *Philosophy and Phenomenological Research 86*(3), 580–589.

Sepielli, A. (2014a). Should You Look Before You Leap? *The Philosophers' Magazine 66*, 89–93.

Sepielli, A. (2014b). What to Do When You Don't Know What to Do When You Don't Know What to Do.... *Noûs 48*(3), 521–544.

Sepielli, A. (2016). Moral Uncertainty and Fetishistic Motivation. *Philosophical Studies 173*(11), 2951–2968.

Sidgwick, H. (1962 (1874)). *The Methods of Ethics* (7th ed.). Palgrave Macmillan.

Singer, P. (1972). Famine, Affluence, and Morality. *Philosophy & Public Affairs 1*(3), 229–243.

Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics 9*(3-4), 331–352.

Smith, M. (1994). *The Moral Problem.* Blackwell.

Smith, M. (2002). Evaluation, Uncertainty and Motivation. *Ethical Theory and Moral Practice 5*(3), 305–320.

Smith, N. J. J. (2014). Is Evaluative Compositionality a Requirement of Rationality? *Mind 123*(490), 457–502.

Smith, N. J. J. (2016). Infinite Decisions and Rationally Negligible Probabilities. *Mind* (500), 1–14.

Soll, J. B. and J. Klayman (2004). Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition 30*(2), 299.

Talbott, W. (2015). Bayesian Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2015 ed.).

Tarsney, C. Normative Uncertainty and Social Choice. unpublished.

Tarsney, C. (forthcoming a). Intertheoretic Value Comparison: A Modest Proposal. *Journal of Moral Philosophy*.

Tarsney, C. (forthcoming b). Moral Uncertainty for Deontologists. *Ethical Theory and Moral Practice*.

Taylor, C. (1999). *The Atomists: Leucippus and Democritus. Fragments, A Text and Translation with Commentary.* Toronto: University of Toronto Press.

Tenenbaum, S. (2017). Action, Deontology, and Risk: Against the Multiplicative Model. *Ethics 123*(3), 674–707.

Thomson, J. J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist 59*(2), 204–217.

Twain, M. (1884). *Adventures of Huckleberry Finn.* Chatto & Windus.

Urmson, J. O. (1958). Saints and Heroes. In A. I. Melden (Ed.), *Essays in Moral Philosophy.* University of Washington Press.

Vallentyne, P. and S. Kagan (1997). Infinite Value and Finitely Additive Value Theory. *The Journal of Philosophy 94*(1), 5–26.

Van Fraassen, B. C. (1989). *Laws and Symmetry.* Oxford University Press.

von Neumann, J. and O. Morgenstern (1947). *Theory of Games and Economic Behavior* (2nd ed.). Princeton University Press.

Weatherson, B. (2007). Disagreeing about disagreement.

Weatherson, B. (2013). Disagreements, Philosophical and Otherwise. In J. Lackey and D. Christensen (Eds.), *The Epistemology of Disagreement: New Essays*, pp. 54. Oxford University Press.

Weatherson, B. (2014). Running Risks Morally. *Philosophical Studies 167*(1), 141–163.

Wedgwood, R. (2013). Akrasia and Uncertainty. *Organon F 20*(4), 484–506.

Williams, B. (1979). Internal and External Reasons. In R. Harrison (Ed.), *Rational Action*, pp. 101–113. Cambridge University Press.

Williams, B. (1981). Persons, Character, and Morality. In James Rachels (Ed.), *Moral Luck*. Cambridge University Press.

Wise, J. (2013, July). Giving Now vs. Later: A Summary.

Wittgenstein, L. (1965). A Lecture on Ethics. *The Philosophical Review 74*(1), 3–12.

Wolf, S. (1982). Moral Saints. *The Journal of Philosophy 79*(8), 419–439.

Zimmerman, M. J. (1997). Moral Responsibility and Ignorance. *Ethics 107*(3), 410–426.

Zimmerman, M. J. (2008). *Living with Uncertainty: The Moral Significance of Ignorance*. Cambridge University Press.