

NORMATIVE UNCERTAINTY & INFORMATION VALUE



Riley Harris

*A thesis submitted in fulfillment of the requirements
for the degree of Master of Philosophy*

*Department of Philosophy
School of Humanities
Faculty of Arts
The University of Adelaide*

October, 2021

Riley Harris: *Normative Uncertainty and Information Value*, © October 2021

SUPERVISORS:

Dr. Antony Eagle
Dr. Garrett Cullity

LOCATION:

Adelaide



TIME FRAME:

February 2019-October 2021

A NOTE ON THE TYPE:

This thesis was typeset in L^AT_EX using `classicthesis`, which was developed by André Miede and Ivo Pletikosić. I used a version adapted by my colleague Danny Wardle. Palatino acts as both the text and display typeface. Monospaced text is typeset in Bitstream Vera Mono.

A NOTE ON NAVIGATION:

In the electronic version of this thesis, internal hyperlinks have been inserted to facilitate navigation. Internal links appear in **this purple colour**; clicking on the purple text will navigate to the relevant part of the document. For example, clicking on a citation will navigate to the relevant entry in the list of references. In all cases issuing a *back* command will navigate back to the point of origin. In most .pdf readers this can be achieved by the key combination  + .

For everyone precious to me; and everyone who could have been.

ABSTRACT

This thesis is about making decisions when we are uncertain about what will happen, how valuable it will be, and even how to make decisions. Even the most sure-footed amongst us are sometimes uncertain about all three, but surprisingly little attention has been given to the latter two. The three essays that constitute my thesis hope to do a small part in rectifying this problem.

The first essay is about the value of finding out how to make decisions. Society spends considerable resources funding people (like me) to research decision-making, so it is natural to wonder whether society is getting a good deal. This question is so shockingly underresearched that bedrock facts are readily discoverable, such as when this kind of information is valuable.

My second essay concerns whether we can compare value when we are uncertain about value. Many people are in fact uncertain about value, and how we deal with this uncertainty hinges on these comparisons. I argue that value comparisons are only sometimes possible; I call this *weak comparability*. This essay is largely a synthesis of the literature, but I also present an argument which begins with a peculiar view of the self: it is as if each of us is a crowd of different people separated by time (but connected by continuity of experience). I'm not the first to endorse this peculiar view of the self, but I am the first to show how it supports the benign view that value is sometimes comparable.

We may be uncertain of any decision rules, even those that would tell us how to act when we face uncertainty in decision rules. We may be uncertain of how to decide how to decide how to... And so on. If so, we might have to accept infinitely many decision rules just to make any mundane decision, such as whether to pick up a five-cent piece from the gutter. My third essay addresses this problem of regress. I think all of our decisions are forced: we must decide now or continue to deliberate. Surprisingly, this allows us to avoid the original problem. I call this solution "when forced, do your best".

DECLARATION

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship, the Adelaide University Master of Philosophy Scholarship, and grants from Open Philanthropy Project, the Centre on Long-Term Risk, and the Australasian Association of Philosophy.

Riley Harris

14 Oct. 2021

*I am strongly inclined to believe that
any future person must be either me or someone else.
And ... there is always a deep difference between these.
... Since I am not a separately existing entity,
these beliefs cannot be true.*

— (Parfit, 1984, p. 277)

ACKNOWLEDGMENTS

I felt odd signing the statements of authorship that begin each chapter. Though they are as true for me as they usually are for others, many people have contributed ideas and discussion that shaped my thinking, and the remaining work is either a minor extension of or combination of ideas already in the literature. It is as if I am a curator rather than an artist. An even more disturbing thought is that, despite my best efforts, I will likely omit someone who provided a vital comment or suggestion—but as I suggest in the body of this thesis, my best is all I have.

I should first thank my supervisors Antony Eagle and Garrett Cul-
lity, who were generous with their ideas, time, encouragement, and
advice. I was incredibly lucky to have them on my team. This thesis
would not be the same without their many substantive contributions
and guidance over the years of this project.

Many helped me out with comments, questions, and sugges-
tions. Thanks to Atheer Al-Khalifa, Francesca Bunkenborg, Krister
Bykvist, Pedro González Fernández, Al Hájek, Brian Hedden, Greg
O’Hair, Henry Phillips, Huon Porteous, Elinor Pryce, Dook Shepherd,
Nicholas Smyth, Jeremy Strasser, Zachary Swindlehurst, Danny
Wardle, Hayden Wilkinson, and Patrick Williamson.

Especially memorable were the contributions of Daniel Muñoz,
Matthew Nestor, John Quiggin, Pamela Robinson, Katie Steele, and
Tim Williamson (the younger). Phil Trammell deserves an extra
special mention for his vital insight and guidance.

I was fortunate to share my various half-formed ideas at several
AAP conferences and Adelaide Philosophy Colloquiums, the ANU
Philosophy Society Seminars, a postgraduate-student-led ‘Hivemind’
session, the Adelaide Philosophy Club, and the Effective Altruism
Club. At each the audience was overflowing with helpful suggestions
that improved my work in almost every respect, too numerous to
thank individually.

A number of others aided my development by giving me advice,
book suggestions, and opportunities to teach or attend courses.

Thanks, Nick Buchdahl, Philip Gerrans, Oscar Peralta-Gutierrez, David Janků (and the Effective Thesis Project), Paul Pezanis-Christou, Craig Tailor, Raymond Vozzo, and (especially) Ralph-Christopher Bayer and Virginie Masson. Michael Lazarou helped me as a budding writer, Luke Keating Hughes helped me as an amateur mathematician, and Ewelina Tur helped me become a better human. Thanks also to the graduate communities at Adelaide and ANU (especially Katie Steele, without whom the visit would not have been possible).

I couldn't have written this thesis without many kinds of support from my family. My mum, Penny, has always supported my interests, loved my odd brain, and given me what she thought I needed. My Dad, Geoff, has always been there to support me; and is one of the most caring and selfless people I know. I've never cared for anyone as much as my brother, Piper, and I'm grateful for his acceptance.

Many were deeply important to me during perilous period of research. Alex, Jaden, Jo, Justina, Jess, Matt, Eris, Alex, Ted, Tommy, Rougá, Gherkin, and Tibby all deserve their own paragraph.

Harry David helped with editing and proofreading the manuscript, and Natasha Madden helped editing chapter 3. Thank you also to Konstantinos Leledakis and Danny Wardle for help with typesetting.

It is hard to think lofty thoughts without food and shelter. The Australian government, University of Adelaide, Open Philanthropy Project (and Good Ventures), Centre on Long-Term Risk, and Australasian Association of Philosophy collectively provided generous funding.

CONTENTS

1 INTRODUCTION	1
1.1 Synopsis and discussion of Chapter 2	3
1.2 Synopsis and discussion of Chapter 3	5
1.3 Synopsis and discussion of Chapter 4	7
STATEMENT OF AUTHORSHIP	9
2 INFORMATION VALUE & DECISION THEORETIC HIERARCHIES	11
2.1 Introduction	11
2.2 A Model of Information Value	15
2.3 Background	32
2.4 How to Compare Decision Theories	41
2.5 Valuable Information	46
2.6 Comparisons of Value	51
2.7 Against the Analogue Principle	54
2.8 A Potential Problem	59
2.9 Conclusion	62
STATEMENT OF AUTHORSHIP	63
3 WEAK COMPARABILITY	65
3.1 Introduction	65
3.2 The Allure of Stake-Sensitivity	66
3.3 The Case against Comparability	70
3.4 Initial Replies in Favour of Comparability	72
3.5 Normalisation: The Best Attempt at Strong Comparability	74
3.6 Against Normalisation	76
3.7 A Final Argument for Weak Comparability	78
3.8 Conclusion	80
STATEMENT OF AUTHORSHIP	81
4 WHEN FORCED, DO YOUR BEST: HOW TO MAKE DECISIONS IN THE FACE OF REGRESS	83
4.1 Introduction	83
4.2 Extant Proposals	88
4.3 When Forced, Do Your Best	93
4.4 Unbounded and Arbitrary Evaluation	97
4.5 Conclusion	100
5 DIRECTIONS FOR FUTURE RESEARCH	101
A APPENDIX	113
A.1 Formalisation of the Examples in §2.7	113
A.2 Formalisation of Example in §2.8	115

LIST OF FIGURES

Figure 2.1 A Natural Hierarchy of Decision Theories (with
value ranges on the x-axis) 23

LIST OF TABLES

Table 2.1	The Intersections of Information Types	14
Table 3.1	Jill’s Decision 1	68
Table 3.2	Jill’s Decision 2	69
Table 4.1	Ada’s Decision 1	86
Table 4.2	Why not always deliberate? (1)	99
Table 4.3	Why not always deliberate? (2)	99

INTRODUCTION

Normative uncertainty is a fact of life.

I write this from my favourite cafe near my home in Adelaide, sipping a flat white. I'm lucky to be in this position, and globally there are many who are less fortunate. The money that I spent on this coffee could have helped someone in need, and although I donate a percentage of my stipend, this coffee constitutes evidence that I could do more. (And indeed a few dollars would significantly help any of the billion poorest people.) Whether I've made the right call depends on the balance of competing moral considerations. To what extent am I obligated to do what is best for all but sacrifice my own personal projects? Indeed, I am uncertain about the norms that ought to govern this decision, so I must decide under *normative uncertainty*.

If I chose to donate, I would have many options. Among other things, I could give the money directly to one of the global poor or fund research that might lower the chance of a pandemic that kills millions of people. I would of course face *empirical uncertainty* about the possible effects of my actions. But I could also face *normative uncertainty* in the form of *value uncertainty* (for example, how to trade off the needs of present and future people) and *risk uncertainty* (for example, which risks are appropriate to take and which aren't).¹

Value claims are the natural domain of moral philosophy.² Though most moral theories fail to distinguish between them, there does seem to be a meaningful distinction between value uncertainty, risk uncertainty, and ordinary empirical uncertainty.³ Risk uncertainty is a natural domain of normative decision theory, though it is also (sometimes implicitly) a domain of moral philosophy.

It goes without saying that moral philosophy has a rich history, but decision-making under moral uncertainty has been much less thoroughly explored. The issue was picked up early on by a few Catholic theologians,⁴ but was largely ignored until the last quarter

¹ Indeed, some risks seem appropriate (leaving the house each morning) and others don't (driving drunk). In other cases I am uncertain which risks to take.

² However, inquiry into value uncertainty can inform other areas. MacAskill, Bykvist, and Ord (2020, §3.2) discuss the relation between uncertainty about the aggregation of values across moral theories, and the aggregation of preferences across people.

³ Of course, the distinction might not be fundamental, and each may reduce to empirical uncertainty ultimately. But the fact that biology is ultimately reducible to physics does little to limit the usefulness of advances in biology.

⁴ For example, Liguori (1785), Medina (1577), and Pascal (1657). See New Catholic Encyclopedia (2003) and Sepielli (2010, §1).

of the twentieth century, and the body of work on it has expanded significantly more recently.⁵ The treatment of decision-making under empirical uncertainty has a rich history, beginning in the eighteenth century with the development of expected-utility theory,⁶ and continuing with a flurry of related developments throughout the twentieth century.⁷ Some of this work paved the way for understanding information value.⁸ Many of the developments in the second half of the twentieth century concerned the actual behaviour of humans, which produced many interesting departures from the standard approach.⁹ These developments, however, were largely not applied to understanding how agents *should* make decisions, and the field of *normative* decision theory is still in its infancy. The first serious normative rival to expected-utility appeared only in the last decade,¹⁰ and the third is still in development.¹¹ Perhaps because of this focus on descriptive decision theory, or perhaps because of the inherent difficulty of deciding without knowing how, decision-theoretic uncertainty has received shockingly little treatment; I know of just a few papers that tackle it.¹²

This thesis touches on an exciting and new body of literature on normative uncertainty. (I review the most relevant literature more thoroughly throughout.) Indeed, many interesting questions occupy this space. *How do we make decisions under normative uncertainty?* naturally leads to such questions as *Are we able to compare the value claims of moral theories, and how can we know?*, *Which moral theories are plausible?*, *What should our credence be in various moral theories?*, *How*

5 For example, Bostrom (2009a), Gracely (1996), Greaves and Cotton-Barratt (2019), Greaves and Ord (2017), Gustafsson and Torpman (2014), Harman (2015), Hedden (2016), Hudson (1989), Lockhart (2000), MacAskill (2016a,b), MacAskill, Cotton-Barratt, and Ord (2020), MacAskill and Ord (2018), Newberry and Ord (2021), Sepielli (2009, 2010, 2017), Tarsney (2017, 2018a,b, 2020), Weatherson (2014), and Zimmerman (2008). See Bykvist (2017) for a brief overview and MacAskill, Bykvist, and Ord (2020) for a recent book-length treatment.

6 See D. Bernoulli (1954) and J. Bernoulli (1975).

7 Especially Friedman and Savage (1952), Jeffrey (1965), Savage (1954), and Von Neumann and Morgenstern (1944). See Steele and Stefánsson (2020) for a brief overview focused on normative decision theory.

8 For example, Blackwell (1953), Blackwell and Girshick (1954), Demski (1980), Good (1967), Keasey (1984), Lawrence (1979), Ramsey (1990), and Wakker (1988). I especially like the summary by Trammell (n.d.).

9 For example, Camerer, Loewenstein, and Rabin (2004), Chew (1983, 1989), Nakamura (1995), Quiggin (1982), Tversky and Kahneman (1992), Wakker (1990), Wakker, Erev, and Weber (1994), and Yaari (1987). See Starmer (2000) for an overview.

10 Buchak (2013). See also the summary in Buchak (2017a), and for discussion see Briggs (2015), Buchak (2010, 2015, 2016, 2017a,b), and Joyce (2017).

11 Bottomley and Williamson (n.d.).

12 See, for example, MacAskill (2016b), MacAskill, Vallinder, et al. (2021), and Trammell (2021); but only Trammell tackles decision-theoretic uncertainty of the kind I'm interested in.

should we solve the special problems of fanatical theories (when a theory takes over the decision-making process no matter how implausible we find it) and infinite regress (which I will explain soon)?, What kinds of “ought” claims should these theories strive to answer?, What is the interaction between these claims and metaethical claims?, and many others. Of course, many of the interesting questions lie in areas that are not the special domain of normative uncertainty—for example, moral philosophy and normative decision theory.

This thesis is not a traditional thesis. Instead of a book-length treatment of a single topic, it is a collection of essays, each of which answers a different question:

- When is information valuable under decision-theoretic uncertainty?
- To what extent can we compare value claims made by moral theories?
- What is the best way to solve the regress problem?

1.1 CHAPTER 2 – INFORMATION VALUE & DECISION THEORETIC HIERARCHIES

My first essay is on the value of information.

Information is learning. We might learn empirical facts about the world, facts about value, or facts about what risks to take. I ignore information about value in this essay ¹³ This simplifies things, and perhaps even avoids fundamental problems for ensuring *the value of information about value* is well defined, one of which I explore in chapter 3.

What should we seek to learn? We must choose what research directions to pursue as individuals, what to Google, and, as a society, what research to fund. So we need to understand the *value* of information. Because many of these decisions are made under normative uncertainty, it is worth understanding information value in this context. But there are also special cases that can only be understood under normative uncertainty. Society spends significant resources on philosophers, economists, mathematicians, and others, and part of their time is spent studying how we should make decisions. It is natural to wonder how valuable this research is, which can only be understood in the context of normative uncertainty.

¹³ MacAskill, Bykvist, and Ord (2020, §9) ignore information about risks and focus on information about value. Many others only explore information about the world. I have more to say in chapter 2.

Gathering information entails a kind of risk: students embark on a thesis without knowing their eventual destination, organisations give grants without knowing the results in advance, and so on. If we knew the results in advance, the inquiries wouldn't be informative. Because gathering information entails a kind of risk, what risks we ought to accept will determine what we ought to seek to learn. Things get dicey and perhaps circular when we consider information *about* which risks to take. I explore problems related to this in chapter 4, but in my first essay I am ecumenical about different solutions to these problems.

For my purpose, we must also be able to compare our different claims about which risks are worth taking. These are the claims of decision theories, so we need to compare choiceworthiness between decision theories. If we don't, then there will be no well-defined answer as to the value of information about which risks to take. As I mentioned, I am ignoring value uncertainty. This makes it natural to assume that all reasonable ways of evaluating risk agree about the value of acts which can only result in one particular amount of value. So decision theories agree on the value of sure outcomes, and disagreement is confined to risky outcomes. Within each decision theory are the seeds of comparability: the theories tell us which sure outcomes are equivalent to which risky outcomes, and so ensuring sure outcomes are comparable across decision theories is all we need in order to compare risky outcomes. When we ensure that, decision theories have an attractive property: no theory will assign a risky outcome a greater value than its greatest possible value nor a lower value than its lowest possible value. I call this property *natural boundedness*.

Some have thought that we ought to accept any free information because such information cannot make us worse off.¹⁴ But when we consider decision theory, it is clear that not all possible information is valuable. We might learn that our acts are worse than we had thought if we learn that a particularly pessimistic decision theory is true; but the information might not yield any gain in empirical information or any change in the *relative* value of our acts (which might have led to a better choice). I generalise this point by showing that all decision theories accept information that satisfies a certain condition I call *weak evaluation dominance*. If this is satisfied, then we will accept costless information.

How might we begin to compare different research directions? We might compare them based on just their overall information value, and I illustrate how. We might also compare them in a more sophis-

¹⁴ This stance is at least somewhat controversial because it is denied by the rivals to expected-utility theory. I discuss this in the paper.

ticated way by looking at each interval of probability and comparing value within that interval. If one source of information dominates for each of these intervals, then it is what I call *stochastically more informative*. If an agent knows some information is stochastically more informative, then they should consider it more informative in general. I show that this provides a compelling restriction on the space of possible decision theories.

In order to derive my results, I appeal to the idea of higher-order decision theories. In the same way that decision theories tell us how valuable our acts are under empirical risk, higher-order decision theories tell us about how valuable our acts are under normative risks (of lower orders). But what are these higher-order decision theories? I argue that an earlier interpretation of higher-order decision theories is mistaken and leads to results that we should not endorse.

A few parts of “Information Value & Decision Theoretic Hierarchies” could have been explored on their own. Namely, I argue that decision theories produce comparable values when we separate issues of belief, value, and risk attitudes (§2.4). I also argue that there may be no deep analogy between higher-order decision theories and regular old decision theories (§2.7). These sections cover issues that are deeply related to the core results of the essay, and I could not separate them in a satisfactory way in the time afforded to me in a master’s program. Regrettably, this makes the essay longer than is ideal. A little solace can be found by the weary reader in the fact that the latter sections can be read separately from each other, as if it’s a choose-your-own-adventure book with multiple endings.

1.2 CHAPTER 3 – WEAK COMPARABILITY

In my second essay, “Weak Comparability,” I relax the assumption of value uncertainty that underlaid my results about information value. In order to extend these results to value uncertainty, we would need a way of comparing value claims. Different things may be appropriate if we can compare value than if we can’t. If we want to act appropriately, which comparisons can be made is decision relevant. Thus, I begin to address what I perceive to be the main barrier I face in extending the results of my first essay from uncertainty about decision theories to uncertainty about value.

Much valuable work has been done on this topic, and so I first review the literature. Some have argued that such comparisons are impossible or unjustified. One argument against value comparisons appeals to particular moral theories in particular decision situations in which it seems impossible to compare the claims of these theories. Another argument is that moral theories do not, by themselves,

have the resources to allow us to compare them. A third argument maintains that approaches that rely on comparisons may be suspect. Since value claims are made by moral theories, this is known as the *problem of intertheoretic value comparisons*. Value claims are also made by utility functions, so this problem is also called *the problem of interpersonal utility comparisons*. The problems are formally equivalent for moral theories that satisfy a set of standard choice axioms (such as those given by decision theorists), but I focus on the moral aspect. I think others have provided compelling rebuttals to each of these arguments.

I then review the most promising approach that would ensure value is always comparable. This approach normalises the functions that represent moral theories (or utility) at their variance.¹⁵ I illustrate that this approach, though promising, violates our intuitions in particular cases.

In fact, I think no approach can account for these cases, and I produce an impossibility theorem. No such approach will be able to satisfy *determinacy* (every act must have an exact, determined evaluation), *consistency* (each theory must be combined consistently), and *intuitiveness* (the approach must respect our intuitions about the strength of our various reasons).

Reviewing this literature makes me pessimistic that value comparisons can always be made but optimistic that they can sometimes be made. I call this position *weak comparability*.

To argue for my position, I start by noting that philosophers believe a number of peculiar things, not the least of which is that our ordinary view of the self is muddled. We tend to think that there is a coherent, stable (yet ever-changing) self. But when we try to defend this view, we get stuck. In exactly what ways is the person typing this sentence the same as Riley at five years old? We have different dreams, desires, memories, and dispositions. We're made of different atoms, organised in a structure that bears at most a passing resemblance.¹⁶ Parfit (1984) famously argues that personal identity is not a deep fact; rather, each of us is a series of people connected by psychological continuity. Without the supposedly deeper facts of

¹⁵ See MacAskill, Cotton-Barratt, and Ord (2020).

¹⁶ I'm also not sure what is supposed to be coherent between these two selves (and perhaps infinite others). I have thoughts and a body, but neither seems satisfyingly self-contained or coherent: My thoughts are in your head as you read this, and they are also contained in the many reminders, notes, and lists that I have scattered around my desk. So perhaps it's just thoughts that are contained within my body, but the delineation of my body too seems somewhat arbitrary. Why not draw the boundary at the edge of my nervous system, rather than at the edge of my skin? Why not draw it at the edge of the biosphere throughout all history, of which I am a part that could not survive on its own? My pattern is, after all, just a small part of the pattern that began with the first self-replicating cell.

personal identity, I am in no deep sense the same person as I was at five; rather, I am merely psychologically continuous with selves that are psychologically continuous with that person. This makes it seem as if I am not a single person continuous through time but many people connected by psychological continuity relations, which are stronger between nearer-in-time selves. I take this view seriously and apply it to the debate about value comparisons.

I use this view of the self to argue for weak comparability. We can sometimes compare our opinions, and the value of our experiences, even though they happen at different times in our lives. For example, I know that my worst experiences of many years ago were much worse than my current ones. I also know, when I change my mind, which attitude is stronger. Because across time it is as if I am many people, I can (sometimes) compare value across different people. Because of the similarities between comparisons between people and comparisons between moral theories, I can also sometimes compare value across theories. Thus I use this peculiar view to defend my mundane position.

My conclusion, that only weak comparability holds, makes me think that the way to apply the model of my first essay to cases of moral uncertainty must be essentially case by case. The kind of mathematical model I construct in the first essay would require the strong-comparability assumption, which I deny with the impossibility theorem of my second essay.

1.3 CHAPTER 4 – WHEN FORCED, DO YOUR BEST: HOW TO MAKE DECISIONS IN THE FACE OF REGRESS

Ordinarily, norms tell us how we ought to act. Even when we are uncertain about norms, it remains natural to think that we ought to act in some way or other. But then how we act under normative uncertainty is governed by a second set of norms, which we may also be uncertain about (and we may appeal to still more norms about how to act under that kind of uncertainty). A natural concern is that this leads to a vicious regress that makes the kind of project I embark on in this thesis, especially in my first essay, inherently suspect. In that first essay I was pluralist about the solution to this problem; now I address the concern directly.

My own solution is that we can only do our best when we face a forced choice; and when we see our options clearly, we realise that we always face a forced choice. The forced choice is between deciding now and deciding after deliberation. If deliberating before deciding seems worthwhile given our current considerations, then we ought to do so; if deciding now seems worthwhile, then we ought

to do that. Normally, it is clear that deliberating forever is not the best option, and so we need not consider infinite levels of the regress. This is how we avoid the regress problem.

I consider two other solutions and argue that mine is superior. I think the proposal that we ought to deliberate as much as we can gets particular cases wrong. When we see five cents in the gutter and wonder whether to pick it up, we should make a quick decision and save our mental energy for more important matters. Another proposal says we ought to consider the hierarchy of potential values that we could place on acts at each level of the regress, and assign acts the value that lies at the infinite intersection of these value ranges (one for each level). This solution, like the first, is overly demanding on real people. We must also make strong assumptions to guarantee a unique value that intersects each of these value ranges.

I also consider potentially problematic cases for my view—cases in which agents are idealised such that they don't face the same costs of deliberating as I do. My solution works for agents that are minimally realistic, and this is enough. We do not need to solve the problems faced by entirely unrealistic agents.

In my first essay, I constructed a hierarchy of decision theories. In order for information value to be well-defined, we need this hierarchy to relate somehow to an overall notion of value given (potentially infinite) uncertainty. My treatment of this issue in my first essay is relatively short, so you could be forgiven for thinking I take this issue to be unimportant. But in fact this is a crucial underlying issue, and my thinking in the third essay led me to my view of information value in the first.

STATEMENT OF AUTHORSHIP

TITLE OF PAPER:

Information Value & Decision Theoretic Hierarchies

PUBLICATION STATUS:

Unpublished and Unsubmitted Work Written in Manuscript Style

PUBLICATION DETAILS:

N/A

NAME OF PRINCIPAL AUTHOR (CANDIDATE):

Riley Harris

CONTRIBUTION TO THE PAPER:

Devised the arguments and wrote, proofread, polished, and formatted the paper.

OVERALL PERCENTAGE (%):

100%

CERTIFICATION::

This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.

Riley Harris

14 Oct. 2021

INFORMATION VALUE & DECISION THEORETIC HIERARCHIES

2.1 INTRODUCTION

An easy way to think about the value of information is as the maximum amount a rational agent should be willing to pay for that information. More precisely, the value of information is the price that would make a rational agent indifferent between how things are and how things would be if they had the information. Although it is a price, the value of information is not a dollar amount. The value of information is equivalent to the value of what a particular dollar amount could buy when spent optimally. This is not meant in a narrow sense: we might ‘spend’ money on having more time off, or on an interesting experience, or on promoting our deepest values. Willingness to pay for information really just means *the value to you of whatever you would be willing to give up in exchange for that information*. Thus, information value is the maximum a rational agent should be willing to give up in order to obtain information.

Information comes in different flavours. It might tell you the truth about decision-critical claims with certainty, or it might only move your beliefs closer to truth. A claim is decision critical if it affects the potential value of some act which is available to the agent. It will be useful to follow an example.

Example 2.1.1 (Drug X). Artemis is deciding whether to take drug X. Drug X has different effects on different people, but these effects are exactly determined by genetics. People with gene A but not gene B experience a pleasant boost to their well-being, which is persistent and without side effects for as long as they continue taking X. People with B but not A experience no noticeable effect from X and usually stop taking it. Finally, people with both the A and B genes experience six months of moderate unpleasant side effects, even when they immediately discontinue taking X. Artemis is unsure whether they have gene A, gene B, or both.

The first distinction worth making is between *perfect* and *imperfect* information. Roughly, perfect information tells you the truth about

all relevant claims with certainty, while imperfect information merely changes your credence in a way that is truth tracking. That is, imperfect information gets you closer to the truth, while perfect information takes you all the way there. Let's continue to develop example 2.1.1.

Example 2.1.1 (continued) Artemis has two ways of gathering information related to X, aside from simply taking the drug. Two genetic tests are available: one tests for only B; the other tests for A and B. Suppose, for simplicity, that these tests are entirely accurate, so there is no chance of a false positive or a false negative.

Only testing for B provides imperfect information. Upon finding out the results of this test, the test taker may remain uncertain about the effects of their actions. Suppose the test came back positive for B; taking the drug may result in either no noticeable effect (if they do not have the A gene) or six months of moderate side effects (if they have the A gene).

In contrast, the genetic test for A and B provides perfect information. Upon receiving the test results, Artemis would know with certainty the consequences of their actions. Suppose the test came back saying that they had both A and B. They would know that they could stick to the status quo or take the drug and experience six months of moderate side effects.

Usually, we formalise our empirical uncertainty as uncertainty about which state of the world will obtain, represented by a probability function over states. This gives a way of understanding imperfect information. We imagine an agent finds out which states of the world do not obtain. Then that agent can assign those states probability 0 and assign other states a probability in accordance with conditionalisation.¹

When information is about states of the world, the relation between the value of perfect information and the value of imperfect information is as follows. The value of perfect information is an upper bound on the value of imperfect information.² In our example,

¹ Some philosophers allow for alternative updating rules. I expect that my framework can allow for these rules, though I do not specifically investigate this issue. If your preferred updating method departs from Bayes's in particular instances, this might provide interesting test cases both for my account of information value and for your updating method. I encourage exploration along these lines.

² This is true for Expected Utility (see the discussion in MacAskill, Bykvist, and Ord (2020, chapter 9)). It is also trivially true for Maximin and Maximax, which are two of the more extreme cases of Risk-Weighted Expected Utility. Clearly it is

the value of knowing whether you have gene B cannot be higher than the value of knowing you have just A, just B, or both.

Information can be about states or about decision theories (or even about higher-order decision theories). That is, information can be about how the world is or about how we should decide. So far we have distinguished between perfect and imperfect information. I now distinguish between information about the state of the world and information about what decision-making norms are true or best.

Information can be about empirical facts, expressed as information about states. I will call this *pure-state* information. Almost all previous research focused exclusively on pure-state information,³ and the above genetic tests are examples of this kind of information. However, when an agent is uncertain, information can be about what norms are best to guide their decision-making. We can find out about decision theories (or about higher-order decision theories). I will call this *pure-normative* information. Consider the following continuation of example 2.1.1.

Example 2.1.1 (continued 2) Artemis is considering two decision theories, t and t' . They think these possibilities are equally likely to be true or best. Artemis has an opportunity to learn more about decision theory by taking a class at their local university. After attending this class, they will be 90 percent sure of the correct way of weighing risk. Before attending the class, they predict that it's equally likely that they will assign 90 percent probability to t' (and 10 percent to t) and that they will assign 90 percent to t (and 10 percent to t') after taking the class.

This is an example of pure-normative information. In this case, taking the decision-theory class would give Artemis *imperfect* information about the correct decision theory. Thus, we can say it is *imperfect pure-normative* information. Here, by 'decision theories' I mean to include such theories as Maximise Expected Utility and Maximise Risk-Weighted Expected Utility.⁴

In my model there is no fundamental uncertainty about value. I assume that agents would be certain about value if they were certain about states, and they are only uncertain about the value

true for a wide class of plausible decision theories, although I do not know the minimal condition that guarantees this property.

³ The only exception I know of is MacAskill, Bykvist, and Ord (2020), whose authors ignore decision-theoretic uncertainty entirely but allow for moral uncertainty (which I have thus far ignored).

⁴ I specifically don't mean to refer to the evidential-versus-causal debate; see MacAskill (2013) on decision-theoretic uncertainty of that kind.

of acts because of their uncertainty about states and about how to resolve uncertainty about states (and higher-order uncertainty about *uncertainty about how to resolve uncertainty about states*, and so on). Thus pure-normative information includes information that does not pertain to states of the world, including higher-order decision theories (see §2.2), but not information about value itself.

Now that we have distinguished amongst perfect, imperfect, pure-state, and pure-normative information, we might ask, ‘How are these categories related?’ Information cannot be both perfect and imperfect, nor pure state and pure normative; however, other combinations are possible. In table 2.1 I provide examples (related to example 2.1.1) of each of the possible combinations.

	Pure-State	Pure-Normative
Perfect	Genetic test for A and B	God tells you t or t' is best, giving you certainty
Imperfect	Genetic test for B only	Research indicates t or t' is best, giving you 90% confidence

Table 2.1: The Intersections of Information Types

Information can also be, at the same time, informative about what state of the world obtains and what decision theory is best. Thus, when we use the term *information*, without qualification, it can refer to states, decision theories, higher-order decision theories, or any combination of these. To clarify, both pure-state and pure-normative information are subsets of information proper. And information, without qualification, is perfect or imperfect. Thus, information is a general notion which subsumes all of the types I have listed in this section.

For those more familiar with other terms, I note, first, that information can also be called *evidence* (though I ignore cases of misleading or opaque evidence)⁵ and, second, that it is not the same thing as *Shannon information*.⁶

⁵ *The value of information* would most likely translate, in this terminology, into *the value of knowledge* (in the sense of Moss (2016)) because the value of our evidence has to come from the knowledge it gives us.

⁶ Shannon and Weaver’s (1949) approach takes seriously both the possibility of error or distortion to messages and the problems in conveying desired information in the least number of bits. I abstract from these details. I am interested in situations in which a rational agent should accept a particular piece of evidence, and I assume they will know exactly what it means when they receive it.

This chapter is about the value of information under decision-theoretic uncertainty. In §2.2 I will set up a mathematical model of information value. Following other authors, I will extend the uncertainty to higher-order decision theories. In §2.3 I will discuss previous work and how it relates to my own. It is well established that the value of information varies with the choice of decision theory, yet no one has explored the value of information under decision-theoretic uncertainty (an omission I hope to remedy). It may seem odd for me to put the background so late in the chapter, but introducing the model first will allow the reader to get extra mileage from the notation.

The early sections (§2.1–2.3) set the stage for the later sections, which can be read largely independently of one another. The main result is the necessary and sufficient conditions for valuable information, presented in §2.5. In §2.6 I will illustrate various measures of comparative value and show how these give rise to constraints on normative decision theory. The core results about information value require comparisons of value across decision theories, so I will address this issue briefly in §2.4. Later, in §2.7 I will argue against Trammell (2021) by showing that the interpretation of higher-order decision theories as exactly analogous to ordinary decision theories is misguided. I will finish with a minor worry about my framework in §2.8; this worry is easily dismissed.

2.2 A MODEL OF INFORMATION VALUE

This section sets up the formal model for the rest of the chapter. We need to define objective and subjective choiceworthiness, decision theory, beliefs, and the relations amongst all of them. To make things sufficiently general, I do not stop at decision-theoretic uncertainty; instead, I extend the discussion to a hierarchy of higher-order decision theories. We must also understand information, formalised as messages and signals. These concepts together will allow us to formalise information value under decision-theoretic uncertainty.

In the first subsections, I will draw heavily from Trammell's (2021) model of choiceworthiness and the hierarchy structure, though the exposition is my own. In §2.2.3 I will depart from Trammell. From §2.2.4 onwards my discussion is inspired by the economics literature on information value.⁷ Later, in §2.7, I will return to Trammell for a critical note.

Building my model requires devising a lot of mathematical structures, which I believe to be illuminating. However, after seeing these results you may still want to modify the framework, extend it, or

⁷ See Lawrence (1979, chapter 1) for a review.

argue against the whole thing in its entirety. I trust that laying out the model with mathematical precision will aid you whatever your later goals.

2.2.1 *Objective and Subjective Choiceworthiness*

Let there be a finite set of feasible acts, $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$, and a finite set of possible states, $\mathcal{S} = \{s_1, \dots, s_{|\mathcal{S}|}\}$. The function u assigns real values to state-act pairs (also known as *outcomes*).⁸ These are an agent's *objective-choiceworthiness* values of outcomes. Following Savage (1954), we can interpret outcomes as good or bad states of affairs. States are features of the world that may affect the state of affairs but are not determined by an agent's actions. There are multiple possible interpretations of utility. It could be that agents know how valuable various outcomes are to them, know facts about their own welfare,⁹ or perhaps know how valuable outcomes are in the moral sense.¹⁰ My model allows for any of these possible interpretations. Depending on your interpretation, choiceworthiness might be objective only in the sense that its value is not defined relative to an agent's beliefs. So objective-choiceworthiness values may differ amongst agents. All I propose is that there are facts about the choiceworthiness (for an agent) of various state-act pairs and that the agent knows these facts.¹¹

Formally, we can define objective choiceworthiness for a given utility function:

Definition 2.2.1. The *objective choiceworthiness* of a , given s , is the i th element of the $|\mathcal{A}|$ -vector $u(\mathcal{A}, s)$.

To simplify the terminology throughout, I denote the objective choiceworthiness of an outcome as $u(a|s)$. We can further denote the

⁸ We can avoid the debate about causal versus evidential decision theory by assuming the probabilities of states are independent of the chosen act. See MacAskill (2016b) for a discussion of these issues in the context of decision-theoretic uncertainty.

⁹ See, for example, Fryxell (n.d.).

¹⁰ We also might extend this to moral uncertainty. The most obvious approach would be to make \mathcal{S} richer by including moral theories and having the agent accept a normalisation method that yields real values that are comparable (see, for example, MacAskill, Cotton-Barratt, and Ord (2020)). In this case, the results will apply with only minor modification to my framework. If values are not comparable (I think they are not; see chapter 3) or moral theories cannot be represented as assigning a single real value to each act, then the model will have to be more complicated. If so, perhaps the best approach will be to model information value in the context of a single interesting case study, rather than in general (this is the approach of MacAskill, Bykvist, and Ord (2020, §)).

¹¹ See Trammell (2021) for further treatment of the relation between objective and subjective choiceworthiness.

$|\mathcal{S}|$ -vector of possible objective-choiceworthiness values of some act a as $u(a)$, and we can denote the $|\mathcal{A}| \times |\mathcal{S}|$ matrix, generated by $u(\mathcal{A}, \mathcal{S})$, that specifies possible objective-choiceworthiness values for all acts as $u(\mathcal{A})$.

In assuming that the set of real numbers, \mathbb{R} , is rich enough to capture objective choiceworthiness, I have ruled out lexicographic preferences. I make the further assumption that objective choiceworthiness is measured on a cardinal scale: the relative distance between pairs of acts is meaningful, and u is unique up to positive affine transformations.¹² This seems plausible and allows for the mathematics to be more simple than they would be otherwise, though I welcome extensions of this framework to allow for lexicographic or merely ordinal objective choiceworthiness.

Throughout, the agent's rational beliefs will be represented by probability distributions that satisfy the Kolmogorov probability axioms.¹³ In addition, I will assume probabilities obey conditionalisation and the principal principle.¹⁴ I will fully define an agent's

¹² Note the difference between this approach and that of axiomatic decision theory. I assume there is a cardinal objective-choiceworthiness scale that may be represented by a function mapping from outcomes to \mathbb{R} . The Von Neumann and Morgenstern (1944) approach is to show that *if* an agent's choices satisfy some axioms of decision theory, *then* there is a representation of their decision such that they maximise the expectation of *some* utility function unique up to positive affine transformation. Other axiomatic approaches are similarly motivated, though they differ in the details. I am implicitly assuming that when an agent is certain about the objective choiceworthiness of an act, the subjective choiceworthiness of that act exactly equals its objective choiceworthiness and that when an agent is uncertain, subjective choiceworthiness is at least constrained by natural boundedness (see §2.5). Trammell's approach is more elegant; in giving structure to subjective choiceworthiness first, he does not need to (even implicitly) assume any additional structure, as it all flows down from the top. Because I think objective choiceworthiness is the primitive notion, I forgo this elegance.

¹³ I need to clarify two points. I consider the probability distribution over states as an example here, but my point applies to probability distributions everywhere in this chapter. First, credences do not act directly on states; instead they act on sets of states (often called 'events'). Thus, we need to generate a collection of subsets of \mathcal{S} containing \mathcal{S} and \emptyset which is closed under complementation and countable union. This collection of subsets is a σ -algebra, so naturally we might call it $\sigma(\mathcal{S})$. To save space, though, I will refer to the 'probability distribution over states' rather than the 'probability distribution over sets of states'. Second, this probability distribution over \mathcal{S} satisfies the Kolmogorov axioms if three conditions are met. (1) For all $o_i \in \sigma(\mathcal{S})$, $p(o_i) \geq 0$. (2) $p(\mathcal{S}) = 1$. (3) For any countable sequence of disjoint o_i , $p(\cup_{i=1}^{\infty} o_i) = \sum_{i=1}^{\infty} p(o_i)$. For those not already familiar with the modern foundations of probability theory, rest assured these are all standard assumptions which generate an intuitive probability measure (see D. Williams (1991, part A), for example). The controversial assumption here is that beliefs can be represented with probability. See Joyce (1998), Joyce (2010), R. Bradley and Stefansson (2017), Pettigrew (2016), and Steele and Stefansson (2020, §5) for discussion. I leave aside these more complicated issues relating to rational belief.

¹⁴ Conditionalisation is essentially the claim that upon encountering some evidence E , an agent comes to believe that their prior beliefs are conditionalised on E .

initial beliefs later in this section. For now it is enough to note that for all $s \in \mathcal{S}$ the agent initially assigns a positive probability to s . We can call this probability distribution over possible states $P(\mathcal{S}) = \{p(s_1), \dots, p(s_{|\mathcal{S}|})\}$.

From this probability distribution over possible states we can construct a probability distribution over possible objective-choice-worthiness values for each act. Let $\mathbb{P}^{|\mathcal{A}|}$ be the set of all finite probability distributions in $\mathbb{R}^{|\mathcal{A}|}$, and let some $P(u(\mathcal{A})) \in \mathbb{P}^{|\mathcal{A}|}$ represent the choiceworthiness distribution entailed by u .¹⁵

Definition 2.2.2. A *choiceworthiness distribution*, $P(u(\mathcal{A}))$, is a probability distribution over possible choiceworthiness values, $u(\mathcal{A})$.

Given this distribution in the *objective*-choiceworthiness values of acts, there is still a sense in which some acts are more appropriate to choose than others. Consider this example.

Example 2.2.1 (An Obvious Choice). Bjørn sits at a desk with two buttons. An enormously powerful alien god has wired these buttons to do incredible things. One button brings about one of the best of all possible worlds for Bjørn, taking full account of all that is valuable to Bjørn. The other button would end Bjørn's life, as would refusing to press either button. Bjørn is almost, but not quite, certain that the left button (L) is the one that would create the best of all possible worlds for them. They are also almost, but not quite, certain that pressing the right button (R) or failing to choose would bring about their early demise.

In this example, Bjørn knows how objectively choiceworthy such state-act pairs as 'pressing L if the alien god wired L to bring about one of the best of all possible worlds' is. But they do not know which *act* is objectively the most choiceworthy, because they are not entirely certain which button is which, and they are therefore uncertain of the effects of pressing the button. However, it is obvious to me that they should still press L despite it not being guaranteed to be the most objectively choiceworthy. Relative to their beliefs, L is their best

In assuming this, I am simplifying things a bit, such that an agent can never receive a later piece of evidence that claims that an earlier piece of evidence was wrong, nor can they misinterpret any piece of evidence or be uncertain which piece of evidence they received. Further work might try to relax this assumption. Meanwhile, the principal principle states that for any state s , an agent's belief in s is their expectation of its objective chance. See Pettigrew (2016) for further discussion.

¹⁵ Many other probability distributions would serve the purpose of $P(u(\mathcal{A}))$. See Trammell (2021, p. 6-11) for discussion.

choice; if they press it, they will with near certainty enter the best possible world. If they don't press it, they will almost certainly or certainly die (depending on whether they press R or fail to choose). Pressing L is, in a belief-relative sense, best.

Examples such as 2.2.1, and more realistic examples, motivate the notion of *subjective* choiceworthiness.¹⁶ Sometimes certain acts are best despite our uncertainty about their objective choiceworthiness.

Notice too that this sense of choiceworthiness depends on our beliefs about the world. If Bjørn were almost certain that pressing R would lead to the best possible world for them, and if they were certain that only one of the buttons would have that effect, then pressing R would be subjectively best (or most choiceworthy). Notice too that these two senses of choiceworthiness need not lead to the same answers: I retain my intuition that the subjectively best thing the agent can do in example 2.2.1 is press L even if the actual state of the world is such that this would kill them.

I will represent subjective choiceworthiness with a function v that assigns to each act in \mathcal{A} a real number which will be sensitive to an agent's utility function and their initial set of beliefs about relevant claims.¹⁷

When agents don't know the objective-choiceworthiness values of their acts, they must choose under uncertainty. Decision theories tell us how to evaluate our options under empirical (or pure state) uncertainty. Decision theories can be *descriptive* of how we actually make decisions or tell us with *normative* force how we should evaluate our options. I am here interested in decision theory as an answer to the question of how a rational agent *should* weigh risky options. When decision theories are normative, they tell us about subjective choiceworthiness. Here I discuss the kind of claims that normative decision theories may give about value in a general way. In §2.3 I will discuss some extant proposals from the literature.

Implicit in decision theories' value claims are deontic claims. When we know how valuable our acts are, we know how to act. It is obvious

¹⁶ In the real world, there are many hard cases, in which we do not know which option is subjectively best. I have abstracted from some of the details of how this might come about: in the model, agents always have fully specified, precise beliefs about states, and they always know exactly how valuable outcomes are. You might relax these two assumptions if you wish to account for hard cases, but there are other ways of doing so. This chapter supposes decision-theoretic uncertainty is an infinite hierarchy, and it gives several ways that subjective choiceworthiness might still be well defined despite this in §2.2.3. Among others, we could allow the agent to be uncertain exactly how to ensure subjective choiceworthiness is well defined. Or we could give the agent other cognitive limitations; for example, we could accept Trammell's solution but give an agent the ability to merely estimate the subjective choiceworthiness for themselves.

¹⁷ The structural assumptions I made about u need not be made again here; they follow directly from the relation between u and v .

that when agents are certain, they should simply pick the option that has maximal value. The same is true when acts have uncertain outcomes. We should take whatever act is most valuable in light of our uncertainty. So in making claims about how valuable our acts are, decision theories also make deontic claims about what an agent should do in a situation—claims about which of their acts are rational and which are irrational. I take it as a primitive that an agent should maximise subjective value when objective and subjective values are not co-extensive.¹⁸ Thus, while my model emphasises the evaluation aspect of decision theory, one could equally well emphasise the choice aspect.

In cases in which we face risk, decision theories are claims about subjective choiceworthiness.¹⁹ Formally, a decision theory assigns to each act a real number representing its subjective choiceworthiness. Decision theories take into account an agent's choiceworthiness distribution—in other words, the probability distribution over possible objective-choiceworthiness values for their acts.

Definition 2.2.3. A *decision theory* is a function $t : P(u(\mathcal{A})) \rightarrow \mathbb{R}^{|\mathcal{A}|}$ that represents claims about subjective choiceworthiness.

So, given some probability distribution over objective-choiceworthiness values (given by an agent's uncertainty about the state of the world), a decision theory makes claims about the subjective choiceworthiness of acts. A decision theory takes into account both how likely various possibilities are and how valuable they are according to the objective evaluations of u . That is, it accounts for how valuable an agent's acts are not only in terms of the objective value of outcomes, but also in terms of risk.

In §2.3 I will discuss in more detail the decision-theoretic proposals. For now, I ask that you take it as given that decision theories are, in their most general form, functions mapping from probability distributions over objective choiceworthiness to real numbers that represent claims about subjective choiceworthiness. If these subjective-choiceworthiness value claims are comparable, and if one decision theory assigns x to a and another assigns y to a , where $x > y$, then we will say both that a is more choiceworthy according to the first theory than the second and that it is more choiceworthy

¹⁸ This is a decision-theoretic claim that sits outside of the hierarchy. I do not aim to model agents who are uncertain about this.

¹⁹ Technically, here and elsewhere, I am being a little loose with my language. Decision theories make claims about what subjective choiceworthiness ought to be relative to an agent's objective-choiceworthiness distribution (but not, for example, relative to their distribution over decision theories). In the language of Trammell (2021), decision theories make claims about what subjective choiceworthiness *1-ought* to be. Because it is always clear from the domain of the function what information these claims are relative to, I do not mention this explicitly each time.

by $x - y > 0$. In §2.4 I will discuss whether the evaluations given by decision theories are comparable. For now, I suppose that they are.

2.2.2 *The Hierarchy Model of Decision-Theoretic Uncertainty*

I am departing from most of the decision-theoretic literature in a very important way. Usually, subjective choiceworthiness is understood to be the value that a decision theory *assigns*. For me, decision theories do not assign subjective choiceworthiness directly, except in the special case in which an agent accepts a single decision theory. Instead, decision theories make *claims* about subjective choiceworthiness. These claims tell us what subjective choiceworthiness *would be* if an agent accepted a certain decision theory (that is, if they knew it to be true). Because I wish to model decision-theoretic uncertainty, these are claims about which an agent who is uncertain which decision theory is true will also be uncertain.²⁰

To make an analogy, first imagine a gymnastics routine that is judged by a single person. After each routine, that person *assigns* a score. That score represents how good the routine is. Now imagine a panel of judges. After a routine, each member of the panel does not assign the final score; instead, they make a *claim* about what the final score should be. Each judge only contributes to determining the overall score; none can assign a score. Similarly, under decision-theoretic uncertainty, each decision theory makes a claim about the value of an act, but none can assign value on their own.

This allows me to clarify what subjective choiceworthiness is. The idea is that there is a sense of subjective ‘all things considered’ value.²¹ Most of the literature considers objective choiceworthiness and uncertainty about states. Normally, decision theories can determine subjective choiceworthiness because they can handle all of the relevant uncertainty. Modelling decision-theoretic uncertainty makes things more complicated, as that uncertainty means decision theories can only make claims about subjective choiceworthiness.

²⁰ I leave out the special case in which an agent is uncertain about decision theories that assign all of their acts the same value. This could arise if the set of available acts is such that theories agree on the value of those acts, while disagreeing about the value of other possible acts that the agent does not actually face. In this case, only a slight modification is needed to amend my claim. Here, assigning values is what decision theories do, so if different theories assign the same values to acts across any possible triple $\langle S, A, u \rangle$, this is a problem with theory individuation.

²¹ Some might be sceptical that there is such a thing as subjective choiceworthiness for agents who are uncertain about decision theories. Although I will assume there is, I note that (the *moral-uncertainty* equivalent of) this claim is controversial; see Harman (2015), Hedden (2016), and Weatherson (2014).

I closely follow Trammell's (2021) approach here.²² I will construct a (potentially infinite) hierarchy of higher-order decision theories that represents all of an agent's uncertainty.

First, I need to motivate the notion of a meta-level theory. What would it mean to have a theory of subjective choiceworthiness under decision-theoretic uncertainty? The answer is that the theory would give us an understanding of how to evaluate options. That is, the theory would make claims about subjective choiceworthiness. These claims would differ from those of decision theories because decision theories take into account empirical risk, while this theory would take into account a different form of risk: decision-theoretic risk. Decision-theoretic risk is the risk we face in virtue of our uncertainty about which decision theory is true. If, for example, we endorse a decision theory that says we should maximise only the maximum possible value, and it turns out that this is not the best way for us to make decisions, then our decisions will be worse. It is beyond the scope of this chapter to discuss in depth what a plausible *meta-level decision theory* would claim, although I discuss an interpretation that fails in §2.7.

How can we represent a meta-level theory mathematically? This is surprisingly simple. A meta-level theory tells us how to evaluate our options given a distribution of subjective-choiceworthiness claims made by decision theories.²³ So it may be represented as a function mapping from a probability distribution over the real-valued subjective-choiceworthiness claims to the set of real numbers. So that this gives a value for each act, we say that the function takes for its argument a probability distribution over a real-valued $|\mathcal{A}|$ -vector and that the output of the function is also a real-valued $|\mathcal{A}|$ -vector.

Although this is all you need to set up a model of decision-theoretic uncertainty, there is an obvious way to generalise such a model. Namely, if we allow an agent to be uncertain about which decision theory to accept, we may also allow them to be uncertain about which *meta-level decision theory* to accept.

This gives us another shift, parallel to the shift from decision theories to meta-level decision theories. Normally, decision theories *assign* subjective-choiceworthiness values, but under decision-theoretic uncertainty, decision theories only make *claims* about subjective choiceworthiness. Under meta–decision-theoretic certainty, meta-level decision theories *assign* subjective choiceworthiness, but under meta–decision-theoretic uncertainty they only make *claims*

²² Note that Trammell draws inspiration from Mertens and Zamir (1985) and Lipman (1991), amongst others.

²³ The claims derive from the decision theories themselves and from an agent's probability function representing their rational beliefs about which decision theories might be true.

about subjective choiceworthiness. When we were dealing with a distribution over subjective-choiceworthiness value claims made by decision theories, a meta-level decision theory assigned subjective-choiceworthiness values that took uncertainty into account. Now we have a distribution over subjective-choiceworthiness claims made by meta-level decision theories, and we want a theory that tells us how subjectively choiceworthy agents' acts are under these conditions. We might call such a theory a *meta-meta-level decision theory*.

Note the formal similarities between a *meta-meta-level* decision theory, a *meta-level* decision theory, and a regular decision theory. Each is most naturally represented by a function mapping from a probability distribution over value claims, with one distribution for each act, to a value claim for each act. When you see clearly the shift from decision theories to *meta-level* decision theories to *meta-meta-level* decision theories, it is apparent that this is part of a pattern. We could talk about *meta-meta-meta-level* decision theories in the same way. We could also extend this to even-higher-order decision theories. A natural hierarchy emerges.

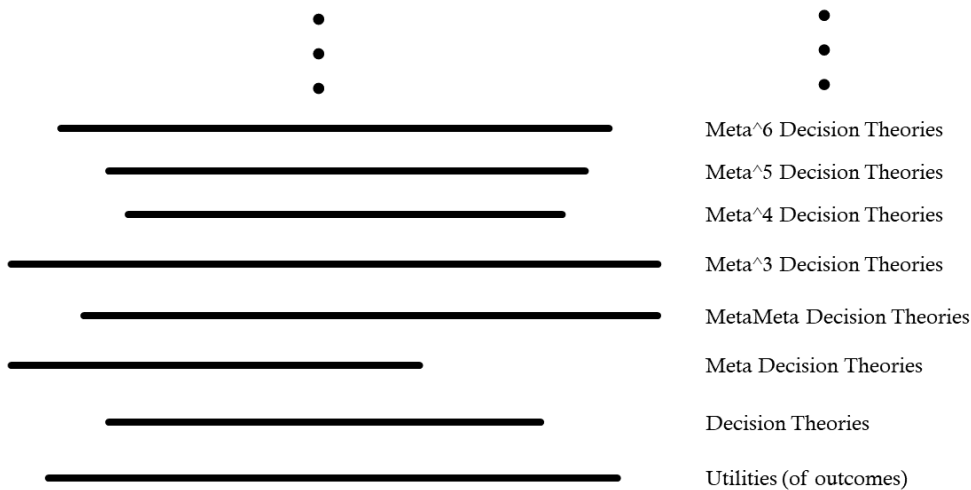


Figure 2.1: A Natural Hierarchy of Decision Theories (with value ranges on the x-axis)

The rest of this section formalises this idea.

Let us quickly do some housekeeping. Talk of *meta-meta-meta-level* decision theories is clunky, and it becomes clunkier for higher-order decision theories. There is perhaps no upper bound on how clunky it might get. We can simplify the terminology by calling decision theories *1-theories*. *Meta-level* decision theories thus become *2-theories*, and so on.

Let us also briefly turn to an agent's beliefs before defining the hierarchy. We can say that the set of all relevant claims is \mathcal{S} , which includes the set of possible states, the set of possible decision theories, and the many sets of possible higher-order decision theories.

Definition 2.2.4. The set of all relevant claims, \mathcal{S} , is the Cartesian product $\mathcal{S} \times \{1\text{-theories}\} \times \{2\text{-theories}\} \times \{3\text{-theories}\} \times \dots$

Definition 2.2.5. The agent's initial beliefs, $P(\mathcal{S})$, specify a probability distribution over \mathcal{S} .

Thus we assume that an agent has a set of rational beliefs about all of the relevant claims. This is a rather strong assumption: not only must the agent's beliefs be represented by a probability function, but this probability function is almost gratuitously complex, as it is defined over potentially infinitely many collections of things. This is not as bad as it appears. Some views of how subjective choiceworthiness relates to this hierarchy maintain that an agent is uncertain only about a finite number of sets of meta-level decision theories. At least some of these views are motivated by the idea that we mortal humans cannot reason about an infinite number of things.²⁴ Thus, the probability space is infinite-dimensional if and only if that is what we require for our view of the relationship between decision-theoretic hierarchies and subjective choiceworthiness. In specifying beliefs, I am being permissive about these views, but a restricted view can be easily captured by my model.²⁵

Back to the hierarchy. At each level of the hierarchy, $(k+1)$ -theories take as their input a distribution over subjective-choiceworthiness claims made by k -theories. We can call subjective-choiceworthiness claims made by k -theories *k-choiceworthiness claims*. We can now define the hierarchy recursively with the concepts of *k-choiceworthiness*, *k-choiceworthiness distributions*, and *k-theories*.²⁶

Definition 2.2.6. The *k-choiceworthiness* of an act a , $v_k(a)$, is what an act's subjective choiceworthiness would be if that agent accepted the true k -theory but remained the same in all other respects.

Specifically, the relevant respects are that the agent has the same utility function u over the same set of available actions \mathcal{A} and the same set of states \mathcal{S} , and their beliefs about $(k-n)$ -theories are the

²⁴ See, for instance, Zimmerman (2008) and chapter 4.

²⁵ One could also extend this to transfinite ordinal orders of uncertainty, as Trammell (2021) does.

²⁶ Technically, to talk about a finite set of acts of arbitrary size, we should modify definition 2.2.8 to speak of a family of functions $\{t_k^n\} : \mathbb{P}^n \rightarrow \mathbb{R}^n$, one for each $n \in \mathbb{N}$. However, it is obvious that we can take \mathcal{A} to be of size $|\mathcal{A}|$, as identical reasoning applies to any other n .

same as those entailed by $P(S)$ for all n such that $1 \leq n < k$. I speak of ‘truth’ in this definition to ensure that k -choiceworthiness is the correct target of the value claims made by k -theories.

Definition 2.2.7. A k -choiceworthiness distribution $P(v_k) \in \mathbb{P}^{|\mathcal{A}|}$ is a probability distribution over k -choiceworthiness values for some finite set of acts \mathcal{A} .

Definition 2.2.8. A k -theory is a function $t_k : \mathbb{P}^{|\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{A}|}$ representing claims about the k -choiceworthiness of the acts in \mathcal{A} given the $(k-1)$ -choiceworthiness distribution $P(v_{k-1})$.

Note that a k -choiceworthiness distribution is derived from k -theories and their probability distribution $P(t_k)$, and $P(t_k)$ comes directly from an agent’s initial beliefs $P(S)$. To make the exposition simpler, we can denote the minimum k -choiceworthiness value of act a (that is, $\min v_k(a)$) as $\underline{v}_k(a)$. Similarly, we can denote the maximum value as $\overline{v}_k(a)$ and the k -choiceworthiness range as $[\underline{v}_k(a), \overline{v}_k(a)]$. These three definitions are interdependent, and together they formally define the hierarchy.

2.2.3 Subjective Choiceworthiness and k -choiceworthiness

Example such as 2.2.1 motivate the idea of subjective choiceworthiness. Here, my goal is to capture how subjective choiceworthiness could be determined by this underlying hierarchy structure.

Recall that each k -theory makes claims about subjective choiceworthiness, which I call k -choiceworthiness claims. We have now constructed a structure in which an agent has a distribution of choiceworthiness values at level $k - 1$ and some theories at level k that make subjective-choiceworthiness claims, and this generates a distribution of choiceworthiness values at level k . And subjective choiceworthiness is an *all-things-considered* evaluation of an agent’s options: it takes into account not only their uncertainty about states but their uncertainty about theories at various levels. There are at least two views of how subjective choiceworthiness relates to the hierarchy structure we have set up.

One view is that this process will not continue forever. Instead, there is a privileged level k^* such that the agent accepts a single k^* -theory.

Definition 2.2.9. An agent *accepts a k^* -theory* when their probability distribution $P(t_{k^*})$ over theories at level k^* assigns positive probability to only one k^* -theory.

There are many different ways that an agent could accept a theory at k^* . For example, they might know the correct k^* -theory;²⁷ or they might accept a single k^* -theory for pragmatic reasons—for example, because they cannot reason at any higher level²⁸ or because the costs of further deliberation are too high.²⁹

If an agent accepts a single theory at k^* , then the subjective choiceworthiness of each act, $v(a)$, will be well defined. Subjective choiceworthiness will simply be k^* -choiceworthiness. On this view, subjective choiceworthiness is the kind of choiceworthiness that takes into account all of an agent's decision-theoretic uncertainty (which is finite) in the appropriate way (which they accept). Subjective choiceworthiness is truly the kind of choiceworthiness that takes into account all of the relevant claims, and we can redefine S as the *finite* Cartesian product $S \times \{\text{decision theories}\} \times \{2\text{-theories}\} \times \{3\text{-theories}\} \times \dots \times \{k^*\text{-theories}\}$.

This is not the only view. On another view, there might not be a privileged k^* such that an agent accepts a single k^* -theory. In this case, there might be unboundedly many possible subjective-choiceworthiness value claims made by unboundedly many orders of k -theories. Instead of thinking about these values individually, we might think of the range of possible values given by the choiceworthiness distribution at each k —that is, the sequence of k -choiceworthiness ranges $\left\{ \left[\underline{v}_k(a), \overline{v}_k(a) \right] \mid k \in \mathbb{N} \right\}$. If a number falls outside the range of possible k -choiceworthiness values for any k , then the agent is certain this is not the subjective-choiceworthiness value of a .³⁰ Thus, subjective choiceworthiness must lie at the intersection of these choiceworthiness values: $v \in \bigcap_{k=1}^{\infty} \left[\underline{v}_k(a), \overline{v}_k(a) \right]$. If, for some act, there is a unique value that lies at the intersection of all possible choiceworthiness ranges, then the agent is certain of the value of that act. We can say that *unique evaluation* occurs when this is true for every act.

Definition 2.2.10. *Unique evaluation* occurs when there is a single real number $v(a) \in \mathbb{R}$ such that $v(a) = \bigcap_{k=1}^{\infty} \left[\underline{v}_k(a), \overline{v}_k(a) \right]$ for each act $a \in \mathcal{A}$.

²⁷ I take it that this is Tarsney's (n.d.) preferred view. However, while he argues that we know the correct decision theory, I only need to have an agent accept a single theory for some $k^* \in \mathbb{N}$.

²⁸ This is the view of Zimmerman (2008).

²⁹ This is my own view. See chapter 4.

³⁰ Trammell (2021) derives this from what he calls the dominance principle. I give the intuition here, and formally I state it in definition 2.2.10. I do not feel the need to justify it with more basic principles because this chapter is not primarily about the regress problem and here I am a pluralist about how to solve it. Note that it could be justified by an extension of the naturally bounded principle (definition 2.5.2).

In a sense, unique evaluation occurs when an agent has certainty across all levels, regardless of their uncertainty at any particular level. At every level k , the agent is uncertain, but they know subjective choiceworthiness cannot fall outside of their possible value range at that k . They thus rule out some possible values for each k . Likewise, when they take their uncertainty in every order of the hierarchy together, they are able to rule out every possibility except one: the unique evaluation. Thus, on the second view, the all-things-considered subjective-choiceworthiness value of an act is what emerges when the agent considers all of their uncertainty and realises there is but one possible value. The necessary conditions for unique evaluation to occur have been explored by Trammell (2021), who outlines three principles that guarantee unique evaluation. In §2.7 I will discuss one of these principles further. Other authors have expressed hope that unique evaluation occurs but have not attempted to specify the conditions that guarantee its occurrence.³¹

So either subjective choiceworthiness is an accepted claim about k^* -choiceworthiness, or it is the unique value across all k -choiceworthiness distributions. For my theorems, I am agnostic about which of these views to take; where not otherwise specified, you can assume either is true. My framework, however, requires at least one of them to be true so that v is well defined.³²

2.2.4 Information: Messages and Signals

I now turn to information, which I will formalise in terms of messages. An agent receives a message, and upon receiving this message, their beliefs change. This message is information. And there is a set of possible messages, called a signal, that an agent could receive.

Recall example 2.1.1. Artemis had to choose whether to try drug X and had a few information sources available. Artemis could have received empirical information about their genes; I named this type of information *pure state*. Artemis could also have attended a class on decision theory that would give them what I named *pure-normative* information. We also saw that for either type of information, it would be *perfect* if it resolved all of the decision-relevant uncertainty, and *imperfect* otherwise. Importantly, information proper is supposed to capture all of the possible intersections of these types of information.³³

³¹ Sepielli (2017) and Tarsney (2017) hope for ‘convergence’, while Trammell (2021) illustrates that what I call unique evaluation is the more general version of convergence in this context.

³² Information value when subjective choiceworthiness is not well defined may well be incoherent, and this case is beyond the scope of this chapter.

³³ I gave examples of these intersections in table 2.1.

Our definition should capture information in this general sense. To achieve this, a message is defined here as a subset of the set of all relevant claims. Recall that we constructed the set of relevant claims as the Cartesian product of the relevant claims about states and theories at every k , and we denoted it \mathcal{S} .

Definition 2.2.11. A *message*, $m \in \mathcal{M}$, is a subset of \mathcal{S} .

If we want to be more specific, we can say that a pure-state message, $m_{\mathcal{S}}$, is a subset of \mathcal{S} while a pure-normative message, $m_{\mathcal{N}}$, is a subset of the Cartesian product of k -theories for all k .

Importantly, an agent does not know what a message says before receiving it; otherwise it wouldn't count as informative. Instead, they know what a collection of possible messages might say, and they know their beliefs about which messages are more or less likely. Thus, we also need to define a set (or collection) of possible messages.

Consider what would happen if some region of \mathcal{S} existed such that no message can induce positive probability for any claims in any part of that region. The agent would know that their beliefs after receiving m will not assign positive probability to claims in that region, no matter the message, and that agent would assign any relevant claims probability 0 prior to receiving the message.³⁴ Thus the signal must cover, without gaps, all of the claims to which an agent assigns positive probability. If there are gaps, the agent cannot assign these claims positive probability.

An elegant mathematical formalism that ensures this condition is met is to treat the collection of messages as a *partition* over \mathcal{S} . This guarantees that messages are mutually exclusive and collectively exhaustive of \mathcal{S} : for every claim to which an agent initially assigns positive probability, there is a message they could receive that would leave them with a positive belief in that claim.

Definition 2.2.12. A *signal*, $\mathcal{M} = \{m_1, \dots, m_{|\mathcal{M}|}\}$, is a partition over \mathcal{S} .

A message is general because the notion of a subset is general. It could be a maximally specific subset that specifies a single state and a single theory for each k . If a message is a maximally specific subset, it will give an agent perfect information. A subset may also be less specific, giving an agent imperfect information. There are many ways for a subset to contain more than a single possibility for at least one relevant claim, and so information can be imperfect in many different ways. There is only one way for information to be perfect, but many ways for it to be less than maximally specific.

We also want to capture cases in which information is imperfect with respect to a single k -theory—for example, when an agent finds

³⁴ So long as they update via conditionalisation and are aware of this fact.

out that Expected Utility is less likely than they thought but they don't definitively rule it out. This can be formalised somewhat awkwardly by splitting some k -theories into two k -theories that each make the same value claims as the original and then assigning them jointly the probability assigned to the original k -theory. Then an agent can find out that only one of these is false and update accordingly.

Exactly how S is partitioned will give us different kinds of information. For example, each m could specify exactly one claim about the state of the world and one k -theory for each k , in which case m would always give perfect information. Alternatively, some m could allow for multiple possibilities, or every m could. Thus there are many possible structures made possible by the notion of a partition.

So far we have defined information as a subset of S and the collection of possible pieces of information as a partition over S . We still have some things to clarify.

If information is valuable, this is in virtue of its ability to change our minds. For any possible message, an agent will know their conditional probability distribution upon receiving that message. Given an agent's initial beliefs $P(S)$,³⁵ we can calculate the probability distribution they will have upon receiving m by assigning probability 0 to the relevant claims not contained in m and assigning probability to the remaining claims according to conditionalisation. I will denote this $P(S|m)$. Given an agent's initial beliefs, we can also calculate their probability of receiving a particular message m as $p(m|S)$. I will denote this $p(m)$. Let $P(\mathcal{M})$ denote the distribution that gives us $p(m)$'s, one for each $m \in \mathcal{M}$.³⁶

2.2.5 Information Value

Upon receiving a message, an agent's beliefs will change. This may change their evaluations of acts via changing the probability of various outcomes, via changing the way risk is subjectively evaluated, or both. Indeed, subjective values are belief dependent. We can thus specify that v denotes subjective choiceworthiness relative to the agent's initial beliefs $P(S)$. v_m , in contrast, denotes subjective choiceworthiness relative to their beliefs conditioned on m , $P(S|m)$. Information can change their probability distribution over states and their probability distribution over theories at every level; so this change is both about the likelihoods of different outcomes given their acts and how they evaluate their risky acts.

³⁵ Recall that this denotes the agent's initial probability distribution over S .

³⁶ Note that this is not the probability of receiving any message m , which is 1.

Earlier I said that the value of information is whatever a rational agent should be willing to sacrifice in order to obtain the information, which I will denote with the symbol ψ . That is, without knowing which message m they will receive, they are indifferent between receiving the information at a price of ψ and not receiving the information, where *price* is defined broadly. Economists might call this the *demand price of information*. Now that we have made the idea of subjective choiceworthiness explicit, we can make this idea more precise.

We can formalise the cost of obtaining information with the aid of the notation $\mathcal{A}_{(-\psi)}$, which denotes the set of acts available to the agent in the nearest possible world in which they pay ψ for m . This set is equal to \mathcal{A} if $\psi = 0$. Acts in \mathcal{A} may correspond to only slightly modified versions in $\mathcal{A}_{(-\psi)}$ —for example, if I pay \$1 for information about whether it will rain or shine and then wear my raincoat as I would have done without the information. Or they may have no corresponding act, as when ψ is high enough to render that act impossible; for example, if I spend all of my money on information about which house is best, I will not have an action that is just like ‘buying a house without information available to me’ available. In cases of negative ψ , the set $\mathcal{A}_{(-\psi)}$ may be more expansive than \mathcal{A} .

The value of information is whatever price ψ makes accepting the information at that price exactly as subjectively choiceworthy, as captured in v , as rejecting that information.

Definition 2.2.13. The *value of information*, ψ , is the solution to the following equation:

$$v \left(\max_{a \in \mathcal{A}_{(-\psi)}} v_m(a) \right) = \max_{a \in \mathcal{A}} v(a)$$

First consider the left-hand side. Overall, the agent pays ψ for information, then, upon receiving that information, chooses their best act. They choose from the acts available after paying ψ , $\mathcal{A}_{(-\psi)}$, given their evaluations of their acts conditional on m , v_m . Outside of the brackets, we realise that they do not know in advance which information they will receive (if they did, it wouldn’t be information!). Given their probability distribution over possible messages, and their knowledge of how they will act optimally upon receiving some information or other, they get a distribution over subjective-choiceworthiness evaluations of their own optimal acts given each possible message (we do not need to take into account varying values of ψ at this stage, because that varies at the level of the

whole equation).³⁷ I propose in this equation that they evaluate that distribution over possible subjective-choiceworthiness values according to their subjective-choiceworthiness function v prior to receiving any information.³⁸ Thus, the left-hand side as a whole is the value of receiving information at the cost of ψ , from the perspective of the agent before they know what that information says but when they do have a fully specified distribution over what information might say and with knowledge about how they will evaluate their acts conditional on m (and thus knowledge about how they will act after receiving m).

On the right-hand side of this equality we have the value of the best act available, as captured in v , without information. This is the value of acting without information.

ψ is supposed to capture the value of information—that is, the most a rational agent would be willing to pay. We set ψ to whichever number ensures the right-hand side and left-hand side of the equation are equal. In other words, ψ is the price of information that would make the agent indifferent between paying for information and then acting (on the one hand) and acting without it (on the other).

Importantly, the actual price of information may be different from ψ . When what we must sacrifice to obtain information is worth less to us than ψ , the information is more valuable than its actual price, and we should obtain it. When we must sacrifice something worth more than ψ to obtain the information, the cost is too great and we should not obtain it.

This model is quite general, and it is worth noting that it coincides exactly with ordinary models of information value when information is pure state and the agent evaluates their options by their expectation of value (that is, $v(a) = \mathbb{E}_s u(a|s)$).

In this section I have developed a model of decision-theoretic uncertainty in which there is a hierarchy of k -theories which all make claims about k -choiceworthiness and about which an agent has well-specified beliefs. I built these notions from an initial discussion of objective and subjective choiceworthiness. In this hierarchy model, I was able to define information as messages which form a partition (or signal) over the set of relevant claims, and I formulated an equation that formalises the value of information.

³⁷ So we should really write $v\left(\mathbb{P}\left(\max_{a \in \mathcal{A}_{(-\psi)}} v_m(a)\right)\right) = \max_{a \in \mathcal{A}} v(a)$.

³⁸ This is a little awkward given that we originally defined v over objective-choiceworthiness distributions. Using v here makes intuitive sense because v captures the agent's *all-things-considered* evaluations of distributions of value when the distributions are not conditional on the agent's acts. In this case, although their acts do affect value, the distribution inside the big brackets is in a sense independent of their acts because they know how they will act given m .

2.3 BACKGROUND

2.3.1 Normative Decision Theory

Here I will survey some of the most relevant literature on normative decision theory. The role of decision theory here is to serve as a minimal account of practical rationality. When we treat issues of rational belief and rational desire separately, the role of formal decision theory is to tell us which risk attitudes are rational and which are not. Exactly what these risk attitudes are will become apparent by the end of this section. As I noted in §2.1, placing this section after the model may seem odd, but it allows us to use the notation set up so far to understand the literature.

The most famous decision theory is *Expected Utility*. Expected Utility (EU) does what it says on the tin: it assigns values which are expected utilities. According to EU, subjective choiceworthiness is the weighted-average objective choiceworthiness. EU has a rich history and is often considered the default view in decision theory.³⁹

Definition 2.3.1. According to *Expected Utility*, the value of act a is its expected utility, $\sum_{i=1}^{|\mathcal{S}|} p(s) \cdot u(a|s)$.

According to Expected Utility, each outcome is weighted according to its probability. However, sometimes we care about the best possible outcome or the worst possible outcome a bit more. In other words, our risk attitudes may be different.

To illustrate this, first consider the extreme cases. First, an agent might care *only* about the best possible outcome, thus seeking to maximise their maximum possible value. Call this *Maximax*. Second, an agent may care *only* about maximising the minimum possible value of the worst possible outcome. Call this *Maximin*.

Definition 2.3.2. According to *Maximax*, the value of act a is its maximum possible value—that is, $\max_{s \in \mathcal{S}} u(a|s)$.

Definition 2.3.3. According to *Maximin*, the value of act a is its minimum possible value, $\min_{s \in \mathcal{S}} u(a|s)$.

Because these are claims about the subjective value of acts given uncertainty in objective value, they are decision theories (per §2.2). Note that these theories entail extreme risk attitudes. Maximax

³⁹ The earliest publication on the topic was Bernoulli's 1738 paper, eventually reprinted in *Econometrica* (D. Bernoulli, 1954). However, it was also discussed by Gabriel Cramer in correspondence a decade earlier; see J. Bernoulli (1975). The most famous axiomatisations are those of Von Neumann and Morgenstern (1944), Savage (1954), and Jeffrey (1965). Ramsey (1926) proposes a similar approach to Savage's; see Steele and Stefánsson (2020, §3).

entails extreme risk seeking, as an agent that accepted this theory would accept any risk for a tiny increase in the maximum possible value they could obtain. Maximin, by contrast, entails extreme risk avoidance: an agent that accepted this theory would forgo any increase in maximum value, or even average value, for a tiny improvement in the worst-case scenario.

The three theories introduced so far correspond to the extremes of *neutrality*, *recklessness*, and *timidity*. A Maximax agent is as reckless as can be, a Maximin agent is as timid as can be, and an Expected Utility agent is a neutral weigher of outcomes by their odds. The extremes of recklessness and timidity might not seem normatively compelling, but perhaps some shift from perfect neutrality is. Consider the following motivating case.⁴⁰

Example 2.3.1 (The Enormously Powerful Alien God from Before Is Giving Out Free Coffee). Bjørn made the correct choice earlier and now lives in one of the best of all possible worlds. The alien god appears before another person, Gjorn. Producing a magical coin from their pocket, the god explains the rules. They will flip the coin twice and give prizes accordingly. If the first flip lands heads, Gjorn will receive a free coffee next Tuesday. If the second lands tails, it's a free coffee on Wednesday. These prizes are equally and independently valuable to Gjorn, and these are the only available prizes. For just a few cents, the alien god is willing to only flip the first coin, which will give Gjorn coffee on Tuesday for heads and coffee on Wednesday for tails. Gjorn pays a few cents for this second option because it guarantees a free coffee next week.

In this example, $EU(\text{option 1}) = 0.5 \cdot u(\text{coffee on Wednesday}) + 0.5 \cdot u(\text{coffee on Tuesday})$, while $EU(\text{option 2}) = 0.5 \cdot u(\text{coffee on Wednesday} - \text{a few cents}) + 0.5 \cdot u(\text{coffee on Tuesday} - \text{a few cents})$. So the expected value of the second option is lower.⁴¹ Gjorn's preference for the second option cannot be captured by taking the expectation of a utility function, so if we want to consider Gjorn's choice to be rational, we must allow departures from perfect risk neutrality. Indeed, Buchak (2013) famously argues that we should depart from EU and allow a spectrum of possible risk attitudes.⁴² Her

⁴⁰ This case is adapted from Buchak (2013, example 1.1.1)

⁴¹ Technically, it's only lower *if* you take losing money to be worse than gaining money in ordinary circumstances, which is hardly a concession.

⁴² This theory (Risk-Weighted Expected Utility) was proposed by Quiggin (1982), developed by Yaari (1987), and presented as a normative theory by Buchak (2013); see also Buchak (2017a). For further discussion treating this theory as normative,

decision theory, Risk-Weighted Expected Utility (REU), can capture a spectrum of risk attitudes, including the extremes of neutrality, recklessness, and timidity.

I now present REU formally. First, we must reindex states such that those on a lower index have lower objective values.⁴³ Because objective values are relative to both acts and states, we must reindex again for each act we evaluate. Next, let there be a risk-weighting function, r . This is a function mapping from the closed unit interval to the closed unit interval⁴⁴ such that r satisfies $r(0) = 0$ and $r(1) = 1$. Following Buchak (2013), we can interpret r as a measure of how much more an agent is concerned about the worst outcome than the best.⁴⁵

Definition 2.3.4. Let r denote a particular risk-weighting function. According to *Risk-Weighted Expected Utility on r* , REU_r , the value of an act a is its risk-weighted expected utility,

$$\sum_{j=1}^{|\mathcal{S}|} \left[r \left(\sum_{i=j}^{|\mathcal{S}|} p(s_i) \right) (u(a|s_j) - u(a|s_{j-1})) \right]$$

There is a lot going on in this formula, so let me explain. First consider $j=1$, and note that we are summing across j . The first term of the sum is

$$r \left(\sum_{i=1}^{|\mathcal{S}|} p(s_i) \right) u(a|s_1). \quad (2.1)$$

As the probability over all states is 1 and by stipulation $r(1) = 1$, we can simplify further to $r(1)u(a|s_1) = u(a|s_1)$. So the first term of the outer sum is just the minimum utility value that an agent can get. It's best to think of this as the utility they are guaranteed to get if the agent acts according to a .

see Briggs (2015), Buchak (2015, 2017b), Joyce (2017), Pettigrew (2015), and Thoma and Weisberg (2017).

43 That is, $u(a|s_j) \geq u(a|s_{j-1})$ for all $s_j, s_{j-1} \in \mathcal{S}$.

44 That is, $r : [0, 1] \rightarrow [0, 1]$.

45 Indeed, part of Buchak's (2013) contribution to normative decision theory was reinterpreting r as a measure of something other than belief and desire that contributes to value, rather than a measure of an agent's irrational beliefs or attitudes about probabilities.

The next term is $j=2$:

$$r \left(\sum_{i=2}^{|\mathcal{S}|} p(s_i) \right) (u(a|s_2) - u(a|s_1)). \quad (2.2)$$

On the right-hand side we have the difference in utility gained by achieving the second-worst state rather than the worst state. We multiply this by the risk-weighted probability that a will lead to that much utility. Overall, this term is the risk-weighted probability of receiving a gain over the worst outcome multiplied by the magnitude of that gain.

Every other term of the outer sum is similar: there is a potential gain relative to the utility we have already weighed, and each possible gain is weighted by the risk-weighted probability that it will occur.

Taking the outer sum as a whole, this means that the value of an act is the value it produces for sure plus each of the possible gains above this minimum multiplied by the risk-weighted probability of that gain. To make this clear, we could have written the whole thing as follows:

$$u(a|s_1) + \sum_{j=2}^{|\mathcal{S}|} \left[r \left(\sum_{i=j}^{|\mathcal{S}|} p(s_i) \right) (u(a|s_j) - u(a|s_{j-1})) \right] \quad (2.3)$$

The key development of REU is that it allows for more complex risk attitudes. r is crucial. If $r(p) = p$, then REU is just the same as EU. If $r(1) = 1$ and $r(p \neq 1) = 0$, then REU is equivalent to Maximin; if $r(p > 0) = 1$, it is equivalent to Maximax. So timidity, neutrality, and recklessness are all special cases of REU, and they can be modelled with particular risk-weighting functions. Further, there are infinitely many functions r which will result in various attitudes towards risk. The flexibility of REU in modelling different risk attitudes is impressive. It can even model several other families of decision theories.⁴⁶ REU represents a crucial development for normative decision theory: not only is it the first proposal for a (normative) departure from EU, but it is also impressive in its own right for its ability to model a wide variety of preferences. However, even more recent work indicates that perhaps even more powerful tools are needed.

⁴⁶ Wakker (1990) and Nakamura (1995) find that REU is roughly equivalent to Choquet Expected Utility. REU can also model select cases of prospect theory; see Buchak (2013, p.59, 66) and Tversky and Kahneman (1992).

According to REU, our rational responses to risk can be more or less risk averse, but they must be diachronically stable: an agent's rational risk attitudes should not change in different situations. However, I think we have good reason to believe that in different situations we should be more or less risk averse. Consider the following example.

Example 2.3.2 (Two Gambles). Suppose Colel buys a lottery ticket that costs \$15 and gives a non-negligible chance of winning \$1 million. Eloy sells her this lottery ticket and keeps all the other tickets. Eloy has a family home worth almost \$1 million, and everything else they own could be sold to make up the difference. Thus, Eloy gains \$15 but may lose everything, while Colel pays \$15 and may win \$1 million. Suppose, for argument's sake, that for both Colel and Eloy, the expected utility of these gambles is exactly \$15.

My intuition is that someone could hold rational risk attitudes such that if they were in Colel's shoes they would be happy to buy the lottery ticket, and if they were in Eloy's shoes, they would not sell an equivalent lottery ticket (or at least, only at a *much* higher price). This intuition says that it is more reasonable to be risk seeking when the stakes are intuitively less significant. I'd be happy to take high expected-value bets that cost me little but unhappy if someone offered structurally equivalent bets about the entire planet.

Indeed, recent work by Bottomley and Williamson (n.d.) argues that it is rationally permissible for agents to be more risk averse when the stakes are intuitively higher and less risk averse when they are intuitively lower. They argue for maximising *weighted linear utility*.⁴⁷ Formally, we require an outcome-weighting function—that is, a function that assigns each outcome a strictly positive real number.⁴⁸

⁴⁷ Drawing on the work of Chew (1983, 1989).

⁴⁸ That is, $w : (a, s) \rightarrow \mathbb{R}^{++}$. The example given by Bottomley and Williamson (n.d.) helps give a feel for what w might look like. Suppose

$$w(s_j) = \frac{1}{1 + \sqrt{u(a|s_j)}}$$

. Where utility has diminishing marginal utility in money, $u(\$x) = \sqrt{x}$. Note that w must be redefined for each a . They write of this example that '*intuitively, an agent with this weight function places greater significance on worse outcomes compared to better ones—rags loom large in comparison to riches*' (p.19).

Next we can define an outcome's *relative weight* given w , $q_w(s_j)$, as

$$q_w(s_j) = \frac{w(a, s_j) \cdot p(s_j)}{\sum_{l=1}^{|\mathcal{S}|} w(a, s_l) \cdot p(s_l)}. \quad (2.4)$$

Following Bottomley and Williamson (n.d.), we can interpret $q_w(s_j)$ as a measure of how much one is concerned with securing some outcomes rather than others. Thus we divide the weighting of the probability of some state by the total weightings of probabilities across all possible states for that act, making $q_w(s_j)$ a measure of *relative concern*.

Definition 2.3.5. Let w denote a particular outcome-weighting function. According to *Weighted Linear Utility on w* , WLU_w , the value of act a is its weighted linear utility, $\sum_{j=1}^{|\mathcal{S}|} q_w(s_j) \cdot u(a|s_j)$.

For each act, we take the expected concern-weighted utility of that act.⁴⁹ This allows an agent's risk aversion to vary with the relative stakes of a situation.

Now that we have seen some of the possibilities in normative decision theory, I would like to clarify a way in which my own position differs from Buchak, Bottomley, and Williamson's. For Buchak (2013), many different risk-weighting functions are rational. I want there to be a single best risk-weighting function.⁵⁰ Indeed, this is why I defined a decision theory to be a single function mapping onto the set of real numbers. I am here in agreement with the early decision theorists⁵¹ and Trammell (2021) The main theoretical virtue of this approach is that it gives more clarity to the project of decision-making under decision-theoretic uncertainty. There are many available decision-theoretic options, and we are uncertain which is best, yet we must decide without resolving this uncertainty.

So some authors say many t 's are permissible, but I say only one is. The difference seems bigger than it is. To me, ' a is irrational'

⁴⁹ Expanding q_w would give

$$WLU_w(a) = \sum_{j=1}^{|\mathcal{S}|} \left(\frac{w(a, s_j) \cdot p(s_j)}{\sum_{l=1}^{|\mathcal{S}|} w(a, s_l) \cdot p(s_l)} \right) u(a|s_j)$$

⁵⁰ This may be confusing for those that are more familiar with representation-theorem first approaches. For me, $r = p$ might count as a single decision theory, even though in an alternative approach it would only specify a family of positive affine transformations. The specificity in my account comes from the single objective-choiceworthiness distribution that all decision theories work with. This may become clearer in §2.4.

⁵¹ Such as D. Bernoulli (1954), Von Neumann and Morgenstern (1944), and Savage (1954).

means ‘*a* is not one of the best available acts’. This might not be a strong criticism—for example, when *a* is the second-best act and differs only a little in value from the best available act. I take it that Buchak, Bottomley, and Williamson mean something different to me when they speak of irrationality and rationally permissible actions. Their usage of those terms, like the colloquial usage, is much more critical of people deemed irrational. They are permissive because they don’t want to make strong criticisms of the people that I would call *irrational but almost rational*. Like them, I don’t want to make strong criticisms of these people, but unlike them, I avoid these strong criticisms through the particular way in which I allow rationality to be a matter of degree.⁵²

2.3.2 *How Normative Decision Theory Relates to Information Value*

In this section, I will show how information value relates to decision theory and motivate my project of finding out about information value under decision-theoretic uncertainty.

In decision theory, a representation theorem says that an agent can be represented as if they are following a particular decision theory if and only if their choices satisfy certain axioms. Von Neumann and Morgenstern (1944) and Savage (1954) give axiomatisations of EU. Expected-utility maximisers satisfy the *independence* condition or, respectively, the *sure-thing principle*. Independence is the condition that adding an equal chance of a particular outcome to any set of acts will not change the preference ordering of those acts to the agent.⁵³ In a sufficiently restricted context, the sure-thing principle is equivalent to independence (Friedman and Savage, 1952). Wakker (1988) illustrates how violating independence necessarily means rejecting free (pure state) information.⁵⁴ In fact, accepting the

⁵² To be clear, they allow rationality to be a matter of degree in a different sense.

⁵³ Formally, an agent satisfies the independence condition if for any three outcomes *A*, *B*, and *C* the following condition holds: if $A \succ B$, then $(A, p; C, 1 - p) \succ (B, p; C, 1 - p)$ for any $p \in [0, 1]$, where $A \succ B$ denotes ‘*A* is strictly preferred to *B*’ and $(A, p; C, 1 - p)$ denotes a p chance of outcome *A* and $(1 - p)$ chance of outcome *C*. For EU theorists, $A \succ B$ is a special case of $(A, p; C, 1 - p) \succ (B, p; C, 1 - p)$ —specifically, the case in which $p = 1$.

⁵⁴ To see this, imagine the independence-violating preferences in which $B \succ A$ yet $(p, A; 1 - p, C) \succ (p, B; 1 - p, C)$. Then imagine the following decision. There are two states of the world such that $p(s_1) = p$ and $p(s_2) = 1 - p$. The player can choose between two acts, the first of which will yield *A* if s_1 obtains and *C* otherwise, and the second will yield *B* if s_1 obtains and *C* otherwise. Notice this choice is initially equivalent to $(p, A; 1 - p, C)$ versus $(p, B; 1 - p, C)$. Note that in this scenario, they will choose $(p, A; 1 - p, C)$ given their stated preferences. Now, imagine that the agent can gain information before choosing, such that they know with certainty which state occurs. In this case, either they will face the choice between *A* and *B*, or they will get *C*. They also know they will choose

independence axiom means always accepting costless information; perhaps surprisingly, EU is the only decision theory that can claim that pure-state information is valuable everywhere.⁵⁵ I call this result the Blackwell-Good-Wakker theorem after those that proved it.⁵⁶

Theorem 2.3.1 (Blackwell-Good-Wakker theorem). Pure-state information has weakly positive value \Leftrightarrow subjective choiceworthiness of an act is its expected utility, $v(a) = \mathbb{E}u(a|s)$, for all $a \in \mathcal{A}$.

According to this theorem, agents will always accept costless pure-state information if and only if they maximise expected utility. Applying this to the hierarchy framework, an agent might maximise expected utility because they accept expected utility at level 1 or because their uncertainty resolves itself such that they act as if they maximise expected value at level 1 (in §2.7 I will discuss cases like this).

Some have proposed that pure-state information must always be valuable; otherwise the agent must be irrational.⁵⁷ The reasoning is that having more information cannot make you worse off. The worst-case scenario is that this does not change your action, and the best-case scenario is that you pick a new, better action.

B, given the preferences above, if they know s_1 will occur. They also know that the probability of s_1 is p . So gaining information is itself an act equivalent to $(p, B; 1 - p, C)$. When given the choice of whether or not to gain information, they will reject information, even when free, given their preference $(p, A; 1 - p, C) \succ (p, B; 1 - p, C)$. So violating independence leads to rejecting costless pure-state information in at least some cases.

- 55 This is true given our earlier assumption that an agent has precise-enough beliefs that you could represent them with a probability distribution. If beliefs are imprecise, then even agents who follow Expected Utility may reject free information. See S. Bradley and Steele (2016) and Kadane, Schervish, and Seidenfeld (2008).
- 56 (\Rightarrow) is proven by Blackwell (1953) and independently by Good (1967) The contrapositive of (\Leftarrow) is proven by Wakker (1988); specifically, any violations of Von Neumann and Morgenstern's (1944) independence axiom will also reject free information. Note that while I have given one specific representation for EU, this theorem will hold true for all positive affine transformations of EU. See also Ramsey (1990).
- 57 Demski (1980, p.37) writes that '*... costless perfect information never "hurts". Either you use it (a strict gain) or ignore it.*' Keasey (1984, p. 648) writes that '*...the acceptance of costless perfect information is a fundamental property of rational behaviour ... to reject costless perfect information ... seems self-evidently irrational.*' Carnap (1947, p. 138-9) writes that '*a principle which seems generally recognized ... says that ... we have to take as evidence the total evidence available to the person in question at the time in question*'. Indeed this idea goes back to the earliest decision theory. Keynes (1921, p. 313) refers to '*Bernoulli's [1713, p. 215] maxim that in reckoning a probability, we must take into account all the information which we have*'. Even A. J. Ayer thought this was one of the two key desiderata for a successful account of science; see Miller (1994). More recently, S. Bradley and Steele (2016, p. 2) write that '*... rationality surely requires that an agent not pay to avoid free evidence. ...*'.

Decision theories that depart from EU must deny that an agent should always accept costless information. This does not present a problem for those who are just trying to model human behaviour; so what if people are irrational! But it may provide an argument against *normative* decision theories that depart from EU.⁵⁸

The non-EU theorist can respond by showing that their theory captures the compelling cases that give normative force to the claim that we should accept costless pure-state information: a theory may get it right in all the important cases without obeying some particular generalisation of those cases.

Indeed, Buchak (2013) argues that there are examples of rejecting costless information that seem rational.⁵⁹ The idea behind these cases, I take it, is that information can sometimes make you take an *ex ante* worse action and that it is unclear that information that may have this effect should be accepted even if it's free.

MacAskill, Bykvist, and Ord (2020, §9) give the first analysis of non-empirical information. Specifically, they try to analyse the value of *moral information*, which is information about moral facts or theories.⁶⁰ In their approach, agents who are morally uncertain should maximise expected moral value (they call this *expected choiceworthiness*). They argue that moral information can be extremely valuable, especially in high-stakes decisions. They write that '*in general, because information brings about a proportional change in the expected choiceworthiness of the options under consideration, if you're dealing with extremely high-stakes issues, then the expected choiceworthiness of gaining new information becomes extremely high as well*' (p. 202). This claim is clearly true of those who accept EU. Indeed, MacAskill, Bykvist, and Ord's book may be considered an extended exploration and defence of this approach, and so the assumption is reasonable. However, for agents who are uncertain about the correct decision theory, as we have seen, information may have negative value. We can think of stakes as multipliers of value. When information value is negative, higher stakes will make information less attractive rather than more attractive.

MacAskill, Bykvist, and Ord (2020) take a step in the right direction by attempting to model information that is not just about states.

58 See, for example, Briggs (2015) for a reply to Buchak's (2013) violation of the sure-thing principle.

59 See also McClennen (1990) and Buchak (2010). Further, Wakker, Erev, and Weber (1994) point out that REU accepts a particular weakening of independence called *comonotonic independence*, and Trammell (n.d.) shows that even accepting comonotonic independence doesn't mean accepting non-misleading information.

60 In the framework of §2.2 this could best be thought of as facts about objective choiceworthiness, and thus they claim that under uncertainty about the objective choiceworthiness of outcomes, information about objective choiceworthiness could be very high.

However, they don't take things far enough. In fact, the arguments they give to motivate moral uncertainty could also motivate decision-theoretic uncertainty.⁶¹ Further work is needed, then, to understand the value of information under decision-theoretic uncertainty.⁶²

Trammell (2021) is the first to give a formal model of decision-theoretic uncertainty, and my own work draws heavily from that work. A few others have explored similar hierarchy structures in different contexts.⁶³ But no one, to my knowledge, has attempted to understand information value in such a framework.

Clearly, information value and decision theory are tightly intertwined. Previous results surrounding information value are clear advancements in our understanding of information and its value, but they leave room for further research. The Blackwell-Good-Wakker theorem (2.3.1) doesn't allow for diversity in decision-making procedures; it only tells us about the value of information that is not about decision theories, and only when we accept EU. Indeed, if rejecting EU means rejecting costless pure-state information, can we say a similar thing about rejecting pure-normative information?

2.4 HOW TO COMPARE DECISION THEORIES

This section argues that we can compare the recommendations of plausible decision theories when we do not face the issue of value uncertainty. If we could not, the results in §2.5 and §2.6 would be meaningless, so it is necessary to address this point now. However, if you are primarily interested in one of the later sections, note that although they rely somewhat on earlier sections (especially §2.2), they can largely be read independently of each other.

Representation theorems for decision theories show that when an agent makes choices that follow some plausible choice axioms (such as independence or its alternatives), their choices can be represented as if they have adopted a particular decision theory.

⁶¹ They argue that one should take moral uncertainty seriously and that moral uncertainty is analogous to empirical uncertainty. It is a natural extension to take into account our uncertainty about the norms of decision-making generally as well as those of our moral decision-making. Indeed, they write in a footnote that '*one might claim, following Lara Buchak, that one ought, in general, to endorse a form of Risk-Weighted Expected Utility. We are perfectly open to this. Our primary claim is that one should endorse maximising risk-weighted choiceworthiness if and only if Risk-Weighted Expected Utility is the correct way to accommodate empirical uncertainty. We don't wish to enter into this debate, so for clarity of exposition we assume that the risk-neutral version of Expected Utility is the correct formal framework for accommodating empirical uncertainty*' (MacAskill, Bykvist, and Ord, 2020, p. 48).

⁶² They might criticise my own approach because I do not take into account moral uncertainty. Indeed, I hope that future work can combine our approaches to give a full account of information value under normative uncertainty.

⁶³ Lipman (1991) and Mertens and Zamir (1985).

These representations are unique to a particular degree of freedom, so representation theorems show there is a class of functions that can represent an agent's decisions. That class is usually the class of all positive affine transformations, but sometimes it is more restricted. Transformations from one function to another, within this class, represent an agent's choices in the following way: if an agent accepts a decision theory, then accepting any permissible transformation of that decision theory will yield the exact same choice in any situation.

A representation theorem, then, only shows that an agent can be represented as making decisions using any decision theory in that class. With this in mind, one might wonder how we can compare decision theories. If decision theories are classes of functions, then the problem is how to determine specific functions for each decision theory that yield comparable values.⁶⁴

In this section I will argue that decision theories are comparable and show the natural way that we can choose representations of each from its equivalence class. My approach begins with the claim that each theory should assign sure outcomes the same values. Next, I will argue that if decision theories make general and coherent claims, then this will be enough to allow us to compare them. Finally, I will show that when we normalise decision theories in this way, we select the most natural representations from the equivalence class. For those who already believe we can compare decision theories, I am merely expounding a theory of exactly how to do this. But for those who object to my findings because I make an implausible claim—inherent in the model I set up in §2.2—about comparability of values, this constitutes a response. After this section, the burden of proof will be on an interlocutor to give an account of why such comparisons are impossible.⁶⁵

My setup in §2.3 is different from the representation-theory approach in a relevant respect. For me, objective-choiceworthiness values are assigned to each outcome (by the utility function). This yields a restriction on utility that is not present for most representation theorems, for which utility is constructed along with the decision theory.⁶⁶ I prefer to keep issues of belief, desire, and risk attitudes separate. This allows me to avoid problems related to belief

64 Equivalently, we might make claims about comparisons between equivalence classes of decision theories. Although I'm not aware of anyone who attempts to do this, I would be excited to see this approach explored (especially in relation to the harder problem of comparing utility functions or moral theories); see MacAskill, Cotton-Barratt, and Ord (2020) and chapter 3.

65 Note that this is different from the problem of comparing utilities across people or across moral theories, which I do not address here.

66 Note that Savage (1954) also constructs the probability measure.

comparisons and utility comparisons and to focus on comparing risk attitudes.⁶⁷

Consider the utility function. An agent knows the objective-choice-worthiness values of state-act pairs, and this provides the basis for comparisons. I claim that the objective choiceworthiness of a sure outcome should be equal across decision theories. After all, a decision-maker knows how much they value a sure outcome. A decision theory, which makes claims about subjective value, should not claim that a sure outcome has less value than its value, nor should it claim that a sure outcome has more than its value. For sure outcomes, subjective choiceworthiness and objective choiceworthiness are co-extensive, and this provides our first restriction on admissible subjective-choiceworthiness claims made by decision theories.

Perhaps surprisingly, this is enough to make decision theories comparable, so long as decision theories make coherent claims about the value of acts whose outcome is risky (exactly the kinds of claims that decision theories must make, even if utility were merely ordinal). Let's discuss what I mean by coherent claims. The kind of claim decision theories make concerns what sure outcomes a rational agent would forgo in order to obtain that risky outcome. Indeed, the value a theory assigns to an act automatically entails ordinal relations between other acts. For representation theorems, decision theories assign values to acts that merely represent deontic properties. There are many ways to represent these deontic claims, and so there is a class of admissible functions that represent a single decision theory. In order for decision theories to play their decision-guiding role, acts that are assigned lower values will be avoided in favour of acts with higher values, and an agent will be indifferent between acts of the same value. Thus, decision theories make broad comparison claims between risky-outcome and sure-outcome acts (a sure-outcome act is an act for which the agent knows the state or an act which doesn't vary in objective value across states). This means that for any value assigned to any risky act, we can consider a hypothetical sure outcome that is the same in value, according to that decision theory.⁶⁸ Any sure outcome with lower value would be given up if that risky act were available, and any available sure outcome with higher value would be chosen over that risky act. Note

⁶⁷ I discuss comparisons of utility functions in chapter 3. My view is that these comparisons are not always possible.

⁶⁸ So long as the space of outcomes is sufficiently fine-grained.

that decision theories in the literature, including all those discussed in §2.3, do in fact make these broad claims.⁶⁹

So any decision theory tells you how to compare risky outcomes to sure outcomes—that is, how to choose acts under risk. I will now defend my claim that decision theories are comparable if the values of sure outcomes are comparable and decision theories make coherent claims about the comparisons between risky- and sure-outcome acts. First, note that the values of sure outcomes are directly comparable across decision theories. Each theory gives a sure outcome its objective choiceworthiness. I'm not sure what a claim that these are not comparable would even amount to, given the assumptions we made about objective choiceworthiness. Next, note that for each theory, there is always a hypothetical sure outcome whose value is exactly that of the value of a risky action.⁷⁰ That sure outcome is comparable to the same hypothetical sure outcome's value according to another theory, and that value is clearly comparable to that other theory's assignments of value over a risky act. Thus, because comparability of values is transitive, risky acts are comparable across theories.

For two theories to be comparable, we need the following to be true:

1. Sure outcomes are comparable between theories.
2. Risky outcomes are comparable to sure outcomes for each theory.

The first condition comes from the fact that objective choiceworthiness is exogenous to our construction of decision theories. The second is necessary because decision theories try to make the exact kind of claim that decision theories try to make.⁷¹ Because these conditions are sufficient to make decision theories comparable, and because they are satisfied, decision theories are comparable.⁷² Indeed, the conditions are guaranteed by our assumptions in §2.2 and satisfied by the examples in §2.3.⁷³

⁶⁹ Hypothetical decision theories may not make more meagre claims, but I am content to compare only the kinds of decision theories that people actually propose rather than anything that one would like to call a decision theory.

⁷⁰ Perhaps there is an actual sure outcome with this property, but a hypothetical one is all we need, and it simply might not be true of all decision situations there is an actual sure outcome.

⁷¹ Proof by tautology!

⁷² In the background, I am assuming transitivity of choiceworthiness relations. This is how I think quantifiers such as 'bigger than' and 'better than' work.

⁷³ One might worry that we are giving too little say to risk-averse theories because they assign lower values and might therefore provide a smaller contribution to overall value. Here I invoke MacAskill, Cotton-Barratt, and Ord's (2020) notion

As I alluded to earlier, comparing decision theories means picking a single representation for each theory from its equivalence class. I now argue there is an obvious and intuitive choice of representation. Decision theory has a rich history and originates with the mathematical study of gambling. Its original foundations lay not in risk about objective choiceworthiness, but risk about monetary amounts. You only need to note the frequency in that early literature of terms such as “bet”, “lottery”, “payoff”, and “game” to see these foundations. The goal of decision theories, then, was to calculate the prices that you would rationally pay for various gambles—that is, the monetary amount you would pay for various chances of various monetary rewards. These foundations ensure that the standard representations of the standard theories satisfy the above two conditions. The subjective choiceworthiness of a sure monetary outcome is that sure monetary amount. For example, the expected value of a gamble is the amount you should pay for it, according to EU. So the representation we choose of expected utility is not $EU' = 99 \cdot \mathbb{E}u(a|s) + 7$, but $EU = \mathbb{E}u(a|s)$. The standard representation of expected utility already satisfies (1) and (2), and so do standard representations of REU and WLU. So the work of picking which representations to use has already been done by earlier decision theorists.⁷⁴ We may use the standard representations.

As Buchak (2010, p. 95) confirms, ‘[According to Risk-Weighted Expected Utility] ... the value of a gamble is always at least its minimum and at most its maximum, and improving the probability of getting a good outcome never makes a gamble worse’. This idea, of course, generalises beyond monetary payoffs. In our case, we have a utility function that assigns to outcomes objective-choiceworthiness values for an

of picking representations that give theories equal say. I do not think this worry is justified. We have established that any plausible theory should give the same evaluations to sure outcomes—that is, their objective values. We can now ask what it means for one theory to be more risk averse than another. Risk aversion, in its most general interpretation, is about giving values to risky acts that are closer to their minimum values than a less risk-averse theory. The most risk-averse theory is Maximin. Similarly, the most risk-seeking possible theory maximises the maximum value of an action. Maximax, as a result, will give any risky gamble a (weakly) higher value than any other theory. Is it getting too much say? A risk-seeking theory claims that value is closer to the maximum value than a risk-averse theory does. A risk-averse theory is not unfairly penalised when it gives a lower value, so it is having an exactly appropriate say. A risk-averse theory cannot give higher or lower values to risky outcomes without also giving inappropriate evaluations of sure outcomes. These low values come directly from the claims they are trying to make; this is just what it means to be risk averse.

⁷⁴ To be clear, there are other choices that may work equivalently well, but this is the most natural.

agent.⁷⁵ We have moved beyond these narrow foundations. This is why I earlier gave the value of information in terms of the value of whatever a rational agent would sacrifice to obtain it. But these foundations mean that every theory, in its most simple and standard form, gives comparable values when applied to monetary amounts.

As a final remark, it is worth pointing out that these comparisons may be context dependent. In §2.3 we saw theories that were sensitive to broader features of an agent's decision problem; for example, WLU allows for sensitivity to the stakes of a particular decision situation. In these cases, comparisons will be appropriately context dependent so they can capture fully the claims made by decision theories.

In this section I have presented a way of comparing the claims made by decision theories. Sure outcomes provide a basis: we first ensure that each decision theory assigns the same value to sure outcomes. Next, we can compare the value claims made by decision theories so long as they are coherent and apply broadly enough. When comparing decision theories in this way, we choose a natural and intuitive representation of each theory: the representations used in §2.3.

2.5 VALUABLE INFORMATION

We have seen that decision theories differ in how they evaluate pure-state information. No theory except EU can even guarantee that the value of information is positive. None of the previous results tell us about pure-state information under decision-theoretic uncertainty. Nor do they tell us about the value of information about decision theories. This section gives the set of general conditions under which information is valuable, which is the core result of the chapter (theorem 2.5.2).

Although the argument of this section may seem far from our actual choices and decisions, we must start with fundamental notions and build on them. Heuristic arguments such as that found in MacAskill, Bykvist, and Ord (2020) about information value, rather than mathematical results, are perhaps more relevant to people's lives. However, as we saw in §2.3, our current heuristic arguments are suspect precisely because under decision-theoretic uncertainty information value might not be positive. I think we need a better foundation for our heuristic arguments, which I provide here.

We first need to define two concepts, which I call the *naturally bounded principle* and *weak evaluation dominance*.

⁷⁵ Another approach might be to define values of lives and then compare the chance p of one life and chance $(1 - p)$ of another with a third life with certainty. This was suggested to me by Tim Williamson, and it is the approach taken by Fryxell (n.d.).

Recall that k -theories yield evaluations that represent claims about subjective choiceworthiness, given a distribution over possible evaluations. The $(k-1)$ -choiceworthiness distribution tells an agent which possible subjective-choiceworthiness values an act could have, given the agent's beliefs at orders lower than k . I now claim that no plausible k -theory will assign acts a lower value than their lowest possible value (in any state) nor a higher value than their highest possible value (in any state). I call this *natural boundedness*.

Definition 2.5.1. A k -theory, t_k , is *naturally bounded* if it only assigns k -choiceworthiness values that are bounded by an agent's distribution over $(k-1)$ -choiceworthiness distribution; that is, $t_k(a) \in [\underline{v}_{(k-1)}(a), \bar{v}_{(k-1)}(a)]$ for all $a \in \mathcal{A}$.

Given our interpretation of objective and subjective choiceworthiness, I take natural boundedness to be an undeniable tenet of rationality. The *naturally bounded principle*, then, is satisfied when an agent does not have positive belief in any theory that violates natural boundedness.

Definition 2.5.2. The *naturally bounded principle* says that an agent only has nontrivial credence in k -theories that are naturally bounded, for each k .

This is an extremely minimal restriction. Indeed, if we accept the way of comparing decision theories I presented in §2.4, natural boundedness will be satisfied by all admissible subjective-choiceworthiness functions.

We can see why this is a normatively compelling constraint by considering the following example.

Example 2.5.1 (A Bad Lottery). An agent rejects the naturally bounded principle: they value $\$x$ with probability p and $\$y$ with probability $(1-p)$ just as much as some sure monetary amount $\$z$, where $z > x$ and $z > y$. This agent will pay $\$z$ for a lottery ticket with a potential prize of $\$x$ or $\$y$, guaranteeing an outcome worse than keeping the $\$z$ for themselves.

In this example, an agent makes a purchase that makes them certainly worse off, which is self-evidently irrational. In particularly bad violations of the naturally bounded principle, there will be many lotteries with these features that an agent will accept (or a single particularly bad lottery), and they may be money-pumped until they are penniless. This argument, I think, applies even more convincingly

to utility than to monetary payoffs: it is irrational in general to make yourself worse off with certainty.⁷⁶

Note that if an agent accepts the naturally bounded principle, then subjective choiceworthiness will also be bounded by utilities.⁷⁷ We do not need to make further assumptions about subjective choiceworthiness to guarantee this.

Theorem 2.5.1. If an agent accepts the naturally bounded principle, then subjective choiceworthiness will be naturally bounded for that agent.

Proof. Suppose that t_k satisfies natural boundedness for every t_k for every k . This implies that choiceworthiness is bounded from above at every level by the maximum possible utility value, $\max_{a \in A} \max_{s \in S} [u(a|s)]$, and bounded from below by the minimum possible utility value, $\min_{a \in A} \min_{s \in S} [u(a|s)]$. There are two possibilities for v . According to unique evaluation, we choose a number at the intersection of each of these value ranges; so we must choose a number weakly greater than the lower bound at every level and weakly less than the lower bound at every level. The other case is that in which an agent accepts a single k^* -theory for some $k^* \in \mathbb{N}$. In this case, the k^* -theory will satisfy natural boundedness as a direct result of the naturally bounded principle. \square

Another important property is *weak evaluation dominance*. This is a property of signals. Weak evaluation dominance means that an agent's prospects do not appear worse after receiving a signal than they did before.

Definition 2.5.3. \mathcal{M} satisfies *weak evaluation dominance* for v if $\max_{a \in A} v_m(a) \geq \max_{a \in A} v(a)$ for all $m \in \mathcal{M}$.

In words, a signal \mathcal{M} satisfies weak evaluation dominance for a subjective-choiceworthiness function v if the subjective choiceworthiness of the best act conditional on m , v_m , is at least as high as the value of the best act according to v , for any possible $m \in \mathcal{M}$.

This condition is far from self-evident.⁷⁸ I will not assume that it is satisfied; instead I note that if it is not satisfied, then information

⁷⁶ Assuming the choice is not forced.

⁷⁷ Natural boundedness does not provide a direct solution to Pascal's wager, which I believe is best thought of as a problem derived from the existence of an outcome to which you assign infinite utility. Rather, you would need an additional assumption that *utility* itself is bounded from above, but this is not what natural boundedness says.

⁷⁸ At least in its current form. As I will show, this is equivalent to the generalised claim that information value is positive. One might argue that this second claim is self-evident, but to me this is not obvious, because of cases such as example 2.5.2.

value may be negative. This is this chapter's key result related to positive information value.

Before presenting this as a theorem, we can get a sense of why violating weak evaluation dominance may create negative information value. The intuition for why information value is everywhere positive, for those that take this to be a tenet of rationality, is that after receiving information you either make a better decision or you make the same decision; in either case, supposedly, you cannot be worse off. Of course, this line of reasoning is contested, as I noted in §2.3. Those that contest it point to situations in which the expected value of your decision is higher after receiving the information but where there is some chance that you make a decision that *ex post* appears worse. They say that it is not irrational to want to avoid feeling regret for choosing an act that appears *ex post* worse than the act they would have chosen without information, at the cost of some *ex ante* expected value. I present the following case as an example of violating weak evaluation dominance that illustrates how it is not self-evident that agents should always accept free information about decision theories.

Example 2.5.2 (A Worse Decision Theory). Suppose an agent is considering whether to find out which decision theory is true. One of the possibilities is *Worse*. The *Worse* theory assigns each act a lower value than their current subjective-choiceworthiness values but gives the acts the same ordering as the agent's initial subjective-value function.

This agent will not choose a different act if they find out *Worse* is true, but the act they choose (and all of their other acts) will seem worse to them.⁷⁹ In cases of pure-state information, even when we find out all of our acts are worse than we thought, this information is useful because it allows us to make a better decision. If I found out I lost my job, perhaps all of my acts would appear worse than they do now, but they would not be *uniformly* worse. If I was about to buy a car on credit, I would do better than I would have done had I not known this, as I would not buy it. Information would benefit me in this way because I could make a better decision. But finding out *Worse* is true would make all of my acts have lower value but would not allow me to make a better decision.

I admit that there may be some chance of finding out that all of one's options are better than they thought, and I admit that this

⁷⁹ Given my approach in §2.3, this can only happen if none of the acts have sure outcomes.

might outweigh the prospect of finding out Worse is true, for some admissible v . I also note that there are other situations in which an agent might change how they act, thus getting some benefit from finding out all of their options are worse, and I do not wish to rule out the possibility that v may allow agents to accept information that could make many or all of their options look worse. I simply claim some subjective-choiceworthiness functions allow agents to reject free information such as that in example 2.5.2. Other constraints may rule these out, but I have so far given no such constraints.

Let us now give the result.

Theorem 2.5.2. Suppose an agent's subjective-choiceworthiness function v satisfies natural boundedness. A signal \mathcal{M} has weakly positive value for any $v \Leftrightarrow \mathcal{M}$ satisfies weak evaluation dominance for v .

Proof. (\Rightarrow Direction) Suppose v satisfies natural boundedness. This implies that the value of information will be bounded from below by $\min_{m \in \mathcal{M}} (\max_{a \in \mathcal{A}} v_m(a)) - \max_{a \in \mathcal{A}} v(a)$. Suppose further that \mathcal{M} satisfies weak evaluation dominance: $\max_{a \in \mathcal{A}} v_m(a) \geq \max_{a \in \mathcal{A}} v(a)$ for all $m \in \mathcal{M}$. Note that this implies that $\min_{m \in \mathcal{M}} (\max_{a \in \mathcal{A}} v_m(a)) \geq \max_{a \in \mathcal{A}} v(a)$, and so the lower bound on information value, $\min_{m \in \mathcal{M}} (\max_{a \in \mathcal{A}} v_m(a)) - \max_{a \in \mathcal{A}} v(a)$, must be a weakly positive number. (\Leftarrow Direction) Suppose that $\exists m \in \mathcal{M}$ s.t. $\max_{a \in \mathcal{A}} v_m(a) < \max_{a \in \mathcal{A}} v(a)$. To prove the contrapositive, I must show that there is some v that has negative information value under these conditions. Consider an agent whose choiceworthiness function is such that the value of an act is always its lowest possible value; for example, $v(a) = \min_{s \in \mathcal{S}} u(a|s)$. And note that this subjective-choiceworthiness function satisfies natural boundedness. For this agent, the value of costless information is $\min_{m \in \mathcal{M}} (\max_{a \in \mathcal{A}} v_m(a)) - \max_{a \in \mathcal{A}} v(a)$. As $\max_{a \in \mathcal{A}} v_m(a) < \max_{a \in \mathcal{A}} v(a)$ for some $m \in \mathcal{M}$, the value of information must be negative. \square

Thus, the necessary and sufficient conditions for positive information value are given. This theorem makes very weak assumptions about v , strictly weaker than those made previously, such as theorem 2.3.1. It is also strictly more general, as it takes into account the value of both pure-state and pure-normative information, while previous results have only focused on pure-state information.

Note that we only need to check the information that produces the worst value for each subjective-choiceworthiness function generated conditional on the information, when an agent chooses their conditionally best act.

Corollary 2.5.2.1. Information will be weakly valuable for $v \Rightarrow v$ is naturally bounded and for $\arg \min_{m \in \mathcal{M}} v(\max_{a \in \mathcal{A}} v_m(a))$, $v(\max_{a \in \mathcal{A}} v_m(a)) \geq \max_{a \in \mathcal{A}} v(a)$.

In this section I introduced the conditions of natural boundedness and weak dominance over signals. I found the relation between them and information value in general cases, which gave the necessary and sufficient conditions for information value conditional on the natural boundedness principle.

2.6 COMPARISONS OF VALUE

Our most important decisions are usually about which of two valuable research projects to pursue, which of two valuable books to read, or which of two valuable experiments to perform. The important decisions are choices between signals. However, if two signals are both of *negative* value, it doesn't matter much which is better: we can usually avoid both. This is why I spent so much time on valuable information in the last section. This section offers a brief discussion of how we might compare information value. In considering these measures of comparative value, I will argue that we may want to constrain decision theories further.

Following Blackwell's (1953, 1954 [with Girshick]) terminology, we can say a signal is *more informative if it is more valuable*. (This odd turn of phrase comes from the early analysis of the expected utility of pure-state information, which is always higher for signals that are more informative in various ways, one of which is demonstrated in theorem 2.6.1 below.) We can also say that one signal is more valuable than the other if it has higher subjective choiceworthiness.⁸⁰ Formally, let's consider two signals, $\mathcal{M} = \{m_1, \dots, m_{|\mathcal{M}|}\}$ and $\mathcal{M}' = \{m'_1, \dots, m'_{|\mathcal{M}'|}\}$, and say the following:

Definition 2.6.1. Signal \mathcal{M} is *strictly more informative according to* v than is signal \mathcal{M}' , or $\mathcal{M} \supset \mathcal{M}'$, if and only if $v(\max_{a \in \mathcal{A}} v_{\mathcal{M}}(a)) > v(\max_{a \in \mathcal{A}} v_{\mathcal{M}'}(a))$.

A rational agent should choose signals that have higher subjective choiceworthiness, all else equal. So if two signals have the same price, an agent should choose the one that is more valuable according to v . There is a clear parallel with the definition of information value in §2.2: if \mathcal{M} is more informative according to v than is \mathcal{M}' , then $\psi \geq \psi'$. That is, the definition of more informative according to v corresponds directly to information being more valuable in the way we saw earlier.

In the literature, Blackwell and Girshick (1954, chapter 12) gives interesting results related to pure-state information.⁸¹ Their results

⁸⁰ Here we are assuming implicitly that the two signals have the same price, and that \mathcal{A} is the set of acts available to the agent after paying that price. Intuitively, A is more valuable than B is only true *all else equal*.

⁸¹ Especially theorem 12.2.2 of that book. See also Blackwell (1953).

cover many different ways of characterising more informative signals for pure-state information in cases in which v is the expectation of utility. I do not wish to present the mathematics and all of their results in a general way. Instead, here is a taste, and the interested reader can read their work in detail.

Theorem 2.6.1. Let P be an $|\mathcal{S}| \times |\mathcal{M}|$ matrix of the probabilities (representing rational beliefs) of receiving certain messages from the signal \mathcal{M} in certain states in \mathcal{S} , and let P' represent the same for another signal \mathcal{M}' . For pure-state information, $\mathcal{M} \supset \mathcal{M}' \Leftrightarrow$ if P' can be expressed as a garbling of P and if v is expected utility.

Here 'garbling' is a kind of randomisation: P' is a garbling of P if P' can be expressed as P' multiplied by a Markov matrix. So this theorem says that if the beliefs induced by \mathcal{M}' are a randomisation of the beliefs induced by \mathcal{M} , then \mathcal{M} is more informative.⁸² Future research might try to come up with equivalent conditions to Blackwell's in the context of uncertainty about information value.⁸³

I note that there are other potential measures of informativeness. Let's consider one. Perhaps each possible message is at least as valuable as a message from another signal, and at least some messages are strictly more valuable. In such a case, a signal stochastically dominates the other. I will call this condition *strictly stochastically more informative* and denote it \sqsupset .

In order to check this stochastic property, we need signals to be ordered and on the same index. First rearrange \mathcal{M} such that $\max_{a \in \mathcal{A}} v_{m_j}(a)_i \leq \max_{a \in \mathcal{A}} v_{m_k}(a)$ iff $j \leq k$, and do the same for \mathcal{M}' . Then, put them on the same index.⁸⁴ Call these rearranged and reindexed signals $\hat{\mathcal{M}}$ and $\hat{\mathcal{M}}'$. Stochastically more informativeness

82 Of course, randomised beliefs might be closer to truth in a particular instance. If it will rain one information source is a reliable weather model and another is the reliable weather models prediction ± 0.2 , then on a particular day it might rain and the second model predicts rain with probability 1 and the other only predicts rain with probability 0.8. However, the garbled signal will clearly be less reliable on average.

83 Indeed, to me it is not obvious exactly how this result will apply, given that information can cut across many dimensions!

84 We may perform a transformation on one or both of \mathcal{M} and \mathcal{M}' to achieve this. For instance, \mathcal{M} may be transformed by replacing one or more messages $m_i \in \mathcal{M}$ with a collection of messages $\{m_{i,1}, \dots, m_{i,n}\}$ such that they jointly form a partition over the region of \mathcal{S} that was originally covered by m . Note that this means $p(m_i) = \sum_{j=1}^n p(m_{i,j})$. Also note that there are many such probability density preserving transformations that will allow us to put \mathcal{M} and \mathcal{M}' on the same finite index, and any of these will allow us to calculate stochastically more informative relations. Finally, a word of caution is in order. While any such transformation allows us to calculate stochastic dominance it may not preserve subjective choiceworthiness, so subjective choiceworthiness evaluations should only be performed on the untransformed signals (though they may be ordered as long as the index is unchanged).

is a property of \mathcal{M} and \mathcal{M}' , but to check whether it holds we must check conditions on $\hat{\mathcal{M}}$ and $\hat{\mathcal{M}}'$.

Definition 2.6.2. A signal \mathcal{M} is *strictly stochastically more informative* than another signal \mathcal{M}' , or $\mathcal{M} \sqsupset \mathcal{M}'$, if $\max_{a \in \mathcal{A}} v_{\hat{m}}(a) \geq \max_{a \in \mathcal{A}} v_{\hat{m}'}(a)$ for all $\hat{m} \in \hat{\mathcal{M}}$ and $\hat{m}' \in \hat{\mathcal{M}}'$ and if for at least one j , $\max_{a \in \mathcal{A}} v_{\hat{m}_j}(a) > \max_{a \in \mathcal{A}} v_{\hat{m}'_j}(a)$.

There is a sense in which a signal is better in every way if it stochastically dominates another signal. For any chance of a subjectively more valuable message in one signal there is an equal or greater chance of an equivalent gain in the other. Every message on a stochastically dominant signal is at least as good or strictly better.

It might seem obvious, then, that an agent should always pick a stochastically dominant signal, all else equal. We might call this the *stochastic-dominance-over-signals principle*.

Definition 2.6.3. A choiceworthiness function v satisfies the *stochastic-dominance-over-signals principle* if $\mathcal{M} \sqsupset \mathcal{M}' \Rightarrow \mathcal{M} \supset \mathcal{M}'$ for all $\mathcal{M}, \mathcal{M}'$.

This principle is intuitive. It says that a signal will be more informative according to v if it is stochastically more informative. Given how compelling this principle is, surely subjective choiceworthiness should follow this principle.

Earlier in this chapter, I wanted to be permissive about the admissible subjective-choiceworthiness functions denoted by v , so I only made the minimal assumption that they satisfy natural boundedness. I have succeeded in giving a general theorem about information value. However, this permissiveness has come at a cost, as some admissible theories violate the stochastic-dominance-over-signals principle.

Theorem 2.6.2. There are naturally bounded subjective-choiceworthiness functions that do not satisfy the stochastic-dominance-over-signals principle.

Proof. Consider $v(a) = \min_{s \in \mathcal{S}} u(a|s)$ and two signals \mathcal{M} and \mathcal{M}' s.t. $\mathcal{M} \sqsupset \mathcal{M}'$. Note that $v(\max_{a \in \mathcal{A}} v_{\mathcal{M}}(a)) = \min_{s \in \mathcal{S}} u(a|s) = \min_v(\max_{a \in \mathcal{A}} v_{\mathcal{M}}(a))$, so $\mathcal{M} \not\supset \mathcal{M}'$. \square

I think this motivates stronger constraints on normative decision theories than I have so far allowed. The proof uses the example of Maximin, which as I mentioned in §2.3 could be defined as REU with a particular risk-weighting function. If we take the stochastic-dominance-over-signals principle to be normatively constraining, this contradicts the permissibility of REU in general. To further motivate it as a normative constraint, consider the following example.

Example 2.6.1 (Almost Free). An agent has access to an information source that will give them perfect information about all relevant claims. Amazingly, this information source is almost free, costing only the smallest chargeable amount, perhaps one cent. Even more amazingly, this signal satisfies weak evaluation dominance. For this agent, v is Maximin, so they evaluate everything according to its minimum possible value. This agent must reject this almost-costless perfect information.

This example makes it clear that natural boundedness falls short of a full specification of the normative constraints on v . Maximin, a theory that satisfies natural boundedness, says that the value of this information is its lowest possible value. The value of the current best act minus the cost of one cent is lower than the value of acting without information (any act available after paying one cent is available before, but its outcomes are all one cent worse), so according to Maximin, this information is not worth one cent. Maximin takes full account of the situation in which information doesn't improve your decision but fails to take into account the many possibilities in which an agent is made better off by gaining information. Natural boundedness guaranteed that information value could not be negative in such cases as example 2.6.1. Here information value is not negative, but only because it is exactly 0. So any small cost of information will make the agent reject it, which is absurd.

In this section, I explained how we can compare signals and showed how this motivates a further constraint on admissible subjective-choice-worthiness functions, which I call the stochastic-dominance-over-signals principle.

2.7 AGAINST THE ANALOGUE PRINCIPLE

In §2.2 we saw that considering decision-theoretic uncertainty leads naturally to a hierarchy of higher-order decision theories. To take full account of an agent's uncertainty, this hierarchy may be unbounded, with multiple possible k -theories for every $k \in \mathbb{N}$. Subjective choice-worthiness may still be well defined if each act has a unique value that lies at the infinite intersection of the value ranges across all k 's. Indeed, if only a single value is possible at every level, an agent knows that this value is the true one. As noted, Trammell (2021) gives principles which, if accepted by an agent, guarantee that only a single value will lie at this intersection.

One of these principles is the analogue principle. Here, I argue against this principle. Although this principle guarantees unique evaluation, it does so in an entirely suspicious way. When we see this principle fail, this failure is instructive: it sheds light on what a higher-order decision theory can and cannot be.

Trammell's analogue principle claims that whichever decision theory is true at some level must be true at any level.

Definition 2.7.1. Let t_k^* denote the true k -theory. The *analogue principle* is the claim that $t_k^* = t_1^*$ for all $k \geq 1$.

The appeal of this idea comes from a deep analogy between decision theories and higher-order decision theories. Recall that a k -theory takes in probability distributions over $(k-1)$ -choiceworthiness values and makes a claim about subjective choiceworthiness. Likewise, $(k-1)$ -theories move from a distribution over $(k-2)$ -choiceworthiness values to a claim about subjective choiceworthiness, and $(k+1)$ -theories move from a distribution over k -choiceworthiness values to another claim about subjective choiceworthiness. Decision theories, on the other hand, move from distributions over objective-choiceworthiness values to claims about subjective choiceworthiness. Mathematically, each is a function that takes a distribution as an input and gives a single real number as the output. So, mathematically, higher-order decision theories are exactly the same objects as decision theories. We can even give functionally similar higher-order theories; for example, EU takes subjective choiceworthiness to be the expectation of utility, and EU_k might take subjective choiceworthiness to be the expectation of $(k-1)$ -choiceworthiness. In this way, decision theories and higher-order theories are exactly the same kind of thing, and the best way of moving from choiceworthiness distributions to single values is the best way everywhere. Indeed, it would be absurd to reject the analogue principle while claiming that higher-order decision theories are just like decision theories in all relevant aspects; and a disanalogy is exactly what a failure of this principle shows.

If the true theory is true at every level, and you believe a theory t has a 50 percent chance of being true at level 1, you must also believe that theory t 's k -theory equivalent has a 50 percent chance of being the true theory at level k . An agent who accepts the analogue principle will have the same beliefs in equivalent claims across levels because these claims are about exactly the same thing, according to the analogue principle.

The appeal of the analogue principle is that it guarantees unique evaluation. To see why it is suspicious, consider the following example.

Example 2.7.1 (An Unstable Equilibrium). Consider an agent who accepts the analogue principle and begins with their beliefs in perfect balance; for each risk-seeking theory they endorse at order k , there is a counterbalancing risk-avoiding theory at that order. With their beliefs in balance, their overall choiceworthiness function assigns values to acts that are their expectation. Next they come to believe very slightly more in some extremely risk-seeking theory at some level—say, level 23. After this change of belief, their subjective-choiceworthiness function assigns values exactly as if they were certain of this extremely risk-seeking theory.

In this example, the agent's beliefs start in perfect balance. For every possible departure from EU in one direction there is an equally likely departure in the opposite direction. When this agent comes to believe just a tiny amount in an extremely risk-seeking theory, they become exactly as risk seeking as that theory. I grant that coming to believe some extremely risk-seeking theory a little bit should perhaps make an agent's overall choiceworthiness assignments *more* risk seeking. However, the agent should only make a shift that is small and proportional to their small change in beliefs.⁸⁵ The problem with throwing caution to the wind and choosing *only* based on this most extreme theory is obvious, and the analogue principle implies the agent must do so.

Indeed, things are even worse. If the agent came to believe just a little in a risk-avoiding theory, then either they would continue to evaluate their options as if the original risk-seeking theory were true (if their belief was stronger in that theory), or they would now act as if the new risk-avoiding theory was true (if their belief was stronger in that theory), or they would now again maximise their expectation of utility (if the beliefs perfectly balanced each other once again).

Suppose that the agent next experienced an even larger shift—for example, becoming 50 percent sure of the extremely risk-avoidant theory (with their beliefs in other theories still held in proportion to maintain balance). We would expect a large shift like this to make an agent more risk seeking, but countintuitively they will become no more risk seeking. They already moved so far to the extreme in the first shift that they cannot move any further. So the analogue principle implies unrealistic shifts in risk attitude when an agent shifts their beliefs in the following way: Sometimes for a small shift, their change in attitude is severely out of proportion. At other times, even for large shifts, they will not change their mind at all. In fact,

⁸⁵ Trammell gives a principle of continuity, which has a similar motivation.

this doesn't rely on the balance I set up at the start. I could have let them believe entirely in taking the expectation before their change of mind in example 2.7.1, and they still would have acted the same way.

Example 2.7.1 is illustrative of an unstable equilibrium. It may help to illustrate the concept of an unstable equilibrium with another example. Imagine a ball resting on the very top of a steep hill. Although the ball is perfectly balanced, a slight nudge will cause it to roll all the way down the hill. The ball on the hill is in equilibrium, but a very small perturbation will cause it to move very far away from its equilibrium. Further, once the ball is at the bottom of a valley, a further perturbation will not make it go further down.⁸⁶ Likewise, decision-makers who use the analogue principle to ensure unique evaluation will always end up at one extreme or the other if their beliefs are not perfectly balanced. Example 2.7.1 generalises, as I show in appendix A.1.

To illustrate how this unstable equilibrium comes about, I give another example. Consider the idea of an *evaluative parliament*.⁸⁷ In an evaluative parliament, each of our theories is represented by a person who argues on behalf of that theory. Each time the parliament meets, members make decisions about the choiceworthiness of different options, taking into consideration the choiceworthiness assignments from the last time they met. Each meeting represents a different order of the hierarchy. The analogue principle implies that an agent's beliefs will be the same at each level. If we apply this to the evaluative-parliament analogy, it means that the parliament consists of the same people. Consider the following example.

Example 2.7.2 (The Parliament with Infinite Sessions). A parliament consisting of the same members meets infinitely many times. It is tasked with making the following decision: how much should we value each of the possible actions? The members are almost completely equally balanced: for each member who argues for a risk-seeking assignment of values, there is an opposing member who argues equally strongly for

⁸⁶ This analogy breaks down in the following way. If you try to nudge the ball back up the hill with a small but slightly larger force than before, the ball will not roll all the way up and then back down to the opposite valley. In the case of the analogue principle, however, the unstable equilibrium magnifies whatever force you add to it, and a small but slightly larger change of beliefs towards an equivalently extreme risk-avoiding theory will cause the agent to be extremely risk avoiding.

⁸⁷ I draw on the idea of a *moral parliament*, introduced by Bostrom (2009a) and developed by Newberry and Ord (2021) and Greaves and Cotton-Barratt (2019). I do not engage deeply with this idea, but use it for illustration.

the opposing risk-avoiding assignment. However, there is an *unbalanced extremist*, a member of parliament that advocates an aggressively risk-seeking assignment of values without having a rival who advocates the opposing risk-avoiding assignment. At each meeting, the unbalanced extremist can only move the assignments of value a little towards their risk-seeking extreme. They are only one of very many voices in the parliament, and so they can only move the value assignments a little. At each session, the parliament moves a little closer towards the evaluation of the unbalanced extremist. Over infinitely many meetings, it moves all the way towards the unbalanced extremist's assignments of value, and in the end, the parliament makes collective decisions as if it were a dictatorship run only by the unbalanced extremist, assigning values exactly as the unbalanced extremist prefers.

Although this method guarantees unanimous decisions by the parliament, it does so in a suspicious way. Likewise, the analogue principle guarantees a unique evaluation in a suspicious way.

The analogue principle relies on the intuition that theories at every level are exactly the same sorts of things as decision theories. It stipulates that rational agents must be consistent in their probability distributions over theories at each level by having equal belief in equivalent theories. If there is a theory that is the equivalent of the unbalanced extremist in example 2.7.2, in that it advocates a risk-seeking policy and there is no plausible alternative that is its equal and opposite risk-avoiding theory, that theory will dominate an agent's decision-making. The way this theory dominates is just like the evaluative parliament: at each level, the risk-seeking theory can only move an agent's assignments of value a little, but this is repeated infinitely, and the infinite tiny shifts towards the risk-seeking extreme result ultimately in extreme risk-seeking behaviour. This is what we saw in example 2.7.1: the agent came to believe in a risk-seeking theory that had no counterbalance, and this resulted in their decisions being dominated by this risk-seeking theory.

What has gone wrong is a sort of double counting. When the parliament met the first time, it shifted a little towards the views of the unbalanced extremist. Then in the next session, with no new information or input, it made another small shift towards the views of the unbalanced extremist. This was the first double counting. But the members met again and made yet another small shift, triple counting the same collection of views. In the end, the infinite sessions of the parliament gave rise to *infinite recounting*, and the infinitely

many small shifts towards the view of the unbalanced extremist led to their views dominating the verdict of the parliament.

In the case of example 2.7.2, we reason up infinite levels of decision-making. At each level, our decision-making may be pushed to be a little more risk seeking. But we do this infinitely many times and end up making decisions just as our most risk-seeking theory (that is not balanced out) does. Here, we have *infinite recounting* as well. According to the analogue principle, each level is supposed to be the same, so we cannot be learning new information as we ascend the hierarchy. We just reapply the same information at each level, and the infinite reapplication of the same information leads to an extreme answer, unless your beliefs are in perfect balance.

The analogue principle relies on an analogy between decision theories and higher-order theories. I think its failure shows a disanalogy. At present, I am unsure exactly what else this implies. Perhaps we learn more as we ascend the hierarchy, or perhaps higher-order theories are not the same sort of things as decision theories.

It is worth noting that the other approaches to ensuring that subjective choiceworthiness is well defined, discussed in 2.2.3, do not suffer the same problem; these solutions work under either of the disanalogies I've suggested, so long as higher-order decision theories are not *so* disanalogous that they no longer make subjective-choiceworthiness claims based on choiceworthiness distributions. In fact, the problem might not be as severe for those other solutions even if they were to accept the analogue principle: *infinite* recounting can behave very differently to *finite* recounting even if the finite number is very large.

In this section I presented a problem with the analogue principle. When we saw how it guaranteed unique evaluation with the help of example 2.7.2, and the implications of this solution in example 2.7.1, these implications were clearly problematic. The analogue principle relies on a deep analogy between decision theories at every level, so when the analogue principle fails it shows there must be a disanalogy. I suggested this disanalogy might arise because we in fact learn as we ascend the hierarchy or because theories are not the same things at each level. Regardless of why exactly the disanalogy happens, this is illustrative of what a higher-order theory cannot be.

2.8 A POTENTIAL PROBLEM

In determining information value, we will need to first evaluate how good an agent's prospects are before receiving information and second evaluate how good their prospects are after receiving information. The value of information is the maximum value a

rational agent should be willing to pay for it, which is the price of information that makes them indifferent between their current prospects and their own evaluations of what their prospects will be when they receive information at its price. For each of these two times at which we evaluate the agent's prospects, we have a choice: we might use the agent's prior evaluations, posterior evaluations, or ex post evaluations. The model I set up in §2.2 uses an agent's prior evaluations of their initial prospects. Then it creates a value distribution of their prospects conditional on a message as evaluated by their conditional evaluations. Finally, to evaluate their prospects after receiving a signal, they take their current evaluation over the possible conditional evaluations.

Although this view is compelling, there is an apparent problem. The problem is that an agent may simply choose information that makes them feel good, and this can lead to what seems, from the outside, like irrational behaviour. This apparent problem is resolved when we realise this agent has irrational beliefs, but to me this resolution was not obvious,⁸⁸ so I give a full explanation below. Consider the following example.

Example 2.8.1 (Choosing Books). Consider an agent who is uncertain in deciding between the following 1-theories: EU and Maximax. She accepts the 2-theory that takes her expectation over 1-choiceworthiness. Initially, her belief in Maximax is vanishingly small. The agent can choose any book stocked by her local library, which has a vast section of decision-theory textbooks. Although most of the books argue in favour of EU₁, she knows there are biased textbooks produced by the propaganda division of the local Maximax Church which will increase her credence in Maximax. If her acts vary in utility across states, then she must, according to our model, choose the books that will increase her belief in Maximax.

So an agent with some belief in Maximax must, in certain decision situations, pursue only evidence that confirms their belief in Maximax. Worse still, it doesn't matter how improbable they find Maximax to begin with, so long as it is a live possibility. They might, for example, read only the books that argue in favour of Maximax, avoiding books that argue against it. After doing this, they will believe in Maximax more strongly than they did before. If there were enough books with compelling enough evidence, they might even

⁸⁸ Thanks to both Antony Eagle and Katie Steele for separately pointing out this solution.

convince the agent to be almost sure that Maximax is true. This may be the case even when the totality of evidence would convince the agent to be almost sure that Maximax is false. This example is formalised in appendix A.2.

To me it is obvious that something has gone wrong. Here are some ways we might try to criticise such an agent, though I do not think they provide solid footing for identifying what has gone wrong in example 2.8.1. We might first claim that their choices are very bad, from our perspective. In the worst-case scenario, the agent would pick actions that are, in light of all of the available information, terrible actions—for example, acts with a very large possible payoff with a very small probability and acts with an equally large negative payoff with very high probability. However, the internal logic is sound, and the agent seems to be acting coherently given their beliefs and their understanding of information value. If the agent has made many choices, we might point out the regret they feel for all of their past decisions. However, even if they were consistently disappointed when they achieved the deeply negative payoffs from choosing such acts, they might be undeterred. Sure, the decision turned out bad all those times, but they just got unlucky. When making decisions under risk, sometimes bad things happen even if you make the right decisions. And to them, they are making all the right decisions. We might argue that they should defer to their more informed self. They know that they are acting irrationally in light of the total evidence because they know that most of the available books give evidence against Maximax. This may be so, they might reply, but it would have been irrational for them to read all the books (as shown in appendix A.2). Given that what is at stake in decision theory is, in part, whether certain information sources are valuable at all, this objection comes dangerously close to deciding on the outcome we like in advance.

The problem that leads to cases like example 2.8.1 is that an agent is violating some important principles of rational belief. Their acts are not inconsistent given their beliefs and their acceptance of our model; they simply have irrational beliefs. The agent's beliefs are inconsistent in that they know that reading certain books will change their mind in a particular direction but they have not already factored that into their current beliefs. They cannot choose information with prior knowledge of what that information will say without already knowing what the information says.

When we make the example more realistic by supposing that when they read books from the Maximax Church there is a $1 - \epsilon$ chance of them being convinced, the case is no longer problematic. In reading those books, they will almost definitely not find out Maximax is

true, and these books are no longer always the books that they must read.⁸⁹ Thus, the apparent problem posed by example 2.8.1 is resolved.⁹⁰

In this section I discussed a potential counterexample for my view and argued that it relied on a failure of the agent to hold rational beliefs.

2.9 CONCLUSION

We have seen that there is a way of comparing decision theories that leads to them satisfying natural boundedness. For agents whose beliefs satisfy the naturally bounded principle, we saw that information will be valuable provided it satisfies weak dominance over signals. Then we explored comparisons of value and saw that a desirable property would align stochastically more informative and be more informative according to a subjective-choiceworthiness function. Next we explored Trammell's analogue principle, and I argued that its failure showed that different orders of decision theory cannot be so deeply analogous. Finally, I explored a potential problem but showed how it evaporated for agents with rational beliefs.

This research could be extended in many ways. One could extend my model to include moral uncertainty or representative decision-making for a constituency, perhaps drawing on the work of MacAskill, Bykvist, and Ord (2020). One could try to show that unique evaluation occurs despite arbitrary decision-theoretic uncertainty, even without the analogue principle (this argument would mirror Trammell's (2021), using different principles. One could see what decision theories satisfy the various constraints I set up in this chapter. One also might try to give a positive account of what higher-order decision theories are, if they are not so deeply analogous to regular decision theories. One could also extend the work on information value in this context—for example, by trying to prove a theorem similar to theorem 2.6.1 in the case of decision-theoretic uncertainty. One might also show what restrictions to decision theories can result in the guarantee of positive information value, perhaps using some of my constraints or perhaps using others.

⁸⁹ This appears to be a version of the *reflection principle*, which claims that we should believe now what we know we will believe later. See Fraassen (1984) and Arntzenius (2003) for discussion.

⁹⁰ Another way of making their beliefs consistent would be to allow them to know that reading books will certainly make them believe in Maximax much more strongly, and then, knowing this, they will currently believe in Maximax that strongly. This would not allow for them to gain information, because information must be *informative* (that is, the agent must not already know it).

STATEMENT OF AUTHORSHIP

TITLE OF PAPER:

Weak Comparability

PUBLICATION STATUS:

Unpublished and Unsubmitted Work Written in Manuscript Style

PUBLICATION DETAILS:

N/A

NAME OF PRINCIPAL AUTHOR (CANDIDATE):

Riley Harris

CONTRIBUTION TO THE PAPER:

Devised the arguments and wrote, proofread, polished, and formatted the paper.

OVERALL PERCENTAGE (%):

100%

CERTIFICATION::

This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.

Riley Harris

WEAK COMPARABILITY

3.1 INTRODUCTION

I do not know which moral theory is true.¹ Perhaps I could look to experts to resolve my uncertainty.² Philosophers sometimes claim to know which theory is true, but each, while claiming to know the truth, asserts a different answer.³ Other philosophers claim that no moral theory could be true. Even with my years of education in philosophy, I do not know which philosophers to trust, if any. It is said that ‘*a wise man proportions his belief to the evidence*’, and when the evidence is indecisive, you should be too.⁴ Thus, I am left with moral uncertainty: I have nontrivial credence in more than one moral theory.⁵

This uncertainty does not cause paralysis, a life of absolute solitude, or an attempt to relinquish agency and act on instinct alone. Instead, I continue to laugh, dance, converse, eat, and hug my friends.⁶ Leftover time is dedicated to procrastination. My actions affect those around me, and I want them to be part of the most appropriate or best actions available.⁷

There are several extant proposals for how to act given this moral uncertainty. These proposals are deeply divided. On the one hand are proposals that require a morally conscientious agent to be sensitive to the stakes of their decision; for example, by maximising expected

¹ I will assume that moral theories are the kinds of things that could be true, and that at least one of them is true.

² MacAskill, Bykvist, and Ord (2020) suggest this, among several arguments for taking moral uncertainty seriously. See also Sepielli (2017) and E. G. Williams (2015).

³ Bourget and Chalmers (2014) find that about two-thirds of philosophers are split roughly equally between deontology, consequentialism, and virtue ethics, while the remaining third are split between endorsing multiple views, an intermediate view, rejecting all theories, or being undecided or unfamiliar with the issues.

⁴ I will sometimes use “belief” and “credence” interchangeably.

⁵ That is, my credence or level of confidence is not vanishingly small. Formally, we might suppose my credence was represented by a probability distribution P . Then, my belief $p \in P$ in some moral theory T , would be *trivial* if for any real number $\epsilon > 0$, no matter how small, $p < \epsilon$.

⁶ Pandemic permitting.

⁷ This framing is already controversial. For example, Harman (2015) would not accept that the sense of “appropriate” used here is a true and interesting sense. I take it that my project here is to decide between competing theories of action under moral uncertainty rather than to argue that we should have any such theories. See Sepielli (2017) for a possible reply to Harman (2015).

moral value. On the other hand, some proposals do not allow such sensitivity.

I find stake-sensitivity intuitively plausible, but it may have problematic entailments.⁸ I focus on just one of these entailments here. In order to be sensitive to the strength of reasons given by various theories, we must compare reasons across these theories. If intertheoretic comparisons are deeply objectionable, but required for stake-sensitivity, this would provide a powerful objection to stake-sensitive approaches to moral uncertainty.

However, consider the least restrictive claim about comparability that could defend stake-sensitive approaches to moral uncertainty, at least some of the time. This would be a very weak claim: the claim that we can sometimes compare the strength of our moral reasons. I argue that this is intuitively true, and that rejecting it will mean rejecting other important truths. Thus, it seems to me that the so-called problem of theoretic value comparisons is no real problem for stake-sensitive views.

Perhaps we can go further, extending the scope of comparability to all theories that could be represented as a rational set of preferences over actions. One particularly interesting attempt to do this constructs comparability from sets of apparently incomparable theories using structural normalisation. We can apparently squeeze or stretch moral theories, equalising some statistical measures across theories, such that they become comparable. Sadly, this attempt cannot respect our most basic intuitions about comparability, and cannot be used to defend a stronger claim about intertheoretic comparisons.

The plan is as follows. §3.2 sets the stage by elucidating the debate about stake-sensitivity. Then, in §3.3 I discuss arguments against intertheoretic comparability, and make the distinction between strong and weak comparability. In §3.4 I respond briefly to arguments in §3.3 to begin building a base of support for (at least) weak comparability. In §3.5 I explore the proposal of using statistical normalisation to generalise comparability. In §3.6 I argue against this account and give an impossibility theorem about statistical normalisation in general. In §3.7 I give a final argument in favour of weak comparability. If we reject weak comparability we must accept other implausible claims. §3.8 is the conclusion.

3.2 THE ALLURE OF STAKE-SENSITIVITY

There are two very distinct classes of approach to decision-making in light of an agent's moral uncertainty. The first class of approach is

⁸ Other problematic entailments are discussed in Harris (n.d.[a]) and MacAskill, Bykvist, and Ord (2020)

sensitive to both an agent's credence function and to the strength of reasons according to each theory. There are several proposals in this class. One is that we might maximise expected moral value (MEMV), choosing an option that maximises the credence-weighted strength of our reasons across moral theories.⁹ Alternatively, we might consider moral theories as members of a moral parliament and give them "bargaining power" in proportion to credence.¹⁰ Finally, we might even consider moral theories as voters, letting them vote on the action they prefer.¹¹ Many of these approaches differ significantly in the details, but they also have a lot in common. On each of them we are required to be sensitive to the stakes according to various moral theories.

An alternative class of approaches is not sensitive to the strength of reasons across multiple moral theories. For example, I might simply act on whichever theory I have the most credence in, choosing My Favourite Theory (MFT).¹² Or perhaps I could choose an option I thought most likely to be permissible, choosing My Favourite Option (MFO).¹³ A conscientious agent would only be sensitive to the credence function and the recommendations of moral theories on both these views without reference to the strengths of reasons according to moral theories. These approaches are not stake-sensitive.

⁹ This view is often referred to as "Maximise Expected Choiceworthiness", and is endorsed by Lockhart (2000), MacAskill, Bykvist, and Ord (2020), MacAskill and Ord (2018), Sepielli (2009, 2010), and Tarsney (2017). It is even preferred by Greaves and Cotton-Barratt (2019, p. 27) who write '*... while the bargaining-theoretic approach is interesting, in the end it seems inferior to at least one version of the standard 'maximise expected choiceworthiness' approach*'. We could even include other similar theories to MEMV, such as the risk-weighted variant. This would be the moral-uncertainty analogue of Quiggin (1982) and Buchak (2013). MacAskill and Ord (2018) admit that many of their arguments could be arguments for a risk-weighted variant if that was the best view of ordinary decision theory.

¹⁰ See Greaves and Cotton-Barratt (2019).

¹¹ It is unclear whether or not voting-theoretic approaches should be counted. On the one hand, we might consider the variance normalisation approach (MacAskill, Cotton-Barratt, and Ord (2020)) as quadratic voting applied to moral theories (see Posner and Weyl (2015)). In this sense, the most compelling versions of MEMV, bargaining-theoretic approaches and voting-theoretic approaches, may coincide. On the other hand, variance normalisation is typically understood as part of MEMV. Other approaches that have been called "voting theoretic" include MacAskill (2016a) and Tarsney (2018b). These approaches may not be sensitive to amounts of value, for MacAskill's proposal effectively cardinalises preferences (see Tarsney (2018b)). Either way, I believe they are susceptible to similar problems on a strong version of the incomparability argument. I am currently unsure if Tarsney's account would count as "stake-sensitive" on my view.

¹² Gustafsson and Torpman (2014) defend a refined version of this proposal, and I also consider it a kind of default view in philosophy (consider phrases like "As a utilitarian I would φ ").

¹³ As considered but rejected by Lockhart (2000) and MacAskill and Ord (2018).

Stake-sensitive approaches are appealing because they capture intuitions we have about trading off value between moral theories. Consider the following.¹⁴

Example 3.2.1 (Jill and the Medicine). Jill is a doctor with two sick patients. The first is Anne, a human patient. The second is Charlotte, a chimpanzee patient. Jill has only one vial of life-saving medicine with her and cannot get more. She could give the entire vial to either patient. Upon taking all the medicine, Anne would recover somewhat but have a permanent disability that diminished her quality of life by about half. Alternatively, Charlotte would make a full recovery if she took all the medicine. Either way, the patient that received no medicine would die. A third alternative would be splitting the vial, giving each patient half. With half the medicine, both patients would recover a quality of life just slightly below half of full health. Jill's beliefs are as follows: Jill is certain that the way to aggregate welfare is by summing it up, but she thinks it equally likely that chimpanzee welfare has no moral value or that chimpanzee welfare has the same value as human welfare.

Jill's options are to give all of the drug to Anne, split the drug, or give all of the drug to Charlotte. Let us assume that the true theory of morality says chimpanzee and human life are equally good. We can now represent Jill's decision numerically, as in table 3.1. If this was the true moral theory, then Jill should give all of the drug to her chimpanzee patient, Charlotte.

	Anne	Charlotte
Give all of the drug to Anne	50	0
Split the drug	49	49
Give all of the drug to Charlotte	0	100

Table 3.1: Jill's Decision 1

¹⁴ This case comes from MacAskill and Ord (2018). Cases like this draw inspiration from Jackson (1991), and similar cases are found in Zimmerman (2006), and Graham (2010).

However, Jill does not know which moral theory is true. She believes it is equally likely that giving all of the drug to Charlotte is either permissible or extremely wrong. Likewise, for giving all of the drug to Anne. If she splits the drug, she can be certain that her act is only slightly wrong. We can represent this as in table 3.2.

	Chimpanzee welfare has no moral value (50%)	Chimpanzee welfare and human welfare have the same moral value (50%)
Give all of the drug to Anne	Permissible	Extremely wrong
Split the drug	Slightly wrong	Slightly wrong
Give all of the drug to Charlotte	Extremely wrong	Permissible

Table 3.2: Jill's Decision 2

My intuition is that Jill should split the drug, given her ignorance about the moral truth. If she gives all of the drug to either patient, she risks committing a grave moral wrong.

Consider what stake-sensitive approaches to moral uncertainty will say here. MEMV plausibly claims that Jill ought to split the drug. Next, consider (informally) theories bargaining with each other. One theory wants Anne to receive all of the drug, but splitting the drug is almost as good. The other theory wants Charlotte to receive all of the drug, but splitting it is almost as good. The theories, given the bargaining power implied by Jill's credence, will compromise and split the drug.¹⁵ Given her uncertainty, I think any plausible version of a stake-sensitive view will claim she ought to split the drug. On the other hand, it is obvious that MFO cannot claim that splitting the drug is most likely to be best. Either giving all of the drug to Anne, or giving all of the drug to Charlotte, will be best. Jill is certain that splitting the drug will not be best, according to whichever theory turns out to be true, and thus MFO cannot recommend this option. Equally, MFT requires some refinement, but however this refinement

¹⁵ See Greaves and Cotton-Barratt (2019, p. 12).

occurs, MFT cannot recommend the intuitively plausible option of splitting the drug.¹⁶

If your intuitions track my own, this provides a strong apparent reason for preferring stake-sensitive approaches. We should only relinquish these approaches begrudgingly, when they are shown to be quite unworkable, and I will argue that incomparability arguments cannot force our hand in this way.

3.3 THE CASE AGAINST COMPARABILITY

Stake-sensitive approaches are appealing, but there may be deep problems with them. To be sensitive to the relative stakes on different theories, we must be able to compare stakes.¹⁷ The exact comparability requirements differ across stake-sensitive views. Some voting-theoretic approaches only use facts about how a theory ranks its options. MEMV is particularly demanding. For MEMV we must always know which theory provides stronger reasons for one option over another, and *exactly* how much stronger these reasons are. This is because MEMV requires that choice be specified not only on an agent's actual credence, but on any possible credence function over the same theories. It is easy to imagine alternative stake-sensitive views that do not require strong assumptions about comparability—for example, applying MEMV where comparisons are possible, and applying MFT otherwise.¹⁸ In any case, at least some comparability is entailed by any stake-sensitive view. We can very coarsely partition the possible space of comparability claims as follows:

- A set of theories are *strongly comparable* if we know which of any two reasons is stronger, and *exactly* how much stronger, in any possible situation.
- A set of theories are *weakly comparable* if we sometimes know which of any two reasons is stronger.

There are several objections to intertheoretical comparisons. First, an appeal to cases. It is indeed easy to think of cases where theories appear not to be comparable. Consider an agent uncertain between a deontological theory that says stealing is morally wrong, and a consequentialist theory that says stealing is required in this case because it will bring about the best outcomes. How does the moral

¹⁶ See Gustafsson and Torpman (2014) for a more refined version of MFT.

¹⁷ In many cases it would be enough to have interval comparability; that is, the difference between any two options on one theory is comparable to the distance between those options on another theory.

¹⁸ Harris (n.d.[a],[b]), MacAskill, Bykvist, and Ord (2020), and Tarsney (2020) explore these possibilities.

value of stealing *according to consequentialism* compare with the moral disvalue of stealing *according to deontology*? It may seem that such a question is mistaken.

Everyone should admit that this is, at least, a very hard case. One case may not be so compelling alone, but consider the diversity of possible moral theories we might want to compare: virtue ethics, utilitarianism, ethical egoism, ethical pluralism, and absolutist deontology. There are many moral theories, defended by great philosophers, that are seemingly impossible to compare.

Another objection appeals instead to the fact that within theories the comparability of options is given by the theories themselves.¹⁹ However, no moral theory could give us the tools we need to compare theories. Each moral theory tells you what to do *if that theory is true*, and no moral theory tells you how to act if you are *uncertain*.

A third objection is that when one theory has higher stakes than another theory in some decision situation, stake-sensitive approaches may recommend that we do what this theory tells us even when we do not have particularly high credence in that theory.²⁰

We might wonder if this problem is solved elsewhere. For example, if we thought about moral theories as rational agents with preferences over acts, we might be able to use something from the economics and social-choice literature to solve this problem. However, without explicit assumptions about comparability, it is not enough to assume that moral theories are “well behaved”; we would also need a way of choosing which of the admissible representations to use.²¹ Several attempts have been made to find the best way to aggregate ordinal preferences, and Arrow famously announced that none of these can satisfy our basic desiderata.²² In fact, some of

19 This is discussed by Gracely (1996), with reference to Hudson (1989). It is also discussed by Nissan-Rozen (2015).

20 This is discussed in Hedden (2016), Greaves and Ord (2017), and MacAskill, Bykvist, and Ord (2020).

21 Here’s a sketch of the proof: consider an arbitrary but finite set of options O , and set T of moral theories, that satisfy the famous Von Neumann–Morgenstern axioms: completeness, transitivity, continuity, and independence (see Von Neumann and Morgenstern (1944)). Thus, theories are unique up to positive affine transformations. This means we could represent any theory T_i by $T'_i = aT_i + b$ just so long as a and b are strictly positive real numbers. Now consider an agent who wants to maximise expected moral value. Let that agent have a nondegenerate probability distribution, P , over theories in T . With only these stipulations, MEMV would be arbitrary in the sense that an agent could choose any option that is most preferred by at least one theory, $T_i \in T$. To see this, suppose a specific option $O \in O$ is most preferred by a specific theory $T_i \in T$. Suppose also that a different option O' satisfies MEMV with the current representation. Then, simply replace T_i with a $T'_i = aT_i + b$ with a large enough a and b , keeping other theories constant, to ensure that O is chosen (there will always be a large enough number as the real numbers are unbounded above).

22 See Arrow (1951) and discussion in Mas-Colell, Whinston, and Green (1995).

this work has been discussed in relation to moral uncertainty, but none addresses the objections to comparability.²³ If there is a perfect solution hiding somewhere deep in the literature, I am not aware of it.

These arguments lead to some very strong conclusions about intertheoretic value comparisons. For example, Gracely (1996) writes that intertheoretic comparisons are ‘*intrinsically invalid*’ (p. 327), ‘*essentially meaningless*’ (p. 328), and ‘*ultimately meaningless and fruitless*’ (p. 332). Gustafsson and Torpman (2014, p. 165) find it ‘*hard to see how any intertheoretic comparison of value whatsoever could be made*’ in light of these arguments. I take it that the view is a very strong one: no intertheoretic value comparisons can ever be made between different moral theories. More precisely, the position is a denial of both strong and weak comparability.

The exact amount of comparability required for different stake-sensitive approaches depends on the details of the approach. However, if no intertheoretic comparisons were possible, this would be a fatal blow to stake-sensitive views *in general*. We cannot be sensitive to the stakes on various moral theories if we can never compare them, which undermines any decision-making procedure that incorporates stake-sensitivity at any point.

3.4 INITIAL REPLIES IN FAVOUR OF COMPARABILITY

This section briefly replies to the arguments made above against intertheoretic comparisons. The first argument appealed to cases. Intertheoretic comparisons are *prima facie* implausible in some situations. However, as Tarsney (2018a) points out, just because we have incomparability *somewhere* does not mean we have incomparability *everywhere*. Thus, if we can reply with different cases, ones that demonstrate the *prima facie* comparability of some theories in some situations, this would favour at least weak comparability. Consider the following.²⁴

Example 3.4.1. I am out to dinner with friends. There are two menu items that appeal. One is a beef burger, the other is vegetarian gnocchi. I have a slight preference for the burger. Suppose I believe in hedonistic utilitarianism but find two forms of it equally likely: Either animals have no moral weight,

²³ Especially by MacAskill (2016a) and Tarsney (2018b).

²⁴ This is similar to the “moral dominance” case in MacAskill and Ord (2018). I share their intuition here.

because suffering requires complex mental states that are only found in humans; or, they have the same moral weight as humans. On both hypotheses eating a “human” burger would be a grave moral wrong.

If I chose the beef burger in this case, I might be doing the equivalent wrong of eating a human burger or I might be doing something permissible (while also satisfying my slight preference for the flavour of beef). Alternatively, I lose very little by choosing the gnocchi, and may avoid a great moral tragedy. It seems to me that I should avoid the burger. But if so, we can make comparisons between how bad eating a beef (or human) burger is, and how good having a slightly preferred dinner is. Again, one case might not be convincing, but I could provide other cases too; for example, Jill and the Medicine indicates how comparisons might be made between a moral patient and a possible moral patient across different moral theories.

The second argument was that no moral theory could tell us how to compare across theories. In light of the *prima facie* evidence—that we can make intertheoretic comparisons—I am suspicious of this line of reasoning. Our theory should be built on the evidence, not the other way around. If theories are comparable there are many deep questions to answer about their nature, but the existence of such questions should not make us doubt comparisons. Everything that is now trivial seemed mysterious at some point.

A better objection would be that there can be no satisfactory account of intertheoretic comparisons. Perhaps if we had just conducted a long, fruitless search for how intertheoretic comparisons might work, this conclusion would be warranted. But moral uncertainty is a new field, which does not look like it is running out of fresh new ideas. I will explore a possible theory of how intertheoretic value comparisons occur in the next section. We might try to argue that such accounts must always fail. Such an argument might even build on the impossibility theorem of the next section. However, no such argument has been made thus far.

Finally, there was the concern that some theory would dominate our decision-making when the stakes are higher according to that theory, when that theory is the one we find most probable. I will put this consideration aside, partly because I find this intuition compelling in the case of regular empirical uncertainty. Consider that it is prudent to insure your home against fire even if the risk of fire is very low. Likewise, in Jill and the Medicine, I thought that splitting the drug was the most prudent option, even though Jill was certain this act would be slightly wrong regardless of how morality

turned out. Later I will explore another case. Others agree with my intuitions here, and rejecting these results may lead to worse problems.²⁵ We could make this objection more compelling if we considered truly infectious theories. Such theories might dominate our decision-making in every decision situation, no matter how unlikely we found them, just so long as we were not certain they were false. This strengthened objection is no longer primarily an objection about intertheoretic value comparisons and seems to be worth addressing separately. I will not address this here.²⁶

3.5 NORMALISATION: THE BEST ATTEMPT AT STRONG COMPARABILITY

As mentioned, different stake-sensitive approaches require different kinds of comparability. In the space of possible stake-sensitive views,²⁷ MEMV is particularly demanding. If we can solve the problem here other approaches will have a much easier time. In this section I will discuss a potential solution strong enough to allow MEMV, but argue that it fails.²⁸

What would count as a solution here? To simplify things, and leave out other issues, we might assume each moral theory is “well-behaved” in the sense that it can be represented by a preference function that satisfies the Von Neumann-Morgenstern axioms.²⁹ However, there will be many functions that represent each moral theory, and we need a way of deciding which representation to use.³⁰ Deciding on a particular representation for each moral theory would be equivalent to deciding exactly how the value of each option compares across moral theories (strong comparability).

25 Greaves and Ord (2017) argue that we should accept swamping results in the case of population axiology. Wilkinson (forthcoming) shows that rejecting swamping leads to serious problems.

26 See Harris (n.d.[a]), MacAskill (2013), MacAskill, Bykvist, and Ord (2020), and Ross (2006) for discussions about infectious nihilism. See also the literature around Pascal’s wager, especially the argument that we need not accept infinite value (Bostrom, 2009b) and may be able to build a decision theory that avoids this problem (Colyvan, 2008; Colyvan and Hájek, 2016).

27 Here I include risk-weighted and other non-expectational approaches, and views that are stake-sensitive at any step in the decision-making procedure (MacAskill, Bykvist, and Ord (2020) and Tarsney (2020)), softer views such as a general trade-off approach (see Harris (n.d.[a])), or ethical pluralism applied to moral uncertainty (see Harris (n.d.[b])).

28 Using a case from MacAskill, Bykvist, and Ord (2020, §5) to do so.

29 See Von Neumann and Morgenstern (1944).

30 Moral theories would be an equivalence class containing all positive affine transformations of some value function that satisfies the von-Neumann-Morgenstern axioms. Our problem is how to choose a particular representation from this equivalence class, for each moral theory. Without a way of choosing we will have severe problems; see footnote 21.

One way of choosing representations would be to squeeze or stretch each function, giving them the same range, or the same mean. In fact, we could *normalise* with respect to any statistical feature. We might try to fairly represent each theory, giving them each equal say in the overall decision-making process. MacAskill, Cotton-Barratt, and Ord (2020) formalise this idea of giving theories equal say in two ways and argue we should normalise theories by their variance. That is, we could choose representations for each moral theory, such that each preference function has equal variance to every other preference function.

The first formalisation of equal say is distance from a uniform theory—a theory that assigns the same value to every option. A uniform theory can be reasonably thought of as having no say in the outcome of a decision process. Thus, we should give each theory equal ability to move the overall expected moral value away from the uniform theory before this ability is scaled by credence. However, the decision may still hang on how we define distance. MacAskill, Cotton-Barratt, and Ord argue that two notions of distance give deeply unintuitive results.³¹ Of the remaining notions of distance, Euclidean distance is most natural, though their arguments do not appear to rule out many of the other possible notions of distance. They find that when we ensure theories are equal Euclidean distance from the closest uniform theory, this is equivalent to normalising by their variance.

The second formalisation of equal say gives theories an equal expected value of being a part of the decision-making process (again, before this is scaled by credence). Of course, expected value requires normalisation (that's how we got into this mess), and so they stipulate that the normalisation method in the setup must be consistent with the resulting normalisation method. Surprisingly, they find that the only consistent method here is to normalise by a theory's variance. They argue overall that this is the best statistical normalisation method. I find this argument convincing; it appears to be the best way of choosing representations based only on statistical features alone.³²

³¹ Specifically the l_1 -norm (which is equivalent to normalising by the mean absolute deviation from the median), and the l_∞ -norm (which is equivalent to normalising by range).

³² MacAskill, Cotton-Barratt, and Ord (2020) argue against range normalisation (Hausman, 1995; Lockhart, 2000), normalisation of the distance between mean and minimum values (Sen, 1970), normalisation of the distance between mean and maximum values Sepielli (2013), and several of their own proposals.

3.6 AGAINST NORMALISATION

This kind of approach, however, cannot respect our intuitions about the strengths of various theories. Consider the following case.³³

Example 3.6.1 (A Change of Mind). Suppose I believe that some form of utilitarianism is the true moral theory. Initially, I believe that humans have significant moral value, while nonhuman animals have severely reduced moral value. Later, I change my mind. I come to believe that both humans and (nonhuman) animals have equal moral value.

It seems there are several ways this change of mind might have occurred. One way involves a loss of apparent value. Perhaps I initially thought that humans had rich experiences and significance that other animals did not have, but then came to believe that only a narrow range of hedonic states were valuable. Then it appears that apparent value is lost by this change of belief; what was most valuable (in my previous conception) is now lost. Consider another way this change could occur. I initially believed that only the hedonic states of humans were valuable, but then came to believe that animals had such states too. Then it appears that the world has much more apparent value than before. There is a third interpretation too. Perhaps apparent value remains constant but now appears to be distributed differently across humans and nonhuman animals.³⁴

Consider again the Dinner case. I initially believed that nonhuman animals have value, but that their value is much lower than that of humans. I have not given enough detail to know which option would be best under considerations of moral uncertainty; relative to my initial beliefs, I have apparent reason to do what I most prefer (to eat the burger), and an apparent reason to avoid eating meat. Later, I come to believe that all animals, human and nonhuman, have

³³ MacAskill, Bykvist, and Ord (2020, §5) use cases like this to argue against the same class of theories, though they focus on the metaphysical aspects of intertheoretic comparisons: what grounds comparisons of value? Ultimately, they argue for a particular view of what grounds comparisons of value. Roughly, it's second-order relations between choiceworthiness properties (this view draws inspiration from an account of other quantities given by Mundy (1987)). However, I am interested in whether we can always make justified comparisons, which depends at least in part on our access to answers to this metaphysical question. Their view implies that we can justifiably compare moral theories in some situations (for example in situations in which you have well-defined credence over various amplifications of each moral theory), but they do not argue how far this extends (and I, not surprisingly, think it extends only to *some* possible situations).

³⁴ In fact, there are many more than three possibilities—there is an entire spectrum of possible ways in which apparent value could have changed.

equal moral value. This change makes my apparent reason to avoid eating meat stronger. But how much stronger would this reason be? The answer, I believe, depends on how my change of mind occurred. If I came to believe that humans were much less valuable than I previously believed, this reason would be weaker than if I came to believe that animals were much more valuable than I'd previously thought. If the strength of my reasons depended on how the change of mind occurred, then this fact could impact our decision-making in some circumstances.

No statistical normalisation method can account for this case adequately. If we choose any normalisation method, and then ask how strong our reason against eating meat is after the change of mind, that normalisation method must give an answer that is independent of any facts about how the change occurred. Statistical normalisation cannot treat these different changes of mind differently. In general, where we have intuitions that theories with the same statistical normalisation have differing strengths, statistical normalisation will not adequately account for this intuition. This hints at an impossibility theorem:

Theorem 3.6.1. No statistical normalisation method will be able to satisfy all of the following:

1. *Determinacy*: every possible act or option will have an exact, determined evaluation after normalisation.
2. *Consistency*: every theory will be combined consistently with reference to its statistical properties.
3. *Intuitiveness*: respecting our intuitions about the strength of various reasons.

Statistical methods in general satisfy (1) and (2). I have argued that they cannot handle (3). This argument would generalise to any statistical normalisation, because no statistical normalisation method can handle A Change of Mind. Rejecting our intuitions about cases here would be especially problematic, as they were part of the reason for believing any comparisons were possible. This seems to me a striking flaw in any statistical normalisation method.

Further work is required to decide whether this blow is fatal. Perhaps other accounts will fare no better.³⁵ Then we could have a way of making consistent comparisons everywhere that violates some of our intuitions about comparability, or we would have to accept noncomparability somewhere. If so, we may have some bullets

³⁵ For example, see Ross (2006) and Sepielli (2010).

to bite, and perhaps this is the least disgusting. In any case, stake-sensitive views may be useful where comparisons are justified. I will make one final argument for weak comparability.

3.7 A FINAL ARGUMENT FOR WEAK COMPARABILITY

Let's turn to a final argument for weak comparability. I have already argued that, intuitively, some comparisons are possible. Consider now what the world would be like if no intertheoretic comparisons were possible. This would imply some very odd conclusions. To deny these odd conclusions, we must accept weak comparability.

The first implication is that interpersonal comparisons would also be impossible. To see this, consider first the following comparison. Person A believes that abortion should be avoided if possible, mainly because of the negative effects on the mother. Person B believes that abortion is as bad as murder. Who believes abortion is worse? The natural answer is B, but this would imply we can compare the badness of abortion across moral theories.³⁶ As noted in the last section, we can treat moral theories as if they are people with preferences. Thus, if interpersonal comparisons of preference strength were possible, this would imply that intertheoretic comparisons are also possible.

If interpersonal comparisons are impossible, moral theories that rely on them are unworkable. Consider, for example, preference utilitarianism. On this view we ought to do whatever has the best consequences in terms of overall preference satisfaction. However, if interpersonal comparisons are impossible, this would make preference utilitarianism unworkable. Admittedly, one might use this as an argument against preference utilitarianism,³⁷ but for some this will be an uncomfortable implication of rejecting interpersonal comparisons outright.

We may need to also reject comparisons of welfare at different times within the same life. Consider my own. I struggle to empathise with my darkest moments of depression now, as I am feeling largely at peace. I also dimly remember that part of the suffering of those moments was the inability to imagine anything ever feeling okay, let alone good. In this case, it is obvious to me when the better experience was. I have higher welfare in and prefer my current state. But how is this not like comparing the welfare of a depressed person other than myself, to my own peaceful state? We might even reject *any* difference between comparisons within a life and between lives.

³⁶ I do not wish to make any normative claims here, but provide this example purely as a comparison of different views.

³⁷ See Coakley (2016) for discussion.

On reductionist views of personal identity, endorsed, for example, by Parfit (1984, p. 275), '*a person is like a nation*'. A person, or a series of selves across time, are in many ways, similar to a nation that has many people. If we accept the reductionist view it should make us reconsider many deeply held beliefs, but it should not make us doubt that our peak experiences are better than our lowest depressions, even when those depressions are far away in time. Even if we don't want to take such a controversial view of identity, it is difficult to see how facts about identity could ground comparisons within lives without grounding them across lives.

Just as rejecting interpersonal comparisons of preferences held problems for preference utilitarianism, so too does rejecting comparisons within a life hold problems for self-interest theory. Based on this theory, I ought to do what is best for myself overall. However, if it seems that the most I could do is whatever is best for me in the present moment, self-interest theory becomes incoherent.

Thus, rejecting weak comparability comes at a considerable cost. Not only would we deny comparisons across theories, but also certain kinds of comparisons between preferences of different people. Worse still, we would need to reject comparisons within a life. These claims are controversial and imply that both preference utilitarianism and hedonic self-interest theory are incoherent. We could use similar reasoning to show problems for hedonic utilitarianism and preference self-interest theory. It is always possible to bite the bullet, but I find it deeply unintuitive that we should reject comparisons within lives.

Note that rejecting strong comparability, but maintaining weak comparability, is enough to avoid the worst of these conclusions. We may be able to make the comparison in the abortion case, but not determine exactly by how much B has the stronger view. Preference utilitarianism may be partly undefined, which would serve as an objection, but it would not be unworkable everywhere. Finally, within my own life, I would be able to claim that my worst states were worse than my current state.³⁸ However, I would perhaps not be able to compare all of my states precisely. There may be implications for moral theories; for example, preference utilitarianism may need to be modified to account for action in situations of incomparability. These implications would be less problematic.

³⁸ Here "my" is used loosely to refer to a self or a collection of selves. I expect that any plausible theory of identity has some (perhaps complicated) way of precisifying the term.

3.8 CONCLUSION

Intuitively, intertheoretic comparisons are sometimes possible, and, at other times, very difficult or impossible. This lends credibility to a weak claim about comparability. I have argued that open questions about comparability do not provide a strong reason to be sceptical about weak comparability. Some argue that any comparability would allow unintuitive consequences, but I did not find these consequences unintuitive. I also argued that rejecting weak comparability would entail a rejection of several plausible positions on other topics. Along the way I examined an attempt to extend intertheoretic comparisons to a wider range of situations. Statistical normalisation may treat theories consistently, but in doing so it gives unintuitive results. Weak comparability is the best position in light of these considerations.

Consider now the implications for decision-making under moral uncertainty. Stake-sensitive approaches entail at least the weak claim about intertheoretic comparability. If we rejected even weak claims about comparability, this would provide a powerful objection to stake-sensitivity. I have argued that we should not reject these claims. If I am right, we should also not reject stake-sensitive approaches to moral uncertainty.

STATEMENT OF AUTHORSHIP

TITLE OF PAPER:

When Forced, Do Your Best: How to Make Decisions in the Face of Regress

PUBLICATION STATUS:

Unpublished and Unsubmitted Work Written in Manuscript Style

PUBLICATION DETAILS:

N/A

NAME OF PRINCIPAL AUTHOR (CANDIDATE):

Riley Harris

CONTRIBUTION TO THE PAPER:

Devised the arguments and wrote, proofread, polished, and formatted the paper.

OVERALL PERCENTAGE (%):

100%

CERTIFICATION::

This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.

Riley Harris

14 Oct. 2021

WHEN FORCED, DO YOUR BEST: HOW TO MAKE DECISIONS IN THE FACE OF REGRESS

4.1 INTRODUCTION

Practical rationality is about responding appropriately to the reasons implied by our beliefs.¹ Consider the following example.²

Example 4.1.1. Toby accidentally causes a small oil fire that is currently contained to a saucepan. Toby knows that if nothing is done, the fire will ignite nearby flammables and eventually consume his house. Toby falsely believes that water puts out fires such as this one; in fact, the evidence he has available justifies his false belief.

It is uncontroversial that if Toby knew that water only makes oil fires more fierce, then he ought to cover the fire with a lid.³ But Toby believes that water will extinguish this fire. Perhaps more controversially, I think Toby ought, given his mental state, to throw water on the fire.⁴ These requirements prescribe what Toby ought to

- 1 This view is defended at length by Parfit (2011, volume 1, part 1), who argues we should do what we would have the most reason to do, were our beliefs true. Similarly, MacAskill and Ord (2018) argue that we should choose the option with the most expected choiceworthiness, where ‘choice-worthiness ... represents the strength of the reasons for choosing a given option’ (MacAskill and Ord, 2018, p. 3, emphasis in original). Others are motivated by similar considerations.
- 2 This example takes inspiration from Parfit’s snake example (Parfit, 2011, volume 1, pp. 34–36).
- 3 Some say this sense of ought isn’t relative to his beliefs at all; instead, they say, it is relative to the (belief independent) facts. As Weatherson (2014, p. 157) writes, ‘The externalist is only committed to the view that the most important evaluative concepts are independent of the agent’s beliefs’. Crucially, on this view there still might be important (but not *most* important) evaluative concepts that *are* relative to an agent’s beliefs, and so there may still be a regress. (I thank Pamela Robinson for pointing this out.) In any case, the debate over which of these concepts is most important can be set aside to focus on how (and whether) these concepts work. Tarsney (n.d.), in contrast, claims the regress gives us motivation to accept the norm₂ that takes expectations over lower-order norms. He argues this is the most natural understanding of rationality, and so we should stop the regress here. I demur. Certainly the regress motivates us to accept *some* norm, at *some* order, but I prefer to vary the order as is appropriate in this context (this is especially apparent in my contrasting intuitions about examples 4.1.2 and 4.2.1).
- 4 Sepielli (2017) argues that the correct beliefs to reference are “epistemic credences”. I agree in principle that in certain situations the mental states to reference are not

do, given what he takes the world to be. The prescriptive part makes them normative, so we can call these requirements *norms*.

The requirements are also conditional on aspects of Toby's mental state: they take the form "When X has mental state M, X ought to φ ". The relevant mental states could be what Toby knows or believes or even his degree of belief.⁵ Although it is interesting to consider whether there are norms that do not take this form, this will not be my focus.⁶

We often say that norms perform their prescriptive role by classifying our acts, perhaps in a coarse-grained way (required, permissible, or impermissible) or in a fine-grained way (assigning them cardinal values or ranking them from best to worst, for example). In classifying our acts, they say what we ought to do and ought not to do.

Everyone agrees that it is sometimes appropriate to be uncertain about empirical matters.⁷ I am uncertain, for example, whether it will rain tomorrow. Norms can take our uncertain beliefs as inputs; famously, one norm of this kind calls for maximising one's expectation of value.

However, we can also be uncertain about norms. Consider the following case.⁸

always the agent's *actual* beliefs but rather some more complicated thing such as the beliefs that are most justified by the available evidence. This is especially apparent when considering moral blame (see Harman (2015)). However, in some situations what concerns me most is the agent's current predicament, in which they can only be guided by their actual mental states. In any case, I will allow the reader to substitute whatever mental states seem most appropriate here, as nothing hangs on this. See Parfit (2011, §1) for a summary of these possibilities.

5 Indeed, Jackson (1991) shows there are certain norms which seem to only apply when agents have intermediate credal states—for example, those that could be accurately represented as 50 percent credence in P (and 50 percent credence in $\neg P$). This shows that we can't make all norms relative to full belief and give an agent degrees of belief in *those* norms. See Parfit (2011, pp. 159–60) and MacAskill and Ord (2018, pp. 4–6) for structurally similar cases.

6 For example, we might think that relative to the fact that covering an oil fire will extinguish it, there is a fact-relative norm that prescribes covering the fire regardless of Toby's mental states. We might even think, first, that the norms I described as conditional on Toby's mental state have special relations to these fact-relative norms and, second, that the knowledge-relative norms have special relations to the belief-relative and fact-relative norms. We might even want to ground knowledge-relative norms in knowledge of facts and fact-relative norms or insist that belief-relative norms are the wider class of knowledge-relative norms (knowing is a special case of believing, namely the case in which your beliefs correspond in the right way to the facts). All of these issues can be set aside for my purposes.

7 Perhaps uncertainty is appropriate when we are not experts in a field, when the field is complex, when we have limited access to evidence, when our evidence is conflicting, or when experts disagree. Regardless of whether it is appropriate, we often are uncertain.

8 MacAskill and Ord (2018, p. 1) mention a similar case.

Example 4.1.2 (Environmental Policy). Ada is in charge of environmental policy and has several options for how to regulate emissions. She is uncertain about the value of future generations relative to the current generation but is certain that her options trade off benefits to those alive today against benefits to those who will be alive in the future.

In a case like this, Ada could be certain of all of the empirical facts about how trade-offs will affect future people yet still not know which policy to choose because of her uncertainty about the comparative value of the welfare of future and current people.⁹ Ada could be uncertain which norm is true. She could face *normative uncertainty*.

Normative uncertainty seems particularly tricky. We might at first try to decide by accepting (adopting as correct) a norm.¹⁰ Perhaps it is the norm most likely to be true,¹¹ or the norm that most experts agree on, or the first norm we think of.¹² But a rational way of choosing which norm to accept is itself a norm.¹³ Although it fits the pattern for a norm, this second kind of norm seems a little different from the first kind. It's a function mapping from our beliefs about facts and about norms to a classification of acts rather than a function mapping from beliefs about facts to a classification of acts. Call it a norm₂. In accepting a norm, we are (perhaps implicitly) accepting a norm₂. Perhaps we ought to accept the norm₂ that we find most plausible, the one that maximises expectation of value across norms over norms, or something else. Perhaps we could even take into account special norm₂ properties, such as coherence between our chosen norm₂ and our chosen norm. Any rational way to choose will be what I will call a norm₃—that is, a rational way of choosing which norms₂ to accept. But any way of accepting a norm₃ (which will be our way of accepting a norm₂, which in turn will be a way of accepting a norm₁, which in turn will be a way of “accepting” an act)

⁹ For a more thorough motivation of the problem of normative uncertainty, see MacAskill, Bykvist, and Ord (2020, §1).

¹⁰ If we accept a norm because we do not face normative uncertainty, there is no problem. If we only accept a norm for pragmatic reasons, there will be a regress.

¹¹ Per Gustafsson and Torpman (2014).

¹² Other second-order norms are suggested by MacAskill, Bykvist, and Ord (2020), Greaves and Cotton-Barratt (2019), Newberry and Ord (2021), and others. As is clear in the early chapters of MacAskill, Bykvist, and Ord (2020), different norms₁ have different structural properties and thus may require different kinds of norms₂. I leave these issues to the side, as my proposal is fairly broad and the details can be filled in for more specific cases.

¹³ This is true regardless of whether reasoning of this type is in the foreground or the background. In the case of Toby, he might not ever explicitly refer to a norm, but he still acts on one.

will itself be a norm₄. There is no obvious point at which accepting a norm_k does not entail that we need to accept a norm_(k+1). This structure could be called a hierarchy¹⁴ or, better, a *regress*.

Back to the example. Perhaps Ada knows option A will maximise welfare for the current generation but cause severe harm to some future generations, while option B will reduce welfare for the current generation but avoid severe harm to future generations.¹⁵ For simplicity, suppose she is a utilitarian, and so she wants to maximise social welfare. But though she knows the welfare facts for each generation conditional on her policy decision, she does not know how their interests should count against each other. We can represent this uncertainty about different versions of utilitarianism as uncertainty about pure time-discount rates (δ).¹⁶

Suppose Ada's decision can be represented by the utilities in table 4.1.

	$\delta = 0.05$	$\delta = 0.01$	$\delta = 0.0125$
A	30	20	10
B	100	10	1

Table 4.1: Ada's Decision 1

At this point, Ada might be uncertain in deciding between two ways of making the decision represented in the table; for example, she might maximise expected value (given her credence about the discount rates). She might know of another way of making that decision: simply pick the discount rate she believes most likely to be true and then act upon that discount rate. These are both norms₂. She might want to choose whichever norm₂ she thinks most likely to be true. This is a norm₃ because it is a norm that governs decision under uncertainty about norms₂. An alternative norm₃ would choose the norm₂ that maximises the maximum possible value across norms₂. But there are also many norms₄ that could govern her choice of norms₃: perhaps she should choose such that her norm₄ is coherent with her lower-order norms, or formalise her norms₃ as if they were

¹⁴ Trammell (2021).

¹⁵ Perhaps policy A will result in a slow transition to a zero-emissions state, perhaps at roughly the status quo rate, while policy B will result in a rapid transformation of the current economy to a zero-emissions state.

¹⁶ For background on this kind of case in particular, see Fleurbaey and Zuber (2015) and Greaves (2017a,b). In the end, there may be dominance arguments that resolve this uncertainty, as suggested by Greaves and Ord (2017), but if Ada does not know this, she may still face the regress.

voters and count their votes according to some voting procedure, or randomly select—according to her credences—a “dictator” from her set of norms₃.

The existence of many possible orders of norms does not by itself entail a problem for rational decision-making, just as the existence of many different measures of length (Euclidean being the most common) does not prevent me from knowing which of two people is taller. Notice the pattern for norms: our acceptance of any norm_k is governed by our accepted norm_(k+1). So to accept a norm at order k we must accept a norm at order $(k + 1)$. This regress appears vicious because it seems we must accept infinitely many norms in order to make a rational decision.

Many, do in fact consider this a problem. Consider the writings of MacAskill, Bykvist, and Ord (2020, p. 23, emphasis in original) ‘*it seems that if one can be uncertain about which first-order moral theory is correct, one can also be uncertain about how to deal with moral uncertainty itself. But it seems like this uncertainty can go even higher: one can be uncertain not only about how to deal with moral uncertainty, but also about how to deal with uncertainty about how to deal with moral uncertainty, and so on ad infinitum.*’, Trammell (2021, p. 1) ‘*this reasoning can launch a regress of ever-higher-order uncertainty, which may leave one forever uncertain about what one ought to do*’, and Sepielli (2017, p. 113) ‘*the possibility of normative uncertainty all the way up makes the uncertainty project look pointless*’. Especially strong is Weatherson’s view that this spells trouble for all mind-state-relative norms (which he calls *internalist norms*). As he puts it, ‘... [*this approach*] is vulnerable to a nasty regress. The problem is that internalists disagree amongst themselves, and there is no internalist friendly way to resolve the disagreement’(p. 17).

I, however, do not think this regress is vicious as relates to our ability to make decisions. If certain norms were rational, then the structural similarities between orders and norms would allow us to accept an infinite variety of orders of norms at once. For example, we could accept at every order the norm that minimises the minimum value across norms of lower orders. This norm would always choose itself, and so, just like that, I could accept infinitely many norms (of different orders) and pick whichever act ensures the lowest minimum value given my empirical (un)certainly.

Of the many possible ways to choose given the existence of this regress, most are not plausibly rational. Anyone conscientious enough to consider various ways of acting rationally under normative uncertainty will find a proposal such as this entirely unappealing. The slogan “When uncertain, always choose the worst option” perhaps has no other appeal than its ability to avoid the regress problem. Consider Ada again. She *could* pick a policy she knows will have the

most catastrophic outcomes for both current and future generations, but she *shouldn't*. So while the proposal just given would allow her to choose, it would more specifically allow her to *choose badly*.

The goal of this essay is modest. I want to solve the regress with a principle that has prima facie normative appeal. I propose a principle which can be captured in the slogan “when forced, do your best”. I argue it is more compelling than two other principles that have been proposed in the literature. I explore these alternative proposals first, in §4.2, then present my solution in §4.3. I address difficulties relating to idealised agents in §4.4.

Fascinatingly, another regress might begin when an agent does not accept “when forced, do your best”. My proposal cannot say anything useful about *this* regress, but nor can my opponents'.¹⁷

4.2 EXTANT PROPOSALS

This section explores two proposed solutions to the regress problem and offers a critical note. First, I explore “reason as far as you can”, the proposal that the regress reaches a natural stopping point when an agent reaches their cognitive limits. At this stopping point, an agent faces only finitely many orders of norms and is able to decide based on the norms at the highest order. The second proposal, a fixed-point solution, says that under certain conditions there is a unique value at the intersection of our value ranges (for each order of norms). This unique evaluation is one which an agent can take to be the value of that act, given their uncertainty at every level, and they can choose the act with the highest such value.

4.2.1 Reason as Far as You Can

Consider the proposal that says we should reason up the regress as far as we can. Imagine taking your uncertainty into account at the level of norms₂, then applying norms₃, and so on. Pretty soon you will be working across norms of several different orders, where perhaps even subtle changes in norms at lower orders will result in important changes at higher orders. The most diligent and intelligent of us may be able to handle a few such levels, but no one can handle arbitrarily many. It is natural to think that while norms at orders 1

¹⁷ In fact, for a fixed-point solution, this problem may be even worse because it is unlikely that an agent would accept the many conditions required to show there is a fixed point.

and 2 are more or less understandable,¹⁸ but that norms of much higher orders are not things we can understand deeply (indeed, perhaps we may even be unable to generate plausible candidate norms at much higher orders). If so, then it is natural to think that any human will reach a point they cannot think beyond. This may come from those higher orders being impenetrable, but it may come more ordinarily from our natural limitations; our cognitive capacities have a limit, and we may be unable to apply higher-order norms to our situation even if we can specify them. When we reach this point, we should stop deliberating. As Zimmerman (2008, p. 42) writes, ‘*For every human agent there will always be a level of evidence, L, such that ... there is no such higher level ... [or] any such level is one at which what maximizes expectable value at it also maximizes expectable value at level L ... I then propose that ... [a]n agent ought to perform an act if and only if it is the option that has the greatest expectable value for the agent at his definite level of evidence.*’¹⁹

On this view, there is some finite level beyond which our biological limits prevent us from understanding fully (perhaps because we cannot form coherent beliefs about that level, perhaps because we can’t even conjure the possible norms in which to have beliefs, or perhaps because we simply cannot calculate what those beliefs imply). We should not be expected to reason any further. Zimmerman (2008) proposes that at the highest feasible order, we take our expectation given our uncertainty.²⁰ If Zimmerman (2008) only meant to show that the regress does not undermine our rational decision-making when we are uncertain in norms₂, then he may not have meant that we *must* reason as far as we can. If so, my proposal only extends and refines this approach. I do not believe that we always should reason as far we can, so on this interpretation of Zimmerman (2008) my

¹⁸ Indeed, much has been written on morality, and a little has been written on normative decision theory and second-order norms that deal with uncertainty about morality.

¹⁹ Zimmerman (2008) goes on to revise and reformulate the view; this is not the final formulation. However, these reformulations and revisions are not chiefly concerned with the regress problem.

²⁰ I have a minor bone to pick with this solution. Suppose at some order, I am uncertain in deciding between taking a particular risk-weighted expectation that is risk averse (call this *complicated risk aversion*) and taking the average of the minimum possible value and the expectation (given by theories at one order below). Suppose I cannot resolve my uncertainty about these two views, and I cannot reason about the level higher. Then, according to this proposal, I ought to take the expectation over these views. However, while I might not be able to reason very conclusively about the next order of the regress, and I might not even be able to specify all of the plausible options and their entailments, it does seem as if I should *not* take the expected value. Indeed, I am quite sure in this case that some sort of risk aversion is in order. This is easy to address: we can take something closer to an agent’s actual credence, such as the theory they find most probable at the highest order at which that they can know this through deliberation.

view is merely a refinement. I leave this interpretation to the side for the rest of this section.

The proposal that we must reason as far as we can is too demanding in some cases. Perhaps it is reasonable to say Ada (in example 4.1.2) should deliberate for a long while on the value of different generations. Hers is an important decision that may substantially affect millions of people today or many more millions in the future. Furthermore, hard trade-offs must be made, and these should not be taken lightly. But consider the following example, in which I do not think so much deliberation is in order.

Example 4.2.1 (Take a Penny, Leave a Penny). Baxter is walking along a street. They are on their way to an important presentation. Baxter is running exactly on time. While walking along the street, Baxter spots a five-cent piece. They can choose to pick up the piece or leave it and continue on their way.

It is not obvious to me what Baxter should do.²¹ If you think it is obvious that they should take the coin, consider that the money is only a small benefit and likely not worth the effort. Consider also that it is largely a distraction from the important task at hand: getting to the meeting on time. If you think it obvious that they should leave the coin, consider that it would likely only take a second or two to pick it up, which may translate into an hourly rate of \$90 per hour. There are considerations on both sides.

Regardless of how he will resolve this uncertainty upon reflection, it is obvious to me that he should not spend a long time reflecting. But according to the “reason as far as you can” principle, Baxter should think hard about whether to take the coin; he should reason up the regress as far as he can. Imagine Baxter does so; perhaps he takes a seat in the gutter beside the five-cent piece and reasons until he is red in the face. Now imagine that as a result he shows up to the important meeting late and with little mental energy for the presentation. Baxter’s decision seems ludicrous. I maintain that while it may be unclear whether he should take the coin, he should certainly not spend his time and energy reasoning as much as he possibly can. The fact that he has an important meeting to attend adds weight to the conclusion that this view is too demanding.

²¹ Many readers have the intuition that it *is* obvious (one way or the other). I think that they are jumping the gun, but I strongly endorse their personal policy of not thinking at all about cases like this, not because the answers are obvious but because they are not usually worth the time it would take to consider them in detail.

Here's another case that offers a similar intuition but has a different underlying structure.

Example 4.2.2 (Muddy Shoes). Peter is walking along the bank of a shallow river when he spots a drowning child. Peter knows he could wade in and save the child, who is seconds away from dying. He also knows that doing so would ruin his brand-new \$300 shoes.

In this case, the situation is important and the cost of delay is enormous.²² The child will die if Peter delays to reason as far as he can. To me, this is a choice situation that should obviously not be considered deeply. Peter should save the child. So what if his shoes get dirty?

Of course, there are other cases in which we should deliberate (example 4.1.2 is one). The amount of deliberation required varies from situation to situation. My own solution, given in §4.3, will be sensitive to this.

There may be another, secondary, issue with “reason as far as you can”: it does not solve the problem for *unbounded agents*. A cognitively unbounded agent is not limited in the orders of regress about which they can compute. Whatever mathematically or logically follows from their beliefs can be known to them through computation.²³ Computation for such agents has no limit. If there were a rational stopping point to their deliberation, they could adopt their expectation over norms of the corresponding order, but “reason as far as you can” might not provide such a stopping point. For unbounded agents, then, this principle appears to offer no solution.²⁴ However, I will discuss unbounded agents in §4.4 and show that (surprisingly) “reason as far as you can” may in fact be a solution even for unbounded agents, though here my own solution gives more plausible results even for unbounded agents.

4.2.2 Fixed Points and Convergent Solutions

Another class of solutions show that evaluations given by norms converge; that is, they reach a fixed point.

²² This case is inspired by Singer (1972).

²³ Specifically, I mean what is computable from these. As Gödel's incompleteness theorem shows, not all logical consequences are accessible even to these unbounded agents (Gödel, 1931).

²⁴ Assuming they are, as a matter of psychological fact, uncertain at every order of the regress.

Consider norms that give a fine-grained categorisation of acts by assigning values to acts available to an agent.²⁵ Then, at each order of the regress, the norms of that order assign values to acts which take into account evaluations by norms at lower levels.²⁶ Converging to X means that the values arrived at by norms of higher order than k will be arbitrarily close to X .²⁷ It is natural to think that if these values converge to X , the value of an act is X , and if the values converge in finitely many orders, considering further orders will not change your decision and so you should assign X as the value of the act. If the convergence is not finite, then the value gets closer and closer to X , and so we can defer to our infinitely more considered position, which assigns the value of X to that act.²⁸

A fixed-point solution is similar but more general. At each order k we get a value range for each act given by the evaluations of norms $_k$.²⁹ If there is a unique value X at the infinite intersection of these ranges, then X can be considered the value of that act, so long as we believe there is an all-things-considered, or '*supersubjective*',³⁰ ought which each norm makes claims about or if we accept Trammell's (2021) dominance principle.

Sepielli (2017) and Tarsney (2017) hope for convergent solutions to the regress problem, while Trammell (2021) generalises to fixed-point solutions and proves the conditions under which these solutions are guaranteed. The conditions include that an agent only has positive belief in a finite number of normative theories at any meta-level, where a normative theory is something like a collection of consistent norms. These theories must all be complete, continuous, cardinal, and compromising, and we must also endorse the analogue principle.³¹

25 I actually think these ideas *loosely* generalise to norms that cannot be represented as assigning values to acts. For example, you might think there is a k such that the act recommended by all norms at orders above k are the same (which does not require value assignments). However, such generalisations have not been explored.

26 This *taking into account* could involve combination, as in the case of taking an expectation over lower-order norms, or it might not, as in the case of accepting the lower-order norm you think most likely to be true.

27 Specifically, you will be able to find some such k for any specified arbitrary distance ϵ .

28 This pattern, in the case of non-uniform or extremely slow convergence, will not be clear unless you stand back.

29 See Trammell (2021) for details, especially the discussion of the difference between fixed-point and convergent solutions in §4.2.

30 In the language of Hedden (2016).

31 Each theory must be *complete*; that is, it must give well-defined (k -level) choiceworthiness claims across the entirety of an agent's uncertainty at each level below. Theories must be *continuous across meta-levels*, in the sense that small changes to choiceworthiness claims at any level result in only small changes in choiceworthiness claims at the level above. Further, we must assume theories, in general, are *cardinal*: there is a universal cardinal choiceworthiness scale. We must

In many cases, people do not hold beliefs that satisfy these tidy assumptions. Although these conditions can be weakened for transfinite convergence, they remain strong and entail the additional assumption that an agent's credences are well specified over transfinitely many orders of uncertainty in norms. Further work may find weaker conditions under which convergence or fixed-point solutions are guaranteed, and I would be happy to see this work. But for now this remains a major concern.

Even when these conditions are met, this theory seems excessively demanding.³² Consider again Baxter's decision of whether to pick up a five-cent piece (example 4.2.1). Should Baxter really be required to calculate the limit of an infinite sequence of norms at various orders or to determine where his higher-order normative considerations reach a fixed point? I think not. In the case of Peter (example 4.2.2), fixed-point solutions perform even worse: Peter will miss his chance to save the child if he calculates the fixed-point solution to resolve his normative uncertainty. This demandingness critique is similar to my critique of "reason as far as you can"; so I will not repeat the point. Suffice it to say Baxter might not know whether his mental state entails a fixed-point solution, and so this proposal may require him to reason further than he can.

4.3 WHEN FORCED, DO YOUR BEST

My own solution to the regress problem draws on the intuition that "reason as far as you can" indicates a hard upper limit to our reasoning but is entirely too demanding. Only sometimes should we reason as far as we can. Including deliberation as one of our options reveals that we always face forced choices, and when we do, we should do our best. I call this principle "when forced, do your best". It allows agents to make decisions despite the regress.

further assume that the agent believes in the same meta-level theories at every level. Trammell (2021) calls this the *analogue principle*. Finally, we must assume that an agent only has credence in *compromising* theories; that is, for a given range of possible values at some level, an agent only has positive credence in theories above that level that assign choiceworthiness judgements strictly within that range. For a critique of the analogue principle, see chapter 2.

³² Another approach might be to calculate what the fixed-point solutions actually are for common situations (or decision structures), then develop approximate decision rules based on the results. These results could say "In situations with features X,Y,Z,..., an agent (usually) ought to φ ". (I thank Matthew Nestor for this suggestion.) However, such an approach would still be limited by the assumptions necessary for fixed-point solutions to get off the ground, and thus much more work would be needed to apply heuristics to problems outside of this limited scope. (Taking this information into account when making some decisions does not necessarily go against the view I am advocating with "when forced, do your best".)

Consider forced choices. In a forced choice, we cannot deliberate for a long time. Even if we remain uncertain, we ought to do our best in the moment. For example, if you are considering whether to switch the train tracks so one person dies rather than five, you face a forced choice: if you “don’t choose”, then the choice is made for you. As the train lurches towards the poor souls tied to the tracks, you ought to do the best you can in the moment, even if your more informed self might choose differently.³³ I offer my principle as a way of deciding in forced choices.

Definition 4.3.1 (When Forced, Do Your Best). If an agent A is forced to decide between options, then A must choose the option O that appears best given their current considerations.

The principle above claims that when a choice is forced, we should do what we think is best given our current mental state, which includes beliefs, knowledge, and credences.

This may seem to not often be applicable and thus not be capable of providing a solution to the regress problem. Crucially, though, *deliberation is one of our options*. Consider Baxter (example 4.2.1). His choice is to pick up the five-cent piece or leave it. It is misleading, though commonplace, to characterise his option set as including just those two choices. There is of course a third option: he might deliberate further.³⁴ He could, for example, deliberate for five minutes and then choose. In fact, he faces a series of choices: take the coin, deliberate further, or leave the coin. When he picks further deliberation, he is again faced with a similar choice after deliberating. Baxter faced a forced choice all along; it just wasn’t obvious from the initial description of the example.

Notice that we always face a forced choice when our options are fully specified.³⁵ Not deciding is just another option. Applying this

³³ I do assume, however, that you can do your best. Initially, without deliberating, this may be a quick, intuitive, nonreflective judgement. If deliberation is in order, these judgements may be more reflective, deliberate, theory driven, and *slow* (in the sense of Kahneman (2011)). This avoids a regress that might begin with your uncertainty about whether deliberation is in order.

³⁴ Here, I mostly wish to focus on the regress you get from considering meta-level norms and meta-meta-level norms, but I take it that “deliberate further” could include gathering evidence, assessing evidence, thinking through alternative heuristics, and so on. Some of these alternate readings of deliberation may lead to a regress which I may be able to offer a solution to, but for now I leave such considerations to the interested reader. For my purposes, “to deliberate further” refers to the metanormative-reasoning kind of deliberation.

³⁵ Note that I have not completely specified their options here; for example, I have excluded waiting for five minutes, then choosing, and so on. But whatever their fully specified options, they will face a forced choice among *those*. We could also add a null option in which he stares off into the distance for a few minutes to decide or deliberate, but note that this option would be dominated by deciding and

fact to regress is problematic because it implies that in order to make a decision, we need to accept ever-higher-order norms. However, we do not need to consult every possible level in forced choices. We can simply accept our initial best guess. When we do so, we decide whether to deliberate further. If our initial best guess says to not deliberate, the regress stops here and we decide. Otherwise, we deliberate a while and then decide whether to continue deliberating or choose an “ordinary” option, again choosing the best we can given our current considerations. *Best* can be made more precise; perhaps it signifies the act we choose based on the norm we find most plausible at the highest order on which we have so far deliberated. Often it might not be as conscious as that, especially at the lowest levels. *Deliberation* too could be made more precise, and in ordinary cases it includes much more than ascending the regress (individuating options, finding out empirical matters, and so on), but we can leave aside these details.

My solution also gives plausible results in the examples discussed so far. Baxter should, I claim, deliberate about the five cents only if he thinks doing so will be worthwhile. However, it is obvious to me that any deliberation that lasts more than a few seconds is not worthwhile. The upper bound on the value of deliberation here is five cents because that is all Baxter has to gain. But the cost of deliberation is running late and arriving flustered at an important meeting—an outcome Baxter would likely pay more than five cents to avoid.³⁶ He should certainly not, as “reason as far as you can” implies, deliberate to his cognitive limits. Nor should he calculate a convergent or fixed-point solution to his normative uncertainty, even if he can. Clearly, my solution can account for the intuition that too much deliberation would be irrational here.

In Muddy Shoes, Peter could save the drowning child but would ruin his new shoes in the process. He might also deliberate about the decision. However, it is obvious that the child could drown at any moment and that the child’s life outweighs Peter’s concern for his own shoes; so it is obvious that a lengthy deliberation is much worse than saving the child. Peter should not be immobilised just because he *can* consider the decision further; he should save the child immediately.

deliberating. So Baxter faces a forced choice when his options are fully specified. Note too that if gathering evidence were included (as it would be in the real world, though I have excluded that consideration for simplicity), then the agent might theorise a small amount, then gather evidence, then ponder some more, then finally decide to go with what they are thinking at the time without further thought. The details would require fleshing out, but for now I am content in presenting my approach in broad strokes.

³⁶ Note that this can be known to Baxter even if he faces uncertainty about whether to take or leave the five cents at every level of the regress.

In Ada's environmental-policy case, she is uncertain of how to trade off the welfare of current and future generations, and she faces an important decision on the matter. Perhaps in this case she feels that the best option is an order of magnitude better than the next-best option. Then she should deliberate. Perhaps she should even reason as far as she possibly can. My principle allows plenty of deliberation where deliberation is likely to be valuable. It gives plausible results in the cases discussed so far, unlike the two proposals in the last section.

It is worth saying a little here about how the value of deliberation usually works, though further work is needed on the topic. The value of deliberation is bounded from above by the value of your best option. It's bounded from above more closely by the value of the difference between the best and worst options. Within these bounds, it is higher the more you can expect to reduce your uncertainty, so long as you can change your decision.³⁷ When the difference between our best and worst options is high but we know which is which, we can choose the best one without deliberating, so the value of deliberation is low relative to the value of our "ordinary" options. Similarly, when our best and worst options have similar value, we cannot gain a lot from choosing correctly, and so the value of deliberation is comparably low. Finally, if we do not expect to reduce our uncertainty, then our decision will be no better after deliberation, and the value of deliberation will be comparably low. (I mean this point to be broad enough to include *knowing how to evaluate our options despite more fundamental uncertainty*; so finding a fixed-point solution reduces our uncertainty about the supersubjective value of our acts without reducing our fundamental normative uncertainty.) All of this means that in many cases we need not deliberate much, and in others, deliberating will either reduce our broader uncertainty (thereby making further deliberation less attractive because there is less residual uncertainty to resolve) or make us expect further deliberation will resolve our uncertainty less than we previously thought (again, making further deliberation less attractive because we expect to resolve less of our uncertainty). In all cases, then, it seems the value of deliberation will gradually decline until we can choose an "ordinary" option following the principle "when forced, do your best." The value of deliberation will also fall off a cliff when we reach our cognitive limit, as Zimmerman (2008) suggested; we need not rely entirely on a broad-strokes understanding about when deliberation is valuable.³⁸

37 Further work is certainly required. See the explorations of information value under normative uncertainty by MacAskill, Bykvist, and Ord (2020, §9) and chapter 2.

38 One concern might be that some people believe that deliberation is *always* their best option. In response, I would say that the solution only requires that an

To sum up, I have given a *prima facie*–plausible principle that allows us to act rationally in forced-choice situations, whether or not we face a regress. I then argued that our fully specified decisions are always forced, thus dispelling the illusion that this principle might only sometimes apply (the illusion resulted from simplifying a forced choice with many options as an unforced choice between a few options). Other solutions did not do well with my examples, largely because they were too demanding.

In the next section, I extend my solution to unbounded agents and argue that under plausible assumptions they too can act rationally despite the regress.

4.4 UNBOUNDED AND ARBITRARY EVALUATION

My proposed solution relies on the intuition that deliberation can become too costly relative to its benefits—perhaps because at some point further deliberation tells us nothing (when we reach our biological limits), perhaps because despite our ability to deliberate further, the benefits are very low (as they were for Baxter), or perhaps because the costs are unreasonably high (as they were for Peter). My examples so far have focused on biological humans. However, what about idealised agents that do not obviously have such hard limits to their cognition? Unbounded agents are agents that have unlimited cognitive capacities. One kind of unbounded agent threatens my view significantly; another does not..

Definition 4.4.1 (Unbounded Agent 1). An unbounded agent, \mathcal{U}_1 , has the ability to instantaneously, and without cost of any kind, compute anything that is provable or deducible from their initial beliefs, knowledge, and other inputs.

These agents thus face absolutely no cost to deliberation. However, when we consider information processing of any kind, we realise there are costs associated with it. Most often these costs are measured in time (it takes a moment for neurons to activate in a pattern, for example) or energy (minds are instantiated as a physical process in which energy is converted and, because the process is not entirely efficient, some is lost). So \mathcal{U}_1 is physically impossible.³⁹

agent believe that the value of deliberation has certain properties and that these properties are plausible. For example, Baxter must grasp that deliberating for a long time will incur a large cost that is not compensated for by deciding correctly whether to pick up five cents. An agent who denies this is either missing the obvious or has such an odd set of values that it really is rational to deliberate as much as possible. Few of us are like this, but this is possible, and so “when forced do your best” is an important addition to “reason as far as you can”.

³⁹ I assume physicalism here. Perhaps the nonphysicalist would posit a substance of mind that allows costless information processing. I do not find it intuitive that

Let's consider more realistic agents.

Definition 4.4.2 (Unbounded Agent 2). An unbounded agent, \mathcal{U}_2 , has the ability to compute anything that is provable or deducible from their initial beliefs, knowledge, and other inputs. There is, however, a small cost associated with information processing in terms of time or energy.

\mathcal{U}_2 can reason infinitely but is minimally bounded by the available energy and time. Can a solution such as the one I've been advocating apply to an agent with such potential? I think it can.

Deliberating as far as possible is unlikely to be an agent's best option. In order to understand why, we must realise that a significant part of the value of deliberation comes from our ability to make better decisions. In Ada's policy decision, she might rationally choose to deliberate a lot because deliberation would allow her to make a better decision. Most of the value of scientific understanding comes from the power it gives us to bring about more of what we value intrinsically rather than from the intrinsic value of understanding the fundamental nature of the world. Likewise for deliberation in general: making a better decision is a key part of the value of deliberation per se ⁴⁰

Let's examine \mathcal{U}_2 with only a time cost. In such a scenario, if \mathcal{U}_2 deliberates as far as they possibly can, they miss out on any opportunity to interact with the world, thereby getting only the intrinsic value of deliberation and forfeiting much of the potential value they could receive by acting. Simply by having the ability to deliberate near costlessly, all they would have to do is convince humanity to adopt improvements that are obvious to more reflective agents—for example, accelerating scientific progress in general, or making progress safer, or influencing the direction of humanity so that they produce more of whatever is valuable. When we examine the costs and benefits of “always deliberating” and “mostly deliberating but occasionally doing particularly impactful bits of influencing humanity”, we get the decision in the following table.

Table 4.2 displays the two options of \mathcal{U}_2 as the rows, and the intrinsic and extrinsic (or use) value of those options. It should be clear from the table that if the agent always deliberates, they gain a slight amount of the intrinsic value of their decision. If, however, they choose to occasionally influence humanity using the knowledge that they have, they can produce a lot of value in the universe for

the same norms that apply to bounded agents should apply to \mathcal{U}_1 . While I have not solved the problem for them, I consider this a different problem from the one discussed here.

⁴⁰ This idea underpins much of the only current exploration of the value of information under normative uncertainty: MacAskill, Bykvist, and Ord (2020).

	Extrinsic Value	Intrinsic Value
Always Deliberate	Status Quo	Highest Possible Value
Mostly Deliberate	Very High Value	Very High Value

Table 4.2: Why not always deliberate? (1)

only a very slight decrease in intrinsic value. I tend to believe that the intrinsic value is a very small part of overall value, but in order to agree that “mostly deliberate” is the clear winner, you only need to think that lots and lots of extrinsic value is worth the very small sacrifice in intrinsic value of knowledge.⁴¹ In any case, \mathcal{U}_2 should do what they think is best, which is likely to involve occasional breaks from deliberation to influence humanity.⁴²

Now consider \mathcal{U}_2 with only an energy cost. Compare always deliberating with deliberating to some level k before transforming the universe into the best state possible according to \mathcal{U}_2 's beliefs at level k . Such a decision would look like table 4.3.

	Extrinsic Value	Intrinsic Value
Always Deliberate	Nil	Highest Possible Value
Deliberate to L	Very High Value	Very High Value

Table 4.3: Why not always deliberate? (2)

In the first example, it was clear that, unless the intrinsic value of knowledge is implausibly high, the agent should not always deliberate. Here, that result is stronger. If \mathcal{U}_2 deliberates as far as they can, they will burn the entire universe, leaving absolutely no extrinsic value.⁴³ Deliberating to almost any finite k and then acting seems more desirable.

⁴¹ This argument is further strengthened by the following consideration: if the extrinsic value of knowledge is thought to be incredibly high, then presumably the agent could help humanity design and build a giant supercomputer—perhaps many of them—that could process even more information for \mathcal{U}_2 , who only has finite time.

⁴² If, however, the intrinsic value of deliberation is infinitely high (though I find this possibility implausible), “when forced, do your best” correctly recommends always deliberating.

⁴³ Manheim and Sandberg (n.d.) argue that the value in the universe is finite unless our understanding of physics is wrong in fundamental ways.

As I noted earlier, I cannot give a convincing version of this story if we disregard the cost of information, as in the case of \mathcal{U}_1 . However, as I noted, such a case is physically impossible. I can sleep well with a solution that may fail in impossible cases. Considering how “reason as far as you can” does for \mathcal{U}_2 will be informative. The rule can solve the regress problem because there is a limit to how far an agent can ascend the regress, but claiming \mathcal{U}_2 ought to reason that far is implausible; but reasoning this far involves either a failure to take any of the many incredible opportunities this agent has to make the universe better or the destruction of all that matters extrinsically. A fixed-point solution may (if the assumptions are met) do better than “reason as far as you can” here, but it will do worse than my approach if the agent should decide *before* the fixed point.

4.5 CONCLUSION

In the introduction I argued that we need to not only solve the regress problem, but also to do so in a (plausibly) rational way.

We saw two potential existing solutions. One gives conditions under which our judgements would converge. Another uses the natural limits to our cognition as the obvious stopping point for deliberation. We could simply deliberate to our cognitive limit and then decide. I argued against both, largely on the grounds that they are too demanding; we need not reason to our cognitive limit when deciding whether to pick up some spare change.

I then offered my own solution. When we are forced to make a decision, it is permissible to simply pick the option we think is best given our current considerations. When we realised that our options include deliberating further up the regress, we found that this principle gives us an answer to how to act at *any* level of the regress we may find ourselves at. We could stop the regress where we cannot reason further, but we often can, and should, decide earlier.

I then addressed unbounded agents. One kind of unbounded agent deliberates costlessly. However, such agents are not physically possible. A second kind of unbounded agent incurs a small time or energy cost for deliberation but has no in-principle computational limit. This second kind of agent, like us, must stop at some maximum level and decide. We also have strong reason for thinking they ought to stop well before this point. Burning the universe in order to learn a little more seems an odd choice at best.

DIRECTIONS FOR FUTURE RESEARCH

It is natural at the end of an extended research project to feel as if you are just beginning to make progress. There are many research threads that could be pursued, and many opportunities to make advances on the topics covered.

The main area my thesis opens up is exploring information value under normative uncertainty (in chapter 2). This area is important because we often do face decisions under normative uncertainty and it is worth understanding how valuable it is to resolve this uncertainty. Because I am the first to give a model of this, it is important to note that that this is not the only possibility, and I encourage exploration and comparison of alternative models (including models of deciding under decision-theoretic uncertainty).

Within my model, several directions seem promising. For one, I feel as if the restrictions I made on decision theories (the way of comparing decision theories, in conjunction with the naturally bounded principle and the stochastic-dominance-over-signals principle) come close to ensuring information is valuable everywhere—precisely what additional assumptions you need to guarantee this would be an interesting result. For another, on my model information value is determined in part by how paying for information would restrict the set of available actions, but I largely ignored the dynamics of this,¹ which could have interesting implications.

The model could also be extended, for example to cases of moral or value uncertainty.²

Applying this model to stylized versions of the important decisions we face would also be a useful extension (as would be eliciting beliefs from experts on decision theory, in order to make these case studies more realistic). In chapter 4 I argued for a solution to the regress problem that relied on stylised facts about information value; further work on information value could also illustrate which class of models can account for these stylised facts or formalise my solution to the regress problem.

There is more work to do in regards to comparability. I explored moral-theoretic comparisons in chapter 3, and decision-theoretic comparisons in a section of chapter 2. I argued moral theories are

¹ Especially in the section on comparative value.

² See in particular MacAskill, Bykvist, and Ord (2020, §9) if you're interested in this direction.

only sometimes comparable, but more work is needed to understand in which cases we are able to compare moral theories, and in which cases we can't. Further, I would like to see the development of theories of moral uncertainty that are sensitive to these facts.³ I argued that plausible decision theories are comparable (under certain assumptions) and gave a procedure for this comparison. I did not give the exact features that make a decision theory comparable according to me, nor did I try to come up with a counterexample (a decision theory that is both plausible and not comparable via my procedure). These would be interesting developments.

In chapters 2 and 4 I argued against fixed-point and convergent solutions to the regress problems. These arguments might be partially avoided by further work, if that work found weaker conditions under which these solutions were guaranteed. Further, I argued that decision theories are not exactly analogous to higher-order decision theories (as these views implied), but did not give a positive view of higher-order decision theory; further work on this topic might be fruitful. Finally, I did not explore other regress problems that begin with either uncertainty in my solution or initial uncertainty about whether deliberation is the best act in the forced choice, which could be explored further.

Finally, very little work has been done with respect to *normative* decision theories that diverge from expected utility (including their moral-uncertainty counterparts), and I think future work on this area will be fruitful.⁴

3 A not particularly well-thought-through example would say that under conditions X and Y choose your favourite option (because you can't compare them); in case Z take your expectation over theories (because you can compare them). See MacAskill, Bykvist, and Ord (2020) and Tarsney (2020) for more thought-through examples.

4 This work could draw off the behavioural-economics literature, just as Buchak (2013) draws on Quiggin (1982) and Yaari (1987), and Bottomley and Williamson (n.d.) draws on Chew (1983, 1989).

BIBLIOGRAPHY

- F. Arntzenius** (2003). "Some Problems for Conditionalization and Reflection." In: *The Journal of philosophy* 100.7, pp. 356–370. ISSN: 0022-362X (cited on page 62).
- K. J. Arrow** (1951). *Social choice and individual values*. Monograph 12 / Cowles Commission for Research in Economics. N.Y: Wiley (cited on page 71).
- D. Bernoulli** (1954). "Exposition of a New Theory on the Measurement of Risk." In: *Econometrica* 22.1, pp. 23–36. URL: <http://www.jstor.org/stable/1909829> (cited on pages 2, 32, 37).
- J. Bernoulli** (1713). *Ars Conjectandi*. Basel (cited on page 39).
- (1975). *Die Werke von Jakob Bernoulli*. [Textkritische Ausg.] Vol. 3. Basel: Birkhauser. ISBN: 3764307137 (cited on pages 2, 32).
- D. Blackwell** (1953). "Equivalent Comparisons of Experiments." In: *The Annals of Mathematical Statistics* 24.2, pp. 265–272. DOI: [10.1214/aoms/1177729032](https://doi.org/10.1214/aoms/1177729032) (cited on pages 2, 39, 51, 52).
- D. Blackwell and M. Girshick** (1954). *Theory of games and statistical decisions*. Wiley publications in statistics. New York: Wiley (cited on pages 2, 51).
- N. Bostrom** (2009a). *Moral uncertainty – towards a solution?* URL: <https://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html> (cited on pages 2, 57).
- (2009b). "Pascal's mugging." In: *Analysis* 69.3, pp. 443–445. ISSN: 0003-2638 (cited on page 74).
- C. Bottomley and T. L. Williamson** (n.d.). "Reasonable Risk-Aversion: Good Things Come To Those Who Weight." [Version of June, 2021] (cited on pages 2, 36, 37, 102).
- D. Bourget and D. J. Chalmers** (2014). "What Do Philosophers Believe?" In: *Philosophical Studies* 170.3, pp. 465–500. DOI: [10.1007/s11098-013-0259-7](https://doi.org/10.1007/s11098-013-0259-7) (cited on page 65).
- R. Bradley and H. O. Stefansson** (2017). "Counterfactual Desirability." In: *British Journal for the Philosophy of Science* 68.2, pp. 485–533. DOI: [10.1093/bjps/axv023](https://doi.org/10.1093/bjps/axv023) (cited on page 17).

- S. Bradley and K. Steele** (2016). "Can Free Evidence Be Bad? Value of Information for the Imprecise Probabilist." In: *Philosophy of Science* 83.1, pp. 1–28. DOI: [10.1086/684184](https://doi.org/10.1086/684184) (cited on page 39).
- R. Briggs** (2015). "Costs of Abandoning the Sure-Thing Principle." In: *Canadian Journal of Philosophy* 45.5, pp. 827–840. DOI: [10.1080/00455091.2015.1122387](https://doi.org/10.1080/00455091.2015.1122387) (cited on pages 2, 34, 40).
- L. Buchak** (2010). "Instrumental Rationality, Epistemic Rationality, and Evidence-Gathering." In: *Philosophical Perspectives* 24.1, pp. 85–120. DOI: <https://doi.org/10.1111/j.1520-8583.2010.00186.x> (cited on pages 2, 40, 45).
- (2013). *Risk and rationality*. Oxford University Press. ISBN: 978-0-19-967216-5 (cited on pages 2, 33–35, 37, 40, 67, 102).
- (2015). "Revisiting Risk and Rationality: A Reply to Pettigrew and Briggs." In: *Canadian Journal of Philosophy* 45.5, pp. 841–862. DOI: [10.1080/00455091.2015.1125235](https://doi.org/10.1080/00455091.2015.1125235) (cited on pages 2, 34).
- (2016). "Why high-risk, non-expected-utility-maximising gambles can be rational and beneficial: the case of HIV cure studies." In: *Journal of Medical Ethics* 2, pp. 1–6 (cited on page 2).
- (2017a). "Precis of Risk and Rationality." In: *Philosophical Studies* 174.9, pp. 2363–2368. DOI: [10.1007/s11098-017-0904-7](https://doi.org/10.1007/s11098-017-0904-7) (cited on pages 2, 33).
- (2017b). "Replies to Commentators." In: *Philosophical Studies* 174.9, pp. 2397–2414. DOI: [10.1007/s11098-017-0907-4](https://doi.org/10.1007/s11098-017-0907-4) (cited on pages 2, 34).
- K. Bykvist** (2017). "Moral Uncertainty." In: *Philosophy Compass* 12.3, e12408. DOI: [10.1111/phc3.12408](https://doi.org/10.1111/phc3.12408) (cited on page 2).
- C. Camerer, G. Loewenstein, and M. Rabin** (2004). "Advances in Behavioral Economics." In: ISBN: 9780691116822 (cited on page 2).
- R. Carnap** (1947). "On the Application of Inductive Logic." In: *Philosophy and Phenomenological Research* 8.1, pp. 133–148. DOI: [10.2307/2102920](https://doi.org/10.2307/2102920) (cited on page 39).
- S. H. Chew** (1983). "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox." In: *Econometrica* 51.4, pp. 1065–92. URL: <https://EconPapers.repec.org/RePEc:ecm:emetrp:v:51:y:1983:i:4:p:1065-92> (cited on pages 2, 36, 102).

- (1989). “Axiomatic utility theories with the betweenness property.” In: *Annals of operations research* 19.1, pp. 273–298. ISSN: 0254-5330 (cited on pages 2, 36, 102).
- M. Coakley** (2016). “Interpersonal Comparisons of the Good: Epistemic not Impossible.” In: *Utilitas* 28.3, pp. 288–313. ISSN: 0953-8208 (cited on page 78).
- M. Colyvan** (2008). “Relative Expectation Theory.” In: *Journal of Philosophy* 105.1, pp. 37–44. DOI: [10.5840/jphil1200810519](https://doi.org/10.5840/jphil1200810519) (cited on page 74).
- M. Colyvan and A. Hájek** (2016). “Making Ado Without Expectations.” In: *Mind* 125.499, pp. 829–857. DOI: [10.1093/mind/fzv160](https://doi.org/10.1093/mind/fzv160) (cited on page 74).
- J. S. Demski** (1980). *Information analysis*. 2nd ed. Reading, Mass: Addison-Wesley. ISBN: 0201012316 (cited on pages 2, 39).
- M. Fleurbaey and S. Zuber** (2015). “Discounting, Risk and Inequality: A General Approach.” In: *Journal of Public Economics* 128. DOI: [10.1016/j.jpubeco.2015.05.003](https://doi.org/10.1016/j.jpubeco.2015.05.003) (cited on page 86).
- B. C. van Fraassen** (1984). “Belief and the Will.” In: *Journal of Philosophy* 81.5, pp. 235–256. DOI: [10.2307/2026388](https://doi.org/10.2307/2026388) (cited on page 62).
- M. Friedman and L. J. Savage** (1952). “The Expected-Utility Hypothesis and the Measurability of Utility.” In: *Journal of Political Economy* 60. URL: <https://EconPapers.repec.org/RePEc:ucp:jpolec:v:60:y:1952:p:463> (cited on pages 2, 38).
- L. Fryxell** (n.d.). “A Theory of Experienced Utility and Utilitarianism.” [Version of November 2019]. URL: <https://lorenfryxell.com/WP/Fryxell-JMP-11-25-2019.pdf> (cited on pages 16, 46).
- K. Gödel** (1931). “Über Formal Unentscheidbare Sätze der Principia Mathematica Und Verwandter Systeme I.” In: *Monatshefte für Mathematik* 38.1, pp. 173–198 (cited on page 91).
- I. J. Good** (1967). “On the Principle of Total Evidence.” In: *The British Journal for the Philosophy of Science* 17.4, pp. 319–321. DOI: [10.1093/bjps/17.4.319](https://doi.org/10.1093/bjps/17.4.319) (cited on pages 2, 39).
- E. J. Gracely** (1996). “On the noncomparability of judgments made by different ethical theories.” In: *Metaphilosophy* 27.3, pp. 327–332 (cited on pages 2, 71, 72).
- P. A. Graham** (2010). “In Defense of Objectivism About Moral Obligation.” In: *Ethics* 121.1, pp. 88–115 (cited on page 68).

- H. Greaves** (2017a). "Discounting for Public Policy: A Survey." In: *Economics and Philosophy* 33.3, pp. 391–439. DOI: [10.1017/S0266267117000062](https://doi.org/10.1017/S0266267117000062) (cited on page 86).
- (2017b). "Population Axiology." In: *Philosophy Compass* 12.11, e12442. DOI: [10.1111/phc3.12442](https://doi.org/10.1111/phc3.12442) (cited on page 86).
- H. Greaves and O. Cotton-Barratt** (2019). "A bargaining-theoretic approach to moral uncertainty." In: *The Global Priorities Institute Working Papers*. URL: <https://globalprioritiesinstitute.org/a-bargaining-theoretic-approach-to-moral-uncertainty/> (cited on pages 2, 57, 67, 69, 85).
- H. Greaves and T. Ord** (2017). "Moral Uncertainty About Population Axiology." In: *Journal of Ethics and Social Philosophy* 12.2, pp. 135–167. DOI: [10.26556/jesp.v12i2.223](https://doi.org/10.26556/jesp.v12i2.223) (cited on pages 2, 71, 74, 86).
- J. E. Gustafsson and O. Torpman** (2014). "In Defence of My Favourite Theory." In: *Pacific Philosophical Quarterly* 95.2, pp. 159–174 (cited on pages 2, 67, 70, 72, 85).
- E. Harman** (2015). "The Irrelevance of Moral Uncertainty." In: *Oxford Studies in Metaethics, Volume 10*. Ed. by R. Shafer-Landau. Oxford University Press. Chap. 3. ISBN: 9780198738695. DOI: [10.1093/acprof:oso/9780198738695.001.0001](https://doi.org/10.1093/acprof:oso/9780198738695.001.0001) (cited on pages 2, 21, 65, 84).
- R. Harris** (n.d.[a]). "Defending Trade-Off Approaches to Moral Uncertainty" (cited on pages 66, 70, 74).
- (n.d.[b]). "Ethical Pluralism as a Practical Solution to Moral Uncertainty" (cited on pages 70, 74).
- D. M. Hausman** (1995). "The Impossibility of Interpersonal Utility Comparisons." In: *Mind* 104.415, pp. 473–490. DOI: [10.1093/mind/104.415.473](https://doi.org/10.1093/mind/104.415.473) (cited on page 75).
- B. Hedden** (2016). "Does MITE Make Right?" In: *Oxford Studies in Metaethics, Volume 11*. Ed. by R. Shafer-Landau. Oxford University Press. Chap. 5. DOI: [10.1093/acprof:oso/9780198784647.003.0005](https://doi.org/10.1093/acprof:oso/9780198784647.003.0005) (cited on pages 2, 21, 71, 92).
- J. L. Hudson** (1989). "Subjectivization in Ethics." In: *American Philosophical Quarterly* 26.3, pp. 221–229 (cited on pages 2, 71).
- F. Jackson** (1991). "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection." In: *Ethics* 101.3, pp. 461–482. DOI: [10.1086/293312](https://doi.org/10.1086/293312) (cited on pages 68, 84).

- R. C. Jeffrey** (1965). *The Logic of Decision*. University of Chicago Press (cited on pages 2, 32).
- J. Joyce** (1998). "A Nonpragmatic Vindication of Probabilism." In: *Philosophy of Science* 65.4, pp. 575–603. DOI: 10.1086/392661 (cited on page 17).
- (2010). "A Defense of Imprecise Credences in Inference and Decision Making¹." In: *Philosophical Perspectives* 24.1, pp. 281–323. DOI: 10.1111/j.1520-8583.2010.00194.x (cited on page 17).
- (2017). "Commentary on Lara Buchak's Risk and Rationality." In: *Philosophical Studies* 174.9, pp. 2385–2396. DOI: 10.1007/s11098-017-0905-6 (cited on pages 2, 34).
- J. Kadane, M. Schervish, and T. Seidenfeld** (2008). "Is Ignorance Bliss?" In: *The Journal of Philosophy* 105, pp. 5–36. DOI: 10.5840/jphil200810518 (cited on page 39).
- D. Kahneman** (2011). *Thinking, fast and slow*. 1st ed. New York: Farrar, Straus and Giroux. ISBN: 9780374275631 (cited on page 94).
- K. Keasey** (1984). "Regret Theory and Information: A Note." In: *The Economic Journal (London)* 94.375, pp. 645–648. ISSN: 0013-0133 (cited on pages 2, 39).
- J. M. Keynes** (1921). *A Treatise on Probability*. Dover Publications (cited on page 39).
- D. Lawrence** (1979). "The quantification of the value of information in decision making." PhD. Iowa State University. URL: <https://lib.dr.iastate.edu/rtd/7287/> (cited on pages 2, 15).
- S. A. Liguori** (1785). *Theologia Moralis*. 9th ed. (cited on page 1).
- B. Lipman** (1991). "How to Decide How to Decide How to...: Modeling Limited Rationality." In: *Econometrica* 59.4, pp. 1105–25. URL: <https://EconPapers.repec.org/RePEc:ecm:emetrp:v:59:y:1991:i:4:p:1105-25> (cited on pages 22, 41).
- T. Lockhart** (2000). *Moral Uncertainty and its Consequences*. Vol. 111. Oxford University Press (cited on pages 2, 67, 75).
- W. MacAskill** (2013). "The Infectiousness of Nihilism." In: *Ethics* 123.3, pp. 508–520 (cited on pages 13, 74).
- (2016a). "Normative Uncertainty as a Voting Problem." In: *Mind* 125.500, pp. 967–1004. DOI: 10.1093/mind/fzv169 (cited on pages 2, 67, 72).

- W. MacAskill** (2016b). "Smokers, Psychos, and Decision-Theoretic Uncertainty." In: *Journal of Philosophy* 113.9, pp. 425–445 (cited on pages 2, 16).
- W. MacAskill, K. Bykvist, and T. Ord** (2020). *Moral Uncertainty*. Oxford University Press. ISBN: 9780198722274. URL: <https://www.moraluncertainty.com/> (cited on pages 1–3, 12, 13, 16, 40, 41, 46, 62, 65–67, 70, 71, 74, 76, 85, 87, 96, 98, 101, 102).
- W. MacAskill, O. Cotton-Barratt, and T. Ord** (2020). "Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons." In: *The Journal of Philosophy* 117, pp. 61–95. DOI: [10.5840/jphil202011725](https://doi.org/10.5840/jphil202011725) (cited on pages 2, 6, 16, 42, 44, 67, 75).
- W. MacAskill and T. Ord** (2018). "Why Maximize Expected Choice-Worthiness?" In: *Noûs*. DOI: [10.1111/nous.12264](https://doi.org/10.1111/nous.12264). (Visited on Apr. 2, 2019) (cited on pages 2, 67, 68, 72, 83, 84).
- W. MacAskill, A. Vallinder, et al.** (2021). "The Evidentialist's Wager." In: *Journal of Philosophy* 118.6, pp. 320–342. DOI: [10.5840/jphil2021118622](https://doi.org/10.5840/jphil2021118622) (cited on page 2).
- D. Manheim and A. Sandberg** (n.d.). "What is the Upper Limit of Value?" [Version of January 2021]. URL: <https://philpapers.org/rec/MANWIT-6> (cited on page 99).
- A. Mas-Colell, M. Whinston, and J. Green** (1995). *Microeconomic Theory*. Oxford University Press. URL: <https://EconPapers.repec.org/RePEc:oxp:obooks:9780195102680> (cited on page 71).
- E. F. McClennen** (1990). *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press (cited on page 40).
- B. de Medina** (1577). *Expositio in primam secundae angelici doctoris D. Thomae Aquinatis* (cited on page 1).
- J. F. Mertens and S. Zamir** (1985). "Formulation of Bayesian analysis for games with incomplete information." In: *International Journal of Game Theory* 14.1, pp. 1–29. DOI: [10.1007/BF01770224](https://doi.org/10.1007/BF01770224) (cited on pages 22, 41).
- D. Miller** (1994). *Critical Rationalism: A Restatement and Defence*. Open Court (cited on page 39).
- S. Moss** (2016). *Probabilistic Knowledge*. Oxford University Press (cited on page 14).
- B. Mundy** (1987). "The Metaphysics of Quantity." In: *Philosophical Studies* 51.1, pp. 29–54. DOI: [10.1007/BF00353961](https://doi.org/10.1007/BF00353961) (cited on page 76).

- Y. Nakamura** (1995). "Probabilistically sophisticated rank dependent utility." In: *Economics letters*. *Economics Letters* 48.3, pp. 441–447 (cited on pages 2, 35).
- New Catholic encyclopedia*. (2003). 2nd ed. Detroit, Michigan: Gale Group in association with the Catholic University of America. ISBN: 0787676942 (cited on page 1).
- T. Newberry and T. Ord** (2021). "The Parliamentary Approach to Moral Uncertainty." In: *Technical Report #2021-2, Future of Humanity Institute, University of Oxford*. URL: <https://www.fhi.ox.ac.uk/wp-content/uploads/2021/06/Parliamentary-Approach-to-Moral-Uncertainty.pdf> (visited on July 15, 2021) (cited on pages 2, 57, 85).
- I. Nissan-Rozen** (2015). "Against Moral Hedging." In: *Economics and Philosophy* 3, pp. 1–21 (cited on page 71).
- D. Parfit** (1984). *Reasons and Persons*. Oxford University Press (cited on pages ix, 6).
- (2011). *On What Matters*. Oxford University Press (cited on pages 83, 84).
- B. Pascal** (1657). *Lettres provinciales* (cited on page 1).
- R. Pettigrew** (2015). "Risk, Rationality and Expected Utility Theory." In: *Canadian Journal of Philosophy* 45.5-6, pp. 798–826. DOI: [10.1080/00455091.2015.1119610](https://doi.org/10.1080/00455091.2015.1119610) (cited on page 34).
- (2016). *Accuracy and the Laws of Credence*. Oxford University Press UK (cited on pages 17, 18).
- E. Posner and E. Weyl** (2015). "Voting Squared: Quadratic Voting in Democratic Politics." In: *Vanderbilt law review* 68, pp. 441–500 (cited on page 67).
- J. Quiggin** (1982). "A theory of anticipated utility." In: *Journal of economic behavior & organization*. *Journal of Economic Behavior & Organization* 3.4, pp. 323–343. ISSN: 0167-2681 (cited on pages 2, 33, 67, 102).
- F. P. Ramsey** (1926). "Truth and Probability." In: *Philosophy of Probability: Contemporary Readings*. Ed. by A. Eagle. Routledge, pp. 52–94 (cited on page 32).
- (1990). "Weight or the Value of Knowledge." In: *British Journal for the Philosophy of Science* 41.1, pp. 1–4. DOI: [10.1093/bjps/41.1.1](https://doi.org/10.1093/bjps/41.1.1) (cited on pages 2, 39).

- J. Ross** (2006). "Rejecting ethical deflationism." In: *Ethics* 116.4, pp. 742–768 (cited on pages 74, 77).
- L. J. Savage** (1954). *The foundations of statistics*. New York: Wiley (cited on pages 2, 16, 32, 37, 38, 42).
- A. K. Sen** (1970). *Collective choice and social welfare*. Mathematical economics texts ; 5. San Francisco: Holden-Day. ISBN: 0816277656 (cited on page 75).
- A. Sepielli** (2009). "What to Do When You Don't Know What to Do." In: *Oxford Studies in Metaethics* 4, pp. 5–28 (cited on pages 2, 67).
- (2010). "'Along an imperfectly-lighted path': practical rationality and normative uncertainty." In: (cited on pages 1, 2, 67, 77).
- (2013). "Moral Uncertainty and the Principle of Equity among Moral Theories¹." In: *Philosophy and Phenomenological Research* 86.3, pp. 580–589 (cited on page 75).
- (2017). *How Moral Uncertainty Can Be Both True and Interesting*. Vol. 1. Oxford University Press. DOI: [10.1093/oso/9780198808930.003.0006](https://doi.org/10.1093/oso/9780198808930.003.0006) (cited on pages 2, 27, 65, 83, 87, 92).
- C. E. Shannon and W. Weaver** (1949). *The Mathematical Theory of Communication*. University of Illinois Press (cited on page 14).
- P. Singer** (1972). *Famine, Affluence, and Morality*. Oxford University Press USA (cited on page 91).
- C. Starmer** (2000). "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice Under Risk." In: *Journal of Economic Literature* 38, pp. 332–382. DOI: [10.1257/jel.38.2.332](https://doi.org/10.1257/jel.38.2.332) (cited on page 2).
- K. Steele and H. O. Stefánsson** (2020). "Decision Theory." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University (cited on pages 2, 17, 32).
- C. Tarsney** (2017). "Rationality and Moral Risk: A Moderate Defense of Hedging." PhD thesis. University of Maryland (cited on pages 2, 27, 67, 92).
- (2018a). "Intertheoretic Value Comparison: A Modest Proposal." In: *Journal of Moral Philosophy* 15.3, pp. 324–344 (cited on pages 2, 72).

- (2018b). “Normative Uncertainty and Social Choice.” In: *Mind*. DOI: [10.1093/mind/fzy051](https://doi.org/10.1093/mind/fzy051) (cited on pages 2, 67, 72).
- (2020). “Vive la Différence? Structural Diversity as a Challenge for Metanormative Theories.” In: *Ethics* 131.2, pp. 151–182. DOI: [10.1086/711204](https://doi.org/10.1086/711204) (cited on pages 2, 70, 74, 102).
- (n.d.). “Metanormative Regress: An Escape Plan.” [Version of June 2019]. URL: <https://www.dropbox.com/s/mqcr8simkfhto3c/Metanormative%5C%20Regress%5C%20-%5C%20An%5C%20Escape%5C%20Plan.pdf?> (cited on pages 26, 83).
- J. Thoma and J. Weisberg** (2017). “Risk Writ Large.” In: *Philosophical Studies* 174.9, pp. 2369–2384. DOI: [10.1007/s11098-017-0916-3](https://doi.org/10.1007/s11098-017-0916-3) (cited on page 34).
- P. Trammell** (2021). “Fixed-point solutions to the regress problem in normative uncertainty.” In: *Synthese* 198.2, pp. 1177–1199. DOI: [10.1007/s11229-019-02098-9](https://doi.org/10.1007/s11229-019-02098-9) (cited on pages 2, 15–18, 20, 22, 24, 26, 27, 37, 41, 54, 62, 86, 87, 92, 93, 113).
- (n.d.). “Weak Betweenness and Misleading Information.” [Version of September 2019]. URL: https://philiptrammell.com/static/weak_betweenness.pdf (cited on pages 2, 40).
- A. Tversky and D. Kahneman** (1992). “Advances in Prospect Theory: Cumulative Representation of Uncertainty.” In: *Journal of risk and uncertainty* 5.4, pp. 297–323. ISSN: 0895-5646 (cited on pages 2, 35).
- J. Von Neumann and Morgenstern** (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press (cited on pages 2, 17, 32, 37–39, 71, 74).
- P. Wakker** (1988). “Unexpected utility as aversion of information.” In: *Journal of Behavioral Decision Making* 1.3, pp. 169–175. DOI: [10.1002/bdm.3960010305](https://doi.org/10.1002/bdm.3960010305) (cited on pages 2, 38, 39).
- (1990). “Under stochastic dominance Choquet-expected utility and anticipated utility are identical.” In: *Theory and Decision: an international journal for multidisciplinary advances in decision sciences* 29.2, pp. 119–132. DOI: [10.1007/BF00126589](https://doi.org/10.1007/BF00126589) (cited on pages 2, 35).
- P. Wakker, I. Erev, and E. U. Weber** (1994). “Comonotonic Independence: The Critical Test between Classical and Rank-Dependent Utility Theories.” In: *Journal of risk and uncertainty* 9.3, pp. 195–230. ISSN: 0895-5646 (cited on pages 2, 40).
- B. Weatherston** (2014). “Running risks morally.” In: *Philosophical Studies* 167.1, pp. 141–163 (cited on pages 2, 21, 83, 87).

- H. Wilkinson** (forthcoming). "In Defence of Fanaticism." In: *Ethics*. URL: <https://globalprioritiesinstitute.org/hayden-wilkinson-in-defence-of-fanaticism/> (cited on page 74).
- D. Williams** (1991). *Probability with Martingales*. Cambridge: Cambridge University Press. ISBN: 052140455X (cited on page 17).
- E. G. Williams** (2015). "The Possibility of an Ongoing Moral Catastrophe." In: *Ethical Theory and Moral Practice* 18.5, pp. 971–982 (cited on page 65).
- M. E. Yaari** (1987). "The Dual Theory of Choice under Risk." In: *Econometrica* 55.1, p. 95. DOI: [10.2307/1911158](https://doi.org/10.2307/1911158) (cited on pages 2, 33, 102).
- M. J. Zimmerman** (2006). "Is Moral Obligation Objective or Subjective?" In: *Utilitas* 18.4, pp. 329–361. DOI: [10.1017/s0953820806002159](https://doi.org/10.1017/s0953820806002159) (cited on page 68).
- (2008). *Living with Uncertainty: The Moral Significance of Ignorance*. Cambridge University Press (cited on pages 2, 24, 26, 89, 96).

APPENDIX

A.1 FORMALISATION OF THE EXAMPLES IN §2.7

Trammell (2021) gives three principles, including the analogue principle, that are sufficient to guarantee a unique evaluation for every act (definition 2.2.10).¹ In order to formalise the examples in §2.7, we need to define several concepts.

Definition A.1.1. A k -theory, t_k , is *weakly risk avoiding* for a if $t_k(a) \leq \mathbb{E}v_{(k-1)}(a)$.

Definition A.1.2. A k -theory, t_k , is *weakly risk seeking* for a if $t_k(a) \geq \mathbb{E}v_{(k-1)}(a)$.

These definitions use the expectation of choiceworthiness as a natural reference point for risk neutrality and then define any theory to be risk avoidant when assigning values lower than the expectation, and risk seeking when assigning values higher than the expectation. Note that these definitions cover every act, whereas many of the theories we discussed in §2.3 vary in their risk avoidance across acts or across decision situations. Note also that these definitions only hold for theories whose evaluations are comparable in the sense of §2.4.

Definition A.1.3. A risk-avoidant k -theory t_k *cancels out* a risk-seeking k -theory t'_k if that agent's subjective-choiceworthiness value of a would not change if they instead believed with probability $p = p(t_k) + p(t'_k)$ in a theory $t''_k(a) = \mathbb{E}v_{(k-1)}(a)$ and assigned no positive probability to either t_k or t'_k .

Note that cancelling out is a symmetrical relation. For example, X cancels out Y iff Y cancels out X . We can generalise example 2.7.1 as follows.

Theorem A.1.1. Consider an agent that accepts the analogue principle and has initial beliefs such that at most one strictly risk-seeking/risk-avoiding theory is not cancelled out by any other theory for any act, a . $v(a) = t^*(a)$, where t^* is their most risk-seeking or most risk-avoidant theory.

¹ He weakens the conditions and shows that they still guarantee unique evaluation if we extend the hierarchy to transfinitely many levels of uncertainty, though I do not discuss this result.

Proof. We can break this down into three cases.

Case 1: All theories cancel out. In this case, v is equivalent to EU. I show this by replacing each k -theory with EU_k , and I note that each order of the hierarchy assigns acts the same single value: $\mathbb{E}u(a|s)$. Since no theory is not cancelled out, this satisfies the theorem.

Case 2: The most risk-seeking/risk-avoiding theory that is not cancelled out is strictly risk avoiding. Consider a sequence of minimum choiceworthiness values for an act. This sequence is constant at the number we will denote as $\underline{v}(a)$. Next, consider the sequence $\{\overline{v}_k(a) | k \in \mathbb{N}\}$ of maximal choiceworthiness evaluations. This sequence is given by the expectation at each level k over $(k-1)$ -choiceworthiness values. The sequence of maximal values must be decreasing, as $\overline{v}_k(a)$ is the weighted average of $\underline{v}(a)$ and some other numbers that are no larger than $\overline{v}_{k-1}(a)$. So this sequence is monotonic and decreasing while being bounded from below by the constant sequence of $\underline{v}(a)$; it therefore converges to $\underline{v}(a)$. Finally, note that the only admissible evaluation is $\bigcap_{k=1}^n [v_k(a), \overline{v}_k(a)] = \underline{v}_1(a) = \underline{v}(a)$. Thus, the agent evaluates options in accordance with their most risk-avoiding theory.

Case 3: The most risk-seeking/risk-avoiding theory that is not cancelled out is strictly risk seeking. The reasoning is the same as in case 2, but we can construct a sequence that converges to the upper bound on value given by the agent's most risk-seeking theory at each k . \square

Example 2.7.2 illustrates the idea behind the proof. Theorem A.1.1 only applies when the agent has up to one theory that is not cancelled out. This indicates that we could make the theorem more general by allowing an agent to replace any k -theory t_k with some set of k -theories that acted equivalently but cancelled out. Then this theorem would be applicable generally.²

Example 2.7.1 is an example of the following corollary.

Corollary A.1.1.1. Consider an agent who begins with initial beliefs such that all of their theories cancel out and then experiences an $\epsilon > 0$ increase in their belief in a theory t^* that is *everywhere* risk avoiding (or risk seeking). For that agent, v is equivalent to expected utility before the shift and equivalent to t^* after the shift.

² Our definition of cancelling out could be made more general if we allowed replacement of some theory or theories at level k with theories at another level; but given that we assume the analogue principle, this isn't necessary.

A.2 FORMALISATION OF EXAMPLE IN §2.8

Theorem A.2.1. We begin with an agent who initially has $\epsilon > 0$ credence in Maximax and $(1 - \epsilon)$ credence in EU at level 1 and who accepts the 2-theory that maximises expectations over 1-meta-level theories. $EU(a) = \mathbb{E}u(a|s)$, and $\text{Maximax}(a) = \max u(a|s)$. This agent has access to an information source, I , which may be interpreted as the collection of books partitioned by Q and indexed by $j = 1, \dots, m$, where each j is interpreted as an individual book. Consider the non-empty set X consisting of each partition q_j s.t. $p(\text{Maximax}|q_j) > p(\text{Maximax})$. Receiving the information contained in X would generate the choiceworthiness function v_X . An agent can choose X given any set of actions but must choose X if the action a^* that they would choose conditional on receiving the information in X varies in utility across states.

Proof. Let $\mathcal{P}(I)$ denote the power set of I . Note that the evaluation of an act according to Maximax is at least as great as the evaluation of that same act according to EU (Maximax says an act's value is its maximum possible value, after all). This and the construction of v_X imply that $v_X(a) \geq v_Y(a)$ for all choiceworthiness functions v_Y generated by receiving the information of some $Y \in \mathcal{P}(I)$, for all a . Further, $v_X(a) > v_Y(a)$ for all a s.t. $\exists s, s' \in \mathcal{S}$ which satisfy $u(a|s) < u(a|s')$, for all $Y \in \mathcal{P}(I)$. Thus, if $a^* = \arg \max_{a \in \mathcal{A}} v_X(a)$ is such that $\exists s, s' \in \mathcal{S}$ which satisfy $u(a^*|s) < u(a^*|s')$, then the agent must choose X . \square