# Choosing for Changing Selves

Richard Pettigrew

July 9, 2018

# Contents

I Aggregating selves 3

1 The problem of choosing for changing selves 4

2 The economists' orthodoxy: expected utility theory 9
- 2.1 Expected utility theory: an example . . . . . . . . . . . . . . . 9
- 2.2 Expected utility theory: the general case . . . . . . . . . . . . 11
- 2.3 Interpreting expected utility theory . . . . . . . . . . . . . . . . 15
- 2.4 Alternative decision theories . . . . . . . . . . . . . . . . . . . . 16
- 2.5 Expected utility theory and the problem of choosing for changing selves . . . . . . . . . . . . . . . . . . . . . . . . . . 18

3 Existing solutions I: the Unchanging Utility Solution 20

4 Existing solutions II: the Utility of Utility Solution 27

5 Existing solutions III: the One True Utility Solution 31
- 5.1 The Reflection Principle . . . . . . . . . . . . . . . . . . . . . . 33
- 5.2 The Weak Reflection Principle . . . . . . . . . . . . . . . . . . . 39
- 5.3 Permissivism and learning from others . . . . . . . . . . . . 41
- 5.4 The objectivist's version of our problem . . . . . . . . . . . . 49
- 5.5 The implausible strength of objectivism . . . . . . . . . . . . 52
- 5.6 Appendix: Proof of Theorem 5.1.1 . . . . . . . . . . . . . . . . 53

6 The Aggregate Utility Solution I: which attitudes to aggregate? 55
- 6.1 Aggregating preferences . . . . . . . . . . . . . . . . . . . . . . 57
- 6.2 Aggregating evaluation functions . . . . . . . . . . . . . . . . . 66
- 6.3 Aggregating credences and utilities . . . . . . . . . . . . . . . 77

7 The Aggregate Utility Solution II: the solution itself 82
- 7.1 The framework . . . . . . . . . . . . . . . . . . . . . . . . . . . 82

7.2 The decision rule . . . . . . . . . . . . . . . . . . . . . . . . . 84
7.3 The instability of preferences . . . . . . . . . . . . . . . . . . . 85
7.4 Responding to instability . . . . . . . . . . . . . . . . . . . . . . 94

**8 Can we compare utilities between different selves? 97**
8.1 Representing values with numbers . . . . . . . . . . . . . . . 100
8.2 Empathetic preferences . . . . . . . . . . . . . . . . . . . . . . . 104
8.3 The Zero-One Rule . . . . . . . . . . . . . . . . . . . . . . . . . . 105
8.4 The Fixed Points Solution . . . . . . . . . . . . . . . . . . . . . 107
8.5 The Fixed Interval Solution . . . . . . . . . . . . . . . . . . . . 110
8.6 The Matching Intervals Solution . . . . . . . . . . . . . . . . . 114

**9 Why aggregate using weighted averages? 115**
9.1 The Argument for Linear Pooling . . . . . . . . . . . . . . . . 118
9.1.1 The Principle of Minimal Mutilation . . . . . . . . . . 119
9.1.2 The Dominance Argument for Weighted Averages . . 121
9.2 The features of linear pooling . . . . . . . . . . . . . . . . . . 129
9.2.1 Arrow's conditions . . . . . . . . . . . . . . . . . . . . . 130
9.2.2 Aggregating credences I: Independence Preservation 131
9.2.3 Aggregating credences II: commuting with conditionalization . . . . . . . . . . . . . . . . . . . . . . . . . 136
9.2.4 Aggregating credences III: independence preservation revisited . . . . . . . . . . . . . . . . . . . . . . 140
9.3 Appendix: Proofs of Theorems 9.2.1 and 9.2.3 . . . . . . . . . 143
9.3.1 Proof of Theorem 9.2.1 . . . . . . . . . . . . . . . . . . 143
9.3.2 Proof of Theorem 9.2.3 . . . . . . . . . . . . . . . . . . 144

**10 Do we know enough to make decisions this way? 146**
10.1 The deliberative conception of decision theory . . . . . . . . 147
10.2 Paul's Utility Ignorance Objection . . . . . . . . . . . . . . . 148
10.3 Paul's Revelation Response . . . . . . . . . . . . . . . . . . . 150
10.4 The Fine-Graining Response . . . . . . . . . . . . . . . . . . . 152
10.5 Paul's Authenticity Reply . . . . . . . . . . . . . . . . . . . . 154
10.6 Moss' No Knowledge Reply . . . . . . . . . . . . . . . . . . . 157
10.7 Assessing Moss' No Knowledge Reply: the Paulian view . . 162
10.8 Assessing Moss' No Knowledge Reply: the independent view 165

**II Setting the weights** **173**

**11 The problem of weighting changing selves** **174**

**12 The weight of the past** **178**
12.1 Deathbed decisions . . . . . . . . . . . . . . . . . . . . . . . . . . 180
12.2 Past values and permissibility . . . . . . . . . . . . . . . . . . . 184
12.3 Past values and obligations . . . . . . . . . . . . . . . . . . . . 188
12.3.1 Three related principles . . . . . . . . . . . . . . . . . . 189
12.3.2 The libertarian objection . . . . . . . . . . . . . . . . . 192
12.3.3 Exchange rates . . . . . . . . . . . . . . . . . . . . . . . 194
12.3.4 Moral blackmail . . . . . . . . . . . . . . . . . . . . . . 196
12.3.5 Applying the Beneficiary Principle . . . . . . . . . . . 197
12.3.6 Obligations and reparations . . . . . . . . . . . . . . . 199
12.3.7 The intrapersonal version of the Beneficiary Principle 200

**13 Discounting the future** **202**
13.1 The epistemic defence of temporal discounting . . . . . . . . 204
13.2 Parfit's defence of temporal discounting . . . . . . . . . . . . 205
13.3 A formal framework for temporal discounting . . . . . . . . 206
13.4 Constraints on discounting functions . . . . . . . . . . . . . . 207
13.5 Ahmed's objection to Parfit . . . . . . . . . . . . . . . . . . . 211
13.6 Amending Ahmed's objection . . . . . . . . . . . . . . . . . . 213
13.7 Exploitability does not imply irrationality . . . . . . . . . . . 215
13.7.1 Time Consistency and Unexploitability . . . . . . . . 215
13.7.2 Changing preferences and changing selves . . . . . . 216
13.7.3 Equal weightings . . . . . . . . . . . . . . . . . . . . . 217
13.7.4 On the alleged badness of exploitability . . . . . . . . 219
13.8 Appendix: Proofs of Lemmas 13.4.1, 13.4.2, and 13.4.3 . . . . 224

**14 The nearer the dearer** **229**
14.1 The Stoic Objection . . . . . . . . . . . . . . . . . . . . . . . . 229
14.2 Proximity considerations . . . . . . . . . . . . . . . . . . . . . 232
14.3 Aneri's career . . . . . . . . . . . . . . . . . . . . . . . . . . . 235
14.4 Fernando's pension . . . . . . . . . . . . . . . . . . . . . . . . 236

**15 I'll be glad I did it—so, I'll do it** **239**
15.1 What is 'I'll be glad I did it' reasoning? . . . . . . . . . . . . . 240
15.2 Two problems with the reasoning . . . . . . . . . . . . . . . . 242
15.3 Deborah's choice . . . . . . . . . . . . . . . . . . . . . . . . . 247

15.4 Self-frustrating choices . . . . . . . . . . . . . . . . . . . . . . . 248

**16 The road ahead 249**

16.1 The Aggregate Utility Solution summarised . . . . . . . . . . 249

16.2 Questions for future selves . . . . . . . . . . . . . . . . . . . . . 251

16.2.1 Sensitivity to risk . . . . . . . . . . . . . . . . . . . . . 251

16.2.2 Imprecise credences and imprecise utilities . . . . . . 254

# Acknowledgments

This book has been forming gradually for four years now—for the first three, just as a series of notes and the occasional talk; in this final year, more as a tangible manuscript. I've benefitted enormously from discussing the issues in it with many different people and I fear I will inadvertently leave some out. I presented an early version of the problem to the Work in Progress seminar in Bristol, and discussed it afterwards with Havi Carel, Samir Okasha, Seiriol Morgan, Katie Monk, and Anya Farennikova. Much later, I presented some of my ideas about a solution to a departmental colloquium at Birmingham, and I am grateful to the audience there and in particular Hanna Pickard both then and later for helping me think through some of the ideas. I must thank: Helen de Cruz for inviting me to talk about this in the Royal Institute of Philosophy series held in Blackwell's bookshop in Oxford; Katherine Hawley for organising a wonderful workshop in St Andrews on transformative experience; Laurie Paul and her terrific students for hosting me at Chapel Hill in their seminar; Laurie again for organising the Ranch Metaphysics workshop in January 2018, where she allowed me to put up a substantial portion of Part I for discussion. At the ranch, I benefitted from the insights of Alex Jackson, Karen Bennett, Jennifer Lackey, Sarah Moss, Eric Swanson, Mike Titelbaum, Branden Fitelson, and many others—insights that were often shouted over the person's shoulder as we rode on horseback through the scrub north of Saguaro National Park. And thanks to Krister Bykvist, who has been a generous supporter and interlocutor throughout.

I received generous and detailed comments on a very rough draft of Part I from James Willoughby, Matt Kopec, and Mike Titelbaum. And Sarah Moss gave generous and brilliant comments on my treatment of her work in Chapter 10.

I owe my greatest debt of gratitude to Laurie Paul. We've been in extended discussion of all of the issues in this book nearly continuously since we first sat down together in Bristol in 2013 to discuss the manuscript of

what was to become her transformative book, *Transformative Experience*. Always motivated by an unflinching desire to get at the truth of the matter, there is no other philosopher from whom I've learned so much through discussion.

Finally, I must thank the generous Leverhulme Trust for funding the year of research sabbatical that has allowed me to write up these ideas and present them in their current form.

# Preface

A philosopher once said to me, whilst discussing the business of doing philosophy: 'We work on what we're bad at, on what we do poorly; we study the philosophy of whatever it is we find difficult.' They were only half joking. There are no doubt exceptions to this generalisation—ethicists who find it easy to do the right thing, competent artists who study aesthetics, or reliable reasoners who study epistemology. But I'm not one of them. For over a decade, I have struggled with the decision whether or not to adopt a child and become a parent. Part of what I find difficult about the decision is that I know it's possible, perhaps even likely, that if I were to become a parent, my values would change between now, as I deliberate over the decision, and the future, when the effects of the decision will be felt—when I sit waiting anxiously at the doctors the first time my child gets sick; and later on, when I comfort them as their first romance breaks down. For instance, there are many things about being a parent that currently I would dislike. To give just one example, if I were a parent I would have less time to spend with friends or on the sort of activities outside work that I value. But I've watched friends and colleagues become parents, and I've seen that, for some of them, what they value changes when their first child is born or when they first adopt. It's not that these people come to like their friends any less; it's just that they value time spent with their new child more than time spent with those friends. Spending time with a child for whom they are the parent has leapfrogged in their estimation from lying below spending time with friends to sitting above it. So I know it's possible that the values I would come to have were I to adopt would be different from the values I have now, and moreover that they would align better with the experiences I would have as a parent better than my current values would. So what am I to do?

Common sense, as well as the orthodox theory of rational choice studied by philosophers and economists, tells us that, when we decide rationally, we decide on the basis of what we believe about the world and what we value in the world. What are we to do then if what we value changes during the course of our life, either as a result of an intentional decision that we make—such as when we choose to become a parent—or just as the result of external factors beyond our control—the culture around us changes and we assimilate our values to it gradually over time? To which values should I appeal when I make a particular choice? My values at the time I face the decision? My values in the future, when the effects of the decision are being felt? My past values? Some amalgamation of all of these? That is the topic of this book—how should I choose on behalf of my changing selves?

While I have been trying to decide whether to adopt for many years, the philosophical treatment of the general problem of choosing for changing selves only came to my attention quite recently, when I read Edna Ullmann-Margalit's paper 'Big Decisions', Krister Bykvist's paper 'Prudence for Changing Selves', and L. A. Paul's book *Transformative Experience* (Ullmann-Margalit, 2006; Bykvist, 2006; Paul, 2014a). The title of the present book is a tip of the hat to Bykvist's paper, which comes closest to doing what I am trying to do.

# Part I

# Aggregating selves

# Chapter 1

# The problem of choosing for changing selves

This book is about how we should make decisions—it's about the rational way to choose what to do when we're faced with a range of options. We'll begin with some examples of the sort of decision that will concern us in what follows. In each of our examples, the choice that the decision-maker faces is a little out of the ordinary. The theory of decision we will eventually propose also covers much more quotidian decisions than these, such as whether or not to take an umbrella when you go for a walk, or which route to take to work. But it will have nothing new to say about such decisions; it will say exactly what our current best theory of decision already says. Where it will have something new to say is in the sort of cases of which the following cases are exemplars:

> **Aneri** is deciding between two career prospects: she has been offered a place on a training programme for new police officers; and she has been offered a position as an conservation officer for her local council. She is trying to decide which offer to accept. Aneri currently values conformity more than she values self-direction, but not much more. She knows that the conservation job provides some scope for self-direction, though not too much—on the whole, it involves following a series of protocols formulated by committees that she won't sit on. A police officer, on the other hand, has very little room for self-direction. If Aneri's values stay as they are, the conservation role will suit her well, while she will find the role of police officer frustrating.

But she also knows that a person's values tend to become 'socialised', at least to some extent—that is, people often take on values that mesh well with their jobs, or the cultures in which they live, or the groups of friends with whom they socialise most frequently. In particular, she knows that she will likely come to value conformity more than she does now if she trains for the police. And, if that's the case, she will not find it frustrating. Indeed, we might suppose that being a police officer will fit with her socialised values very slightly better than being a conservation officer will fit with her current values. Which career should Aneri choose?[1]

**Blandine** is also pondering her career. For years, she wanted to be a musician in a band. She always placed enormous value on the emotional side of life, and she wished to devote her career to exploring and expressing those emotions through music. Recently, however, she has abandoned this desire completely. She no longer wishes to be a musician, and no longer values the emotional side of life. Indeed, she is now committed to pursuing studies in particle physics. Some friends ask her to join a new band that they are putting together; and on the same day she receives an offer to study physics at university. Which path should Blandine choose?[2]

**Cheragh** is deciding whether or not to write the great novel that has been gestating in her imagination for five years. But she faces a problem. If she writes it, she knows she will come to have higher literary standards than she currently has. She also knows that while her own novel would live up to her current standards, it will not live up to these higher ones. So, if she writes it, she'll wish she'd never bothered. On the other hand, if she doesn't write it, she'll retain the same literary standards she has now, and she'll know her novel would have attained those standards. So, if she doesn't write it, she'll wish she had. Should Cheragh

[1] For related examples, see (Bardi et al., 2014; Bricker, 1980; Bykvist, 2006; Ullmann-Margalit, 2006; Paul, 2014a). We will discuss Aneri's case at many points throughout the book.

[2] For related examples, see Parfit's example of someone who always wanted to be a poet, but then changed their mind (Parfit, 1984); or Hare's example of someone who, as a young boy, wanted to be a train driver (Hare, 1989); for a discussion of the normative force of past preferences, see (Bykvist, 2003). We will discuss such cases in greater detail in Chapter 12.

write her book?[3]

**Deborah** has decided to have a baby, but she needs to decide when to try to become pregnant: now, or in three months' time. Currently, she has a virus, and she knows that, when people become pregnant whilst carrying this virus, their child will have an extremely high chance of developing a very aggressive cancer around the age of forty. However, if she becomes pregnant in three months' time, once her body is rid of the virus, there will be no risk to her child. Currently, she values having the child with the prospect of aggressive cancer very much less than she values having the child without. However, if she becomes pregnant now and has a child with that prospect, she will, most likely, form a bond with them so strong that she would value having that particular child, with their tragic prognosis, more than having any other child, including the child without that prognosis that she would have had if she had waited three months. After all, the alternative child would have been a different child, created from different gametes; they would not be the child with whom Deborah has formed the bond. When should Deborah try to become pregnant?[4]

**Erik** is contemplating an offer that his pension scheme is advertising. If he pays an extra £50 into the scheme this month, he will receive an all-expenses-paid trip to a white-knuckle, high-octane theme park when he is ninety years old, should he live that long. While he'd enjoy such a trip enormously now, he will probably not when he is ninety. Should he take up the offer?[5]

**Fernando**'s pension scheme is offering something rather different. If he opts in to their scheme, they will donate 10% of his pension payments to effective charities once he retires. If he opts in now, there is no way to reverse this decision—it is binding. Considering it now, he would like to do this. Fernando thinks it's important to give money to charity, particularly those that will use it effectively. However, he also knows that, when he re-

[3] For a related example, see Bykvist's example of someone contemplating marriage (Bykvist, 2006). We will discuss Cheragh's decision in more detail in Section 15.4.

[4] For related examples, see Harman's example of a young woman deciding whether or not to become pregnant (Harman, 2009); or Parfit on the non-identity problem (Parfit, 1984); see also (Paul, 2014a, 2015b). We will discuss Deborah's case in more detail in Chapter 15.

[5] We will discuss Erik's decision in Chapter 3.

> tires, his values will have changed and he'll prefer to give that money to his children, not to charity. Should he opt in to the scheme and bind himself to giving the money to charity?[6]
>
> **Giang** values seriousness in all things— in the literature he reads, the films he watches, the projects he pursues in his life, the careers he chooses, and the conversations he has with friends and colleagues. He doesn't value frivolity or fancy or whim at all. He can't understand how anyone else can value these things. He watches friends joking amongst themselves, and he knows that he could join in and even enjoy it—he's quite witty, in a dry sort of way. But he also thinks it would be a waste of time—there are more serious things to discuss. But, as well as all this, he doesn't want to be the way he is. Now, he certainly doesn't want to give up his preference for seriousness, but he'd like to let in frivolity a little more—he'd like to value seriousness slightly less than he does at the moment and frivolity a little more. His friend Gail seems to strike the right balance—Giang would like to emulate Gail. And if he hangs out with her more, he feels sure he will. His preferences will gradually migrate to match Gail's. Should Giang hang out with Gail more?

What these examples share in common is that, for the person making the decision, what they value or desire or enjoy or dislike might change throughout the course of their life in ways that seem relevant to the decision. This might happen as a result of a decision they make. Deborah's decision to become pregnant at one time rather than another will determine which of several different sets of values she will have—whether she values having *this* child or *that* child more. And Cheragh's decision to write her novel will lead to her values changing, as will Aneri's decision to pursue a career as a police officer.[7] Or, a person's values might change as a result of external factors, such as a change in the ideologies that dominate in the culture in which they live, or because of experiences they have that are not of their choosing—the experience of receiving a terminal or chronic diagnosis, for instance, can lead a person to change their values, but they do not choose

[6] For a related example, see Parfit's Russian nobleman case (Parfit, 1984). We will discuss Fernando's decision in Section 14.4.

[7] Decisions of this sort are particularly the subject matter and focus of Edna Ullman-Margalit's treatment of this topic, as well as Krister Bykvist's and L. A. Paul's (Ullmann-Margalit, 2006; Bykvist, 2006; Paul, 2014a).

to have this experience.[8] Alternatively, a person's values might change as a result of simple developments in their outlook and character as they move through life: Erik expects to move naturally away from valuing excitement and risk; Fernando anticipates that he will shift from wishing to donate to charity to wishing to preserve his children's inheritance; and Blandine experiences a sudden change from would-be musician to would-be particle physicist that is not occasioned by any choice she makes.

When a person's values have changed in the past or might change in the future, this poses a problem for decision making. After all, we ought to make our decisions on the basis of what we believe about the world and what we value in the world. Suppose, for instance, that I am deciding whether or not to take an umbrella when I go for a walk. Then my decision should depend on how likely I think it is that it will rain, but also on how much I value staying dry if it does rain, how much I value being unencumbered when I'm walking, and so on. Or suppose I am trying to decide which route to take to work. My decision should depend partly on how likely I think it is that each route has various features—it is quiet, or quick, or quaint—and partly on how much I value those features. But if rationality requires that we make our decisions based on what we believe about the world and what we value in the world, then we face a puzzle when what we value changes over time. To which values should we appeal when we make our decision? Those to which we are committed at the time we make the decision? Those we will have when the main effect of the decision is felt? The most enduring, which we have held or will hold for the longest time? Perhaps some amalgamation of all of our values, past, present, and future, some given greater weight than others? But, if this, how should we determine the weights? This is the central question of this book. How should we choose for changing selves?[9]

Hopefully, this gives a sense of our problem in an informal context. Throughout the book, however, we will pursue it in a particular formal context. The orthodox formal theory of rational decision making is expected utility theory, and we will work primarily in that in what follows. However, in Chapter 16, we will consider how our solution to the problem of choosing for changing selves might be adapted to alternative theories.

[8] See (Carel et al., 2016).

[9] Of course, you might be thinking that it isn't just what I value in the world that changes over the course of my life—what I believe about the world changes too. If changes in value pose a problem, why don't changes in belief? Hold that thought! We'll discuss this at length in Chapter 5.

# Chapter 2

# The economists' orthodoxy: expected utility theory

In this chapter, I describe the orthodox theory of decision-making—expected utility theory—and explain how the problem of choosing for changing selves arises for this theory.[10]

## 2.1 Expected utility theory: an example

Consider the following example of a decision problem: I have been learning to drive for some time. My test is only four weeks away. Should I practise or not? Intuitively, which choice I should make depends on a number of factors. Firstly, there are my beliefs: how likely I think I am to pass if I practise, and how likely if I don't. Second, there are my values: how much do I value the various situations that might results from the acts before me—the situation in which I practise and pass, in which I practise but fail, in which I don't practise but pass all the same, and in which I don't practise and I fail. Other things being equal, the more strongly I believe I'll pass if I practise, then more I'll lean in favour of practising. Other things being equal, the more I value situations in which I don't practise, the more I'll lean in favour of not practising. And so on.

Orthodox decision theory makes this intuitive approach precise in the following way. There are two acts between which I must choose: I practise (*Practise*) or I don't (*Don't Practise*). And there are two possible states of the world that I care about: I get my license (*License*) or I don't (*No License*). To

[10] For alternative introductions to expected utility theory, see (Joyce, 1999; Briggs, 2017).

choose between *Practise* and *Don't Practise*, I rate both of these options as means to my ends and I pick the one I rate most highly—or, if I rate them equally, I am free to pick either. I record my ratings for the options in my evaluation function $V$, which assigns to each option my estimate of how good that option is as a way of getting me what I want. Thus, $V(\textit{Practise})$ is the number that measures my subjective assessment of practising as a means to my ends, while $V(\textit{Don't Practise})$ does the same for not practising.

Let's first consider *Practise*. To evaluate it as a means to my ends, I begin by asking how much I value the outcome in which I choose *Practise* and I receive my license (that is, *License*), and how much I value the outcome in which I choose *Practise* and I do not receive my license (that is, *No License*). Let's begin with the state of the world, *License*, in which I receive my driving license. How much do I value the outcome in which I receive my license having practised for my driving test? That is, how much do I value *Practise* & *License*, which says that *License* is true and I performed *Practise*. I measure how much I value this outcome, all things considered, and I record it in my current utility function $U$. Thus, $U(\textit{Practise} \,\&\, \textit{License})$ is the real number that measures the extent to which I value *Practise* & *License*, all things considered.[11] And similarly for the state *No License*: my utility for the outcome *Practise* & *No License*, which we write $U(\textit{Practise} \,\&\, \textit{No License})$, measures how much I value the outcome in which I practise for my test but do not receive my license.

Now, according to expected utility theory, my evaluation $V(\textit{Practise})$ of the option in which I practise for my test is given by my *subjective expectation* of the utility of practising. That is, $V(\textit{Practise})$ is the weighted average of $U(\textit{Practise} \,\&\, \textit{License})$, the utility that I assign to practising and receiving my license, and $U(\textit{Practise} \,\&\, \textit{No License})$, the utility I assign to practising and not receiving my license, where the weights are given by my credences on the supposition that I practise. That is, I weight $U(\textit{Practise} \,\&\, \textit{License})$ by my credence, on the supposition that I practise, that I will pass and receive my license—we write this $P(\textit{License}||\textit{Practise})$. And I weight $U(\textit{Practise} \,\&\, \textit{No License})$ by my credence, again on the supposition that I practise, that I will not get my license—we write this $P(\textit{No License}||\textit{Practise})$. Thus, the value of the option *Practise* is:

$$\begin{aligned} V(\textit{Practise}) = {} & P(\textit{License}||\textit{Practise}) \times U(\textit{Practise} \,\&\, \textit{License}) \\ & + P(\textit{No License}||\textit{Practise}) \times U(\textit{Practise} \,\&\, \textit{No License}) \end{aligned}$$

I then do the same for *Don't Practise*:

[11] As we'll see below, it's not quite right to call it *the* real number that measures this, but let's indulge in this fiction for the moment.

$$\begin{aligned} V(\textit{Don't Practise}) = {} & P(\textit{License}||\textit{Don't Practise}) \times U(\textit{Don't Practise} \,\&\, \textit{License}) \\ & + P(\textit{No License}||\textit{Don't Practise}) \times U(\textit{Don't Practise} \,\&\, \textit{No License}) \end{aligned}$$

Expected utility theory then says that I am rationally required to pick whichever of the options I evaluate as the better means to my ends; or, if I evaluate them as equally good, I am rationally permitted to pick either. In short, I am required to maximize my subjective expected utility, which is what $V$ measures; I am required to pick from among the options that have maximal subjective expected utility.

## 2.2 Expected utility theory: the general case

Having seen expected utility theory in action in a particular case, let's see how it works in general. Like every formal theory of rational decision making, expected utility theory takes a real-life, concrete, flesh-and-blood decision that you face, and provides a formal model of that decision, which we might call the corresponding *decision problem*. Decision problems contain representations of what you must choose between: alternative actions in decision theory, strategies in game theory, and so on. And it is the job of the decision theory to separate out those possible choices into those that rationality permits you to choose, and those it doesn't.

Let's consider the formal model that expected utility theory offers. In that theory, a decision problem consists of the following components:

- $\mathcal{A}$ is the set of possible acts (or options).

  *In our driving example,* $\mathcal{A} = \{\textit{Practise}, \textit{Don't Practise}\}$.

- $\mathcal{S}$ is the set of possible states of the world (or possible worlds).

  These form a partition of the possible ways the world might be. That is, the states are exclusive, so that, necessarily, at most one is true; and they are exhaustive, so that, necessarily, at least one is true.

  *In our example* $\mathcal{S} = \{\textit{License}, \textit{No License}\}$.

- $P$ is our agent's credence function.

  This is the component of our formal model that represents the agent's *doxastic state*—that is, it represents her beliefs, her levels of confidence, the attitudes that record how she takes the world to be. $P$ is a function that takes an act $a$ from $\mathcal{A}$ and a state $s$ from $\mathcal{S}$ and returns the agent's current credence that $s$ is the actual state of the world under the supposition that she performs act $a$, which we denote $P(s||a)$ or $P^a(s)$.[12] An agent's credence in a state, such as $s$, is the strength of her belief in it; it is her degree of belief in it; it measures how confident she is in it. It is measured on a scale from 0% to 100%, or 0 to 1. Thus, a credence of 100% (or 1) is certainty—it is the highest possible confidence. 0% (or 0), on the other hand, is the lowest. We might have 0% credence in something we're certain is false, for instance.

We assume that, for each $a$ in $\mathcal{A}$, $P(-||a)$ (or $P^a(-)$) is a probability function. That is, under the supposition that $a$ is chosen, an agent's credences in each state of the world, taken together, sum to 1; and her credences in any proposition, on the supposition of $a$, is just the sum of her credences, on the supposition of $a$, in each state of the world in which that proposition is true.[13]

- $U$ is our agent's utility function. This is the component of our formal model that represents our agent's *conative state*—that is, her desires, her values, what she wants, her likes and dislikes, attitudes that record how she would like the world to be. $U$ is a function that takes an act $a$ from $\mathcal{A}$ and a state $s$ from $\mathcal{S}$ and returns the agent's utility, $U(a \,\&\, s)$ (or $U^a(s)$), for being in that state having performed that act. As mentioned above, $U(a \,\&\, s)$ measures how much she values the outcome $a \,\&\, s$; how much she desires it or wants it to be the case.[14]

[12] Nothing will turn on whether the supposition in question is indicative or subjunctive, and thus whether the decision theory is evidential or causal, so I leave this unspecified. For more on this question, see (Gibbard & Harper, 1978; Lewis, 59; Joyce, 1999).

[13] For arguments that credences with this property are required by rationality, see (Hájek, 2008; Joyce, 1998; Pettigrew, 2016a), but also chapter 9.

[14] Thus, initially, my utility function is defined only on conjunctions $a \,\&\, s$, which specify which act from $\mathcal{A}$ I perform and which state of the world from $\mathcal{S}$ is actual. Given a proposition $X$, which might be represented by the set of states of the world at which it is true, I can also define my utility for the conjunction $a \,\&\, X$, which tells me that act $a$ is performed and proposition $X$ is true. My utility for $a \,\&\, X$ is my conditional subjective expectation of my utility for $X$ under the supposition of $a$ and conditional on $X$: that is,

$$U(a \,\&\, X) = U^a(X) = \sum_{s \in \mathcal{S}} P^a(s|X)U^a(s)$$

This ensures that my decision theory is *partition invariant*. That is, the recommendation that my decision theory makes is not sensitive to the level of grain at which I define my decision problem. For more on this feature, see (Joyce, 1999, 178). As we will see in chapter 7, however, even this assumption won't ensure that my favoured solution to the problem of choosing for changing selves is also partition invariant.

Thus, I might assign a utility of 2 to the outcome in which I don't practise and don't pass—that is, $U(\textit{Don't Practise} \ \& \ \textit{No License}) = 2$—while I assign a utility of 8 to receiving my license having practised—that is, $U(\textit{Practise} \ \& \ \textit{License}) = 8$—and so on.

In fact, there is some subtlety here, which will become important in chapter 8. Utility functions assign real numbers to outcomes. But consider the following two utility functions:

| | $U$ | $U'$ |
|---:|:---:|:---:|
| *Practise* & *License* | 8 | 17 |
| *Practise* & *No License* | 10 | 21 |
| *Don't Practise* & *License* | 1 | 3 |
| *Don't Practise* & *No License* | 2 | 5 |

The utility that $U'$ assigns to an outcome is obtained by doubling the utility that $U$ assigns to it and adding 1; and the utility that $U$ assigns to an outcome is obtained by subtracting 1 from the utility that $U'$ assigns to it and halving the result. In such a case, where one utility function is obtained from another by multiplying by a positive constant and adding a constant, we say that one is an *positive linear transformation* of the other.[15] In the formal model offered by expected utility theory, we take one utility function to be just as good as a representation of an agent's conative state—her desires, her values, her likes and dislikes—as any other that is obtained from it by an positive linear transformation. In this sense, utility is like temperature: the Celsius and Fahrenheit scales are equally good representations of temperature, and they are positive linear transformations of one another.[16] This means that we take there to be no sense in saying that an agent assigns four times as much utility to one outcome as to another, just as it makes no sense to say that it's four times hotter in Bristol than in Irkutsk today, since such relationships are not preserved under positive linear transformation. *Practise* & *License* has four times more utility than *Don't Practise* & *No License* relative to the representation $U$, but not relative to the equally valid representation $U'$; Bristol

[15] One utility function, $U'$, is a positive linear transformation of another, $U$, if there are real numbers $\alpha$ and $\beta$, with $\alpha > 0$ such that $U'(-) = \alpha U(-) + \beta$.

[16] Given a temperature measured on the Celsius scale (°C), you obtain the same temperature measured on the Fahrenheit scale (°F) by multiplying by $\frac{9}{5}$ and adding 32. To move from Fahrenheit to Celsius, you subtract 32 and multiply by $\frac{5}{9}$.

may currently be twice as hot as Irkutsk according to the Celsius representation (20°C vs 10°C, say), but it won't be relative to the Fahrenheit representation (68°C vs 50°C). Only relationships that are preserved by positive linear transformation make sense. So it would make sense to say that our agent assigns more utility to one outcome than to another, or that the difference between their utilities for two outcomes is twice as great as the difference between their utilities for two other outcomes, since such relationships are preserved by positive linear transformations.

Now, even to say that the utilities are defined up to positive linear transformation is quite a substantial assumption. It is equivalent to the axioms for decision making under risk that were formulated by John van Neumann and Oscar Morgenstern (von Neumann & Morgenstern, 1947). I will simply assume for the moment that such an assumption is justified. In chapter 8, I will consider it in more detail; and in chapter 16, we will ask how my favoured solution to the problem of choosing for changing selves fares if we drop this assumption, such as we do in Richard Jeffrey's decision theory, or in Lara Buchak's (Jeffrey, 1983; Buchak, 2013).

- $V$ is our agent's evaluation function.

  This component represents the agent's doxastic and conative states together. It takes an act $a$ in $\mathcal{A}$ and it measures the extent to which the agent judges $a$ to be a good means to her ends.

- $\preceq$ is our agent's preference ordering.

  This component of our formal model also represents the agent's doxastic and conative states together, but whereas $V$ provides a cardinal representation, $\preceq$ provides only an ordinal one. It orders the acts in $\mathcal{A}$ according to their choiceworthiness.

Some of the components of the formal model are related. In particular, we require:

(EU1) $V(a)$ is the agent's subjective expectation of the utility of $a$.

That is,

$$V(a) = \sum_{s \in \mathcal{S}} P(s||a)U(a \,\&\, s) = \sum_{s \in \mathcal{S}} P^a(s)U^a(s)$$

(EU2) $a \preceq b$ just in case the agent evaluates $b$ at least as highly as she evaluates $a$.

That is,

$$a \preceq b \text{ iff } V(a) \leq V(b)$$

With the formal model laid out, we are ready to state in full generality the way in which expected utility theory categorises acts in $\mathcal{A}$ into those that are permissible and those that are not:

> **Maximise Subjective Expected Utility (MSEU)** It is irrational to choose an act from $\mathcal{A}$ that has less than maximal subjective expected utility.
>
> That is, $a$ in $\mathcal{A}$ is irrational if there $b$ in $\mathcal{A}$ such that $V(a) < V(b)$.
>
> That is, $a$ in $\mathcal{A}$ is irrational if there $b$ in $\mathcal{A}$ such that $a \prec b$.[17]

## 2.3 Interpreting expected utility theory

There are two interpretations of decision theory. I will call them the *realist interpretation* and the *constructivist interpretation*.[18] Both agree on the ingredients of a decision problem: a set of acts $\mathcal{A}$, a set of states $\mathcal{S}$, a preference ordering $\preceq$, a credence function $P$, a utility function $U$, and an evaluation function $V$. But they disagree on which ingredients are more fundamental than which others. Thus, the constructivist claims that the preference ordering is fundamental, and the credence and utility functions are determined by that preference ordering via a representation theorem, which establishes that, if $\mathcal{S}$, $\mathcal{A}$, and $\preceq$ satisfy certain conditions, there are credence and utility functions such that the preference ordering is as it would be if the agent were to have these credences and utilities and were to determine their evaluation function on the basis of them via (EU1), and their preference ordering on the basis of that via (EU2). The realist, on the other hand, says that the credences and utilities are fundamental and they determine the evaluation function via (EU1), and the preference ordering via (EU2). Throughout, we adopt a realist position.

Why? One reason is that this seems better to reflect how we deliberate about our decisions. We think about what the world is like—thereby setting our credences—and we think about what we value—thereby setting our utilities. On the basis of these, we set our preferences and we make our

[17] By definition: $a \prec b$ iff $a \preceq b$ and $b \not\preceq a$.

[18] I borrow the terminology from Buchak (2013). Okasha (2016) uses 'mentalistic' instead of 'realist', and 'behaviouristic' instead of 'constructivist'. Eriksson & Hájek (2007) argue for a realist account of credences.

decision. When new evidence arrives, we change our credences first, and that then often determines a change in our preferences. And when we have a new experience, perhaps, we change what we value, and that might then also determine a change in our preferences. On the constructivist view, new evidence, or a change in your values, initially affects your preference ordering, and only secondarily does it affect the credences and utilities you are represented as having. In this book, I'm interested in providing a decision theory that we might actually use to deliberate about choices we face. So I go realist.

## 2.4 Alternative decision theories

This, then, is expected utility theory. Other decision theories represent further or fewer features of your mental state, or they represent the same ones in a different way. Three examples:

- *Von Neumann-Morgenstern decision theory* represents fewer features of your mental state (von Neumann & Morgenstern, 1947). In particular, it does not represent your credal state. The acts between which you choose are lotteries over states of the world. That is, each act specifies the objective chance of each state of the world coming about as a result of you choosing that option. Thus, to each act $a$ corresponds an objective chance function $C^a$ over the possible states of the world—given a state $s$, $C^a(s)$ is the objective chance of $s$ given that you choose $a$. The objective expected utility of an action $a$ is $\sum_{s \in S} C^a(s)U(s)$. And you are then required to pick a lottery with maximal objective expected value.

- *Imprecise decision theory* represents the same features of your mental state as expected utility theory, but it represents them differently (Elga, 2010; Joyce, 2010; Moss, 2015; Rinard, 2015). It represents your credal state as a *set* of probability functions, rather than a single one; and it represents your values as a *set* of utility functions, rather than as a single one. The idea is that the true features of your credal state are those shared by all probability functions in the set, while the true features of your conative state are those shared by all utility functions in the set. Thus, you think that $s$ is more likely than $s'$ given $a$ iff $P(s||a) > P(s'||a)$, for all $P$ in the set that represents her doxastic state. And you value $a$ & $s$ more than $a'$ & $s'$ iff $U(a \,\&\, s) > U(a' \,\&\, s')$ for all utility functions $U$ in the set that represents your conative state. This allows us to represent you even if you don't have a determinate opinion as to whether $s$ is more or less likely than $s'$ given $a$, for instance. The set of probability functions that would then represent your doxastic state includes $P_1$ such that $P_1(s||a) < P_1(s'||a)$ as well as $P_2$ such that $P_2(s||a) > P_2(s'||a)$ and $P_3$ such that $P_3(s||a) = P_3(s'||a)$. And it allows us to represent you even if you don't have a determinate opinion as to whether situation $a$ & $s$ is more or less valuable than $s'$ & $a'$. The set of utility functions that would then represent your conative state includes $U_1$ such that $U_1(a \,\&\, s) < U_1(a' \,\&\, s')$ as well as $U_2$ such that $U_2(a \,\&\, s) > U_2(a' \,\&\, s')$ and $U_3$ such that $U_3(a \,\&\, s) > U_3(a' \,\&\, s')$.

Having represented the doxastic and conative states differently, we also need a new decision rule. After all, for each credence function from the set of probability functions representing your doxastic state and for each utility function from the set of utility functions representing your conative state, there is a set of acts that maximise subjective expected utility relative to that credence function and that utility function. Usually, those will be different sets. Which acts are rationally permissible? You might think it is those that occur in *every one* of those sets; or those that occur in *any one* of those sets; or something different. Each faces a difficulty. We will consider this at greater length in chapter 16.

- *Risk-weighted expected utility theory* represents further features of an agent's state (Buchak, 2013; Quiggin, 1993). In particular, it represents her attitudes to risk. It does this using a function $r$, which transforms the agent's credences. Roughly speaking, if she is risk-averse, $r$ transforms the probabilities of the worse outcomes by magnifying them, and it transforms the probabilities of better outcomes by shrinking them. The former thus feature more prominently in the resulting risk-weighted expected utility calculation than they would in a standard expected utility calculation. And Buchak's theory demands that agents maximise risk-weighted expected utility.

We will stick with expected utility theory for most of the book. Why? Well, largely because it is simple and familiar. Once we have seen how to accommodate choosing for changing selves in this framework, we will then explore other frameworks in chapter 16.

Having said that, we offer an argument in favour of expected utility theory in chapter 9. It assumes that we have at least the ingredients of expected utility theory. Thus, it does not tell against von Neumann-Morgenstern or imprecise decision theory; but it does tell against Buchak's risk-weighted expected utility and other risk-sensitive decision theories, such as prospect theory.

## 2.5 Expected utility theory and the problem of choosing for changing selves

How does the problem of choosing for changing selves arise for expected utility theory? In that theory, the agent's utility function measures how much she values various outcomes. But, as we saw in the examples of Aneri, Blandine, and so on, the values that agents assign change over the course of their lives. Thus, it might seem that an agent can have different utility functions at different moments throughout their lives, reflecting their changing values as their life progresses. But if that's the case, which of these should the agent use to make a decision at a particular time? Her utility function that reflects her values at the time she makes the decision? Or her utility function at some later time, when the effects of the decision are felt? Or perhaps she should use none of the individual utility functions, but some other one that aggregates the value judgments encoded in the individual ones, perhaps giving different weights to different selves based on their similarity to the current self, who is making the decision, or based on their proximity in time, or some other criterion. Or perhaps, instead of aggregating the utility functions of her various past, present, and future selves, she should instead aggregate their evaluation functions; that is, she should aggregate their expected utilities, rather than their utilities. We will discuss which, if any, of these our agent should do at much greater length in chapter 6.

Of course, you might respond to me thus: 'Look, my utilities at the present time encode *everything* I value now, and that includes not just what you are calling my values now, but also how much consideration I want to give to what you are calling my values at other times. So there is really no problem here. My ordinary utilities already achieve what you are trying to do.'[19] Now, if I were a constructivist, this would surely be correct. On that account, preferences and the utilities that represent them encode not just what you value now, but also what you value at other times and your attitudes to those other values. But I am not a constructivist—I am a realist. And it is clear from the realist point of view that we do have values that change over time—the examples of Aneri, Blandine, Cheragh, etc. from the

[19] Thanks to Sarah Moss for pushing me to address this challenge.

opening chapter make that clear. And that requires us to determine a further set of values that we will use when we make decisions. These further, decision-making values will somehow relate to the values we have that change over time. How they relate is precisely our question here. Thus, I distinguish between *local* values and utilities, on the one hand, and *global* or *decision-making* values and utilities, on the other. For Aneri, for instance, her local utilities now represent her current moderate value for conformity and moderate value for self-determination, while her local utilities later if she becomes a police officer represent her future stronger value for conformity and weaker value for self-determination. Her global or decision-making utilities at a time, on the other hand, are those she should use to make her decisions at that time. The question of the book can then be restated: how should local values and utilities relate to global or decision-making values and utilities?

# Chapter 3

# Existing solutions I: the Unchanging Utility Solution

In this chapter and the next two, we consider three existing solutions to the problem of choosing for changing selves: the *unchanging utility solution* (Chapter 3), the *utility of utility solution* (Chapter 4), and the *one true utility solution* (Chapter 5). My complaint in the first two cases is that they are not complete solutions to the problem. While each is successful in certain sorts of case, neither is successful in all, and there are cases that neither treats correctly.

The Unchanging Utility Solution gives the correct account of Erik's case from the Introduction and the Utility of Utility Solution identifies how to respond to Giang's case. But neither treats Aneri's case correctly, nor Blandine's nor Cheragh's nor Deborah's. The One True Utility solution has the potential to treat all cases correctly. But it assumes an extreme and implausible version of objectivism about value.

According to this solution, contrary to appearances, our values don't in fact change over time, and so our utilities don't change either. Thus, there is no problem of choosing for changing selves, either for our informal account of decision-making, or for our formal account in the guise of expected utility theory. According to this account, we should make our decisions based on our single, unchanging set of values, which are reflected in our single unchanging utility function—we might call this unchanging utility function our *ur-utility function* and the values it represents our *ur-values*.[20]

[20] Nagel (1978) considers a solution along these lines to a related problem. Stigler & Becker (1977) also attempt to account for economic phenomena that are usually explained by posit-

The idea is best introduced using an example.

> **Ice cream** On Monday, you kindly offer to deliver a tub of ice cream to me on Friday, either lemon sorbet or dark chocolate. Which flavour should I ask you to deliver? At the start of the week, I enjoy the refreshing sharpness of lemon sorbet; indeed, on Monday, I enjoy that more than the richness of dark chocolate. On Fridays, in contrast, I enjoy the decadence of dark chocolate ice cream; indeed, on Fridays, I enjoy that more than the acidity of the lemon sorbet. On Monday, you ask me which you should deliver to me on Friday. What should I say?

Clearly, I should ask you to deliver the dark chocolate. After all, the ice cream will be delivered on Friday, and I will enjoy the dark chocolate more on Friday. At first glance, this seems like a paradigm case of choosing for changing selves. Indeed, it is analogous to the example of Erik from the introduction, who has to decide whether or not to buy a bargain price ticket now to a white knuckle theme park for his future elderly self, who would not enjoy it, even though his current self would enjoy it greatly. According to the Unchanging Utility Solution, however, in neither case do your values, nor the utilities that record them, change over time. In the ice cream case, you value gustatory pleasure throughout, and at exactly the same level. What changes between Monday and Friday is what gives you that pleasure. Thus, the change is in the world, or in the world's influence on your affective states. Nothing changes in your conative state. According to the Unchanging Utility Solution, the ice cream case is analogous to the following case. On Monday, £10 is worth more than $10, but you know that, by Friday, $10 will be worth more than £10. On Monday, you'd like to receive £10 on that day more than $10; and on Friday, you'd like to receive $10 on that day more than £10. But here we would not say that your values have changed. Rather, we would say that the world has changed so that, on Monday and on Friday, different elements of it give you more of what you value, namely purchasing power.

How do we calculate the utility of a state of the world, according to the Unchanging Utility Solution? Consider two cases: in both, I prefer lemon sorbet at the beginning of the week and dark chocolate ice cream at the end, and with exactly the same intensity in each case. In the first state of

ing changes in value—addiction, for instance, and the efficacy of advertising—while maintaining that people have what they call "stable tastes". Becker has more recently extended this to a book-length treatment (Becker, 1998).

the world, I receive lemon on Monday and dark chocolate on Friday; in the second, these are reversed. Then, according to the Unchanging Utility Solution, my utility for the two states will be different, even though they contain the same bundles of commodities—a tub of lemon sorbet and a tub of dark chocolate ice cream—and my values don't change. The point is that my utility is determined by, for each time, how much the world gives me what, at that time, produces what I timelessly value. So I assign higher utility to the state *lemon-on-Monday-and-dark-chocolate-on-Friday* than to the alternative *dark-chocolate-on-Monday-and-lemon-on-Friday* because lemon sorbet on Monday gives me, on Monday, more of what I timelessly value, namely, gustatory pleasure, than dark chocolate ice cream on Monday, while dark chocolate ice cream on Friday gives me, on Friday, more of what I timelessly value than lemon sorbet on Friday.

Now, this is a very plausible analysis in the case of the ice cream dilemma, and the white knuckle theme park ticket decision, and the choice between £10 and $10. Indeed, its insight reminds me of conversations with my brother, who loathed broccoli. I told him that that he might learn to like it, and he responded: 'But why would I want to like broccoli? Then I would end up eating it and I hate it!' The joke works because his argument is obviously absurd, and the Unchanging Utility Solution captures what is absurd about it. Even if my brother hates broccoli now, he shouldn't spurn the opportunity to learn to like it. If he'll like it in two months' time, he should assign a high utility to being served a plate of it in two months' time, since then it will give him gustatory pleasure, and it is such pleasure that he timelessly values; he shouldn't assign a low utility to being served it in two months' time on the basis that it would give him no such pleasure now. (Though, of course, explaining it in this way rather spoils the joke.)

In the end, I agree with the Unchanging Utility Solution in the cases we have been considering, and I will accommodate its insight into the solution I favour. However, the Unchanging Utility Solution takes its analysis much further. Indeed, according to that solution, the same analysis holds in *any* case in which it seems that our values change. So, for instance, it holds for all of the examples with which we began this book. Take, for instance, Aneri, who is deciding whether to become a police officer or a conservation officer. Let's suppose she opts for the former, and suppose that, as she settles in to the training and eventually settles in to the job, her values seem to change: to begin with, she seems to value conformity more than self-direction, but not much more; by the end, she seems to value conformity very highly and self-direction only a very small amount. According to the Unchanging Utility Solution, this change in her values is only apparent. In fact, Aneri's values remain the same throughout. What does change is what procures for Aneri what she values. Thus, just as I continued to value gustatory pleasure throughout the week leading up to my Friday ice cream delivery, while what changed was what gave me that pleasure—lemon sorbet on Monday, dark chocolate ice cream on Friday—so, for Aneri, there is something that she values throughout the period from deciding to join the police force to becoming an established officer, while what changes is what gives her whatever that thing is that she values.

There are two problems with the Unchanging Utility Solution, the second following on from the natural solution to the first. The first is that it's hard to specify exactly what that thing is that Aneri values throughout her career with the police force such that all that changes is what features of the world procure for her that thing. In the example of the ice cream, it was easy to identify what it is that I value timelessly throughout—it is gustatory pleasure. And similarly for the case of the currencies—it is something like purchasing power. And again for the case of the theme park ticket—it is physical or bodily pleasure. But in Aneri's case, it is less obvious.

I think the most natural thing to say is this: throughout, Aneri unchangingly values getting what she prefers, or what she endorses, or being in situations of which she approves. At the beginning, she endorses or prefers or approves of activities that require more conformity than self-determination, but only a little more; at the end, once she is an established police officer, she endorses or prefers or approves of activities that require much more conformity than self-determination.

The first apparent problem with this solution is that it doesn't seem to move us forward. After all, according to the solution, my values don't change—I continue to value getting what I endorse or prefer—but what I endorse or prefer does change. But the Unchanging Utility Solution seems then to face a dilemma. On the first horn, endorsing and preferring and approving are just species of valuing. Thus, while we might say that my ur-values don't change, since my single ur-value is to get what I endorse or prefer and I never lose that value, my lower-level values do change, since those lower-level values are just what I endorse or prefer. If that's the case, then our values, taken as a whole, do in fact change and the problem of choosing for changing selves returns. On the second horn, endorsing and preferring are not just species of valuing. But in that case, we realise that, while we presented the problem of choosing for changing selves as the problem of how to make decisions when your *values* might change over time, the problem is really one of how to make decisions when what your *prefer* or *endorse* might change over time. Either way, the problem of choosing for changing selves remains unsolved.

In fact, this problem is only apparent. Recall the ice cream case: there, we determined the utility of an entire state—either lemon-on-Monday-and-dark-chocolate-on-Friday or dark-chocolate-on-Monday-and-lemon-on-Friday—by looking at each moment the state contains, asking to what extent what is happening to the agent at that moment gives her gustatory pleasure at that moment, which is what she unchangingly values, taking that to be that moment's contribution to the utility of the entire state, and then determining the utility of the state from the utilities of the moments in some manner—we might just add them up, for instance. We can apply that strategy quite generally, and in particular in Aneri's case. We look at each moment in her trajectory from deciding to join the police force to becoming an established officer; we look at the extent to which her experience at each of those moments gets her something she endorses or prefers *at that moment*, and then we take that to be that moment's contribution to the utility of the entire state.

So the Unchanging Utility Solution does offer us a solution to our problem. And indeed it is admirably democratic amongst Aneri's different selves, past, present, and future. It simply makes a choice between options by looking at how good or bad each moment of their outcome would be for the self that exists at that moment.

However, I anticipate two problems with this. First, it seems to place too much emphasis on satisfying preferences *in the moment*. Second, it seems to make it too easy to choose to change your preferences.

To illustrate the first problem, consider the following example:

> **Chinara**, a friend of Cheragh, is also deciding whether to embark on writing the novel she's been planning for years. But she's spoken to a few writers and she thinks the following is very likely: throughout the process of writing it, she'll have a very strong preference in favour of having written a book—she'll think of that as a real achievement that is worth celebrating. She won't much prefer the process of actually writing it—she has little time for half-written novels and only endorses the situation of having actually completed one. However, she also knows that, after a writer has been through the process of completing a novel, they no longer have that strong preference for having completed a novel.

In this situation, it seems to me, it would be quite rational for Chinara to embark on writing the novel—whether or not rationality *requires* her to start writing it, it must certainly *permit* it; if Chinara does start writing, we would surely not call her irrational. After all, throughout the process, she'll be working towards something that she values, every day making it more likely that she'll achieve the thing she values most. And surely working towards something you currently value is not irrational, even if you will end up valuing it much less when it is completed, or more poignantly when you know you will never live to see it completed. But the Unchanging Utility Solution gives the opposite result. Suppose Chinara does decide to write her novel. Then, at each moment from the first time she puts pen to paper until the day she sees copies of the book on the shelves of her local bookstore, she will not, at that moment, get anything that, at that moment, she prefers or endorses.[21]

That's the first problem. The second is also best illustrated by example. Think again about Aneri. Recall that, as we first spelled out that example, there was a trade off. If she chooses to become a conservationist, her values will stay close to those she currently has—indeed, they will remain the same—but her job will not fully satisfy those values; if, on the other hand, she becomes a police officer, her values will change considerably, but her job will satisfy those values better. In such a situation, the Unchanging Utility Solution requires that Aneri become a police officer. Doing so, she will obtain more of what she values at the time she values it. And indeed the Unchanging Utility Solution says the same for any analogous case. However much your values will change as a result of a choice you make, if by making that choice you'll get more of what you'll come to value than you'll get of what you currently value by choosing the alternative, the Unchanging Utility Solution says you should do it. But that seems too strong. It seems to pay too little attention to our current values. What would it mean to value something if you were happy to do anything that will radically change your values in order to get more of what you would then come to value? We will return to this question again and again in what follows. Choices that can lead to changes in value are difficult precisely because there are two

[21] Of course, you might interpret the Unchanging Utility Solution as saying that the value that Chinara assigns to a state of the world is not the sum, over all moments, of the value she assigns at that moment to that moment, but rather then sum, over all moments, of the value she assigns at that moment to the whole state. If that is the case, the Unchanging Utility Solution might well render it rational for Chinara to write her novel. But if this is the Unchanging Utility Solution, then I have no quarrel with it—indeed, it is a particular version of the solution that I favour and for which I will argue later in the book. What it says is that, while my local utilities do change over time, my decision-making utilities never change because they are simply the unweighted average (or sum) of my local utilities. Thanks to James Willoughby for pushing me to clarify this.

competing considerations. One consideration is that you wish to obtain what you currently value; the other is that you wish to make it rationally permissible to escape your current values, lest a certain sort of parochialism or conservatism sets in. The second problem with the Unchanging Utility Solution to the problem of choosing for changing selves is that it ignores the first consideration completely.

# Chapter 4

# Existing solutions II: the Utility of Utility Solution

The second potential solution to our problem that I would like to consider is the Utility of Utility Solution. Edna Ullmann-Margalit (2006), in one of the few published philosophical discussions of choosing for changing selves, proposes something close to this. Again, we illustrate using an example.

> **Adoption** I am deciding whether or not to (apply to) adopt a child. Currently, I assign a lower value to adopting and becoming a parent than I do to remaining child-free. However, these are not the values I would most like to have. I would prefer to assign higher value to adopting than to remaining childless. That is, while I currently prefer not being a parent, I'd prefer to be someone who prefers being a parent. What should I do?

This is a familiar sort of situation: I currently enjoy watching reality television shows and reading spy fiction; but I'd prefer to enjoy listening to Handel and reading Virginia Woolf. Or recall Giang from the Introduction: he values seriousness completely and frivolity not at all, but would like to value seriousness a little less and frivolity a little more. Quite often, we have preferences concerning which preferences we have; we assign different values to having various alternative sets of values. On the Utility of Utility Solution, we represent these as our *higher-order utilities*. Our first-order utilities represent the values that we have concerning different ways the world might be that do not involve which values we have; and our second-order utilities represent the values that we have concerning different first-order utilities. Thus, I might assign a particular first-order utility to the state of the world in which I take an umbrella when I go for a walk and it rains, while I assign a second-order utility to the state of the world in which I assign that particular first-order utility to that state of the world, and so on. Of course, most people haven't thought about higher-order utilities above the third level. Thus, I might wish I were a person who valued Handel above the *X Factor*, but hate myself for this self-hating snobbery and wish I were the sort of person who is at peace with the low-brow tastes I in fact have. That involves a third-order utility. Anything higher is a little far-fetched.

In any case, according to the Utility of Utility Solution, when choosing for changing selves, we should defer to our higher-order utilities. Thus, in our example, where I am choosing whether or not to have a child, where I know that I will prefer having a child if I have one and continue to prefer remaining child-free if I do not, and where I assign a high second-order utility to the state of the world where I prefer having a child to the state in which I prefer remaining child-free, the Utility of Utility Solution says that I should choose the actions that will bring my first-order utilities into line with my second-order utilities. I should choose to maximise second-order utility; or, perhaps better, I should choose to maximise a weighted sum of the first-order utilities I could have, each weighted by my second-order utility in my having that first-order utility. Either way, in the case described, I should choose to adopt a child.

There are at least two reasons why we should not take the Utility of Utility Solution as our complete solution to the problem of choosing for changing selves. First, higher-order utilities often change in lock-step with first-order utilities. So, for instance, while I have described myself as having current second-order utilities that do not endorse my current first-order utilities concerning parenthood, we might easily imagine someone else who currently has pro-child-free first-order utilities together with second-order utilities that endorse them, and who will then, having become a parent, have pro-parent first-order utilities together with second-order utilities that endorse those. But, for this person, the problem of choosing for changing selves arises again: to which first- and second-order utilities should they appeal when they make their decision? Their current ones? Their future ones? Or some amalgamation of them all, perhaps with different weightings for different times? If we appeal to the Utility of Utility Solution again, looking this time to third-order utilities to adjudicate, we begin on what might turn out to be an infinite regress.

A similar problem arises if, instead of having second-order utilities that always endorse the changing first-order ones, we have second-order utilities that always reject them. Thus, I might currently have pro-child-free first-order utilities and second-order utilities that prefer pro-parent first-order utilities, but if I were a parent, I'd have pro-parent first-order utilities, but second-order utilities that prefer pro-child-free first-order ones. Again, in this situation, our central question arises for the Utility of Utility Solution: to which second-order utilities should I appeal? My second-order utilities now, at the time of the decision, or my second-order utilities in the future, after I've become a parent or remained child-free? Thus, the Utility of Utility Solution is, at best, incomplete—it says nothing about cases in which second-order utility might change in the future.

The second problem with the higher-order utilities solution suggests that it might also give the wrong answer in the cases that it does cover. The problem is that it assumes that higher-order utilities have a certain normative priority over lower-order utilities. When my current second-order utilities endorse first-order utilities other than those I currently have—when they assign higher second-order utility to those alternative first-order utilities than they assign to my current first-order utilities—then my first- and higher-order utilities fail to *cohere*. But, as with any other sort of incoherence, when it is revealed, there are always a number of different rational responses. If my beliefs are incoherent because I believe the temperature is below 15°C, but disbelieve that it's below 30°C, then I might respond by throwing out my belief that it's below 15°C, or abandoning my disbelief that it's below 30°C, or by changing my attitudes to both propositions in a way that results in coherence between them. Similarly, when I realise that my first- and higher-order utilities do not cohere—because the higher-order utilities endorse first-order utilities other than those I have—there are a number of ways in which I might restore rationality. I might stick with the higher-order utilities and try to effect change in my first-order utilities, just as the higher-order utilities solution suggests. But, equally, I might stick with my first-order utilities and change my higher-order utilities. Or I might change them both in some way that restores coherence between them. The second problem with the higher-order utilities solution is that it gives priority to resolving the incoherence by sticking with the second-order utilities and changing the first-order utilities, or at least ignoring the first-order utilities in decisions, and instead appealing to the first-order utilities that the second-order utilities endorse. But there is no principled reason to award higher-order utilities such normative priority.

It may be that reflecting on your values and forming higher-order attitudes towards them will help to remove any irrationality that lurks within them—perhaps your preferences are cyclical, and by taking a step back to consider them carefully, you will realise that and rectify the situation. But it is hard to even make sense of the claim that reflecting on your values improves them when they are already rational—by which lights would we judge whether they had been improved? by the lights of the old values or the new ones? So it is not obvious how we could ever identify a reason why higher-order values should have normative priority. And absent such a reason, we should not take the Utility of Utility Solution to dictate how an agent should choose in all situations.

# Chapter 5

# Existing solutions III: the One True Utility Solution

We come now to the third putative solution to the problem of choosing for changing selves—the One True Utility Solution.[22] Consider again the examples from the beginning of the book—Aneri, Blandine, Cheragh, etc.—where it seems initially as if the individuals' values change over time. According to the One True Utility Solution, while there is a sense in which it is true that the *subjective* values of each of these individuals change over time, what really happens is that their opinions about what are the *unique objectively correct* values change over time. That is, while there is a change in their conative state, it comes about as a result of a change in their doxastic state. And, indeed, on this solution, the components of an individual's conative state—her subjective values, desires, likes and dislikes—are all simply shadows cast by certain of her doxastic states—in particular, her credences concerning what is objectively valuable. In particular, her subjective utilities at a particular time are just the expectations of the objective utilities by the lights of her credences at that time. That is, for any act $a$ and state $s$,[23]

$$U^a(s) = \sum_{OU} P^a(OU \text{ gives the objective utilities}) OU^a(s)$$

[22] Ralph Wedgwood defends something like this solution (Wedgwood, 2007, 2017).

[23] Recall: $P^a(X) = P(X||a)$ is the individual's credence in $X$ on the supposition of $a$; $U^a(s) = U(a \,\&\, s)$ is her subjective utility for the situation in which $a$ is performed and $s$ is the state of the world; and $OU^a(s) = OU(a \,\&\, s)$ is the objective utility for that same situation.

where the sum ranges over every utility function that, as far as the individual knows, might be the objective utility function.[24] For obvious reasons, we will refer to the proponent of the One True Utility Solution henceforth as the objectivist.

Recall the example of Blandine from the opening chapter. For many years, she valued being a musician in a band above all else; but now she values that much less and values being a particle physicist much more. As we described her above, she has changed her subjective values. And the objectivist agrees. But according to the objectivist, a person's subjective values are determined entirely by their opinions about the objective values; they are her expectations of those objective values. Thus, according to the objectivist, Blandine's subjective values have shifted because she has moved from being very confident that the one true objective utility function assigns a very high utility to being a musician to being rather sure that this isn't the case, and instead being quite sure that the one true objective utility function in fact assigns a low utility to that and a high utility to being a particle physicist. And similarly for Cheragh, who currently has certain literary standards, but who knows that she would come to have different literary standards were she to write her novel. According to the objectivist, what underlies such a change is primarily a shift in her credences concerning the one true objective utility function.

Notice: just as the subjectivist says that a change in your values can occur as a natural part of growing older, or because of some external influence, or because of a choice you made yourself, so the objectivist can say that the underlying changes in credences concerning the one true objective utility function can occur for the same reasons. We might imagine the newly-published author Cheragh saying: 'Completing a novel taught me that it isn't very worthwhile to write one.' And indeed many new parents express the change in their values in this way: 'Holding my baby for the first time taught me the importance of family.'

Now, the problem of choosing for changing selves arises because (i) our subjective utilities play a central role in determining which action we should choose when we face with a decision, and (ii) our subjective utilities often change over the course of our lives. For the objectivist, then, the problem of choosing for changing selves arises because our credences concerning the objective utilities—and therefore our expectations of those objective utilities, and therefore our subjective utilities—change during the course of our lives.

[24] Of course, we must ensure that each of the possible objective utility functions measure value on the same scale. For more on this, see Chapter 8.

For the subjectivist, the problem is this: for them, what it is rational for you to do depends on what you believe about the world and what you value in the world. But your values change during your life. So we must ask: to which values should we appeal when we make a decision? Those you have at the time of the decision? Or your past values? Your future value? An amalgation of them all? For the objectivist, the problem is this: for them, what it is rational for you to do depends on what you believe about the world and what you believe about objective value. But given that your subjective values change over the course of your life, your beliefs about objective value must change, for the former are just the shadows cast by the latter. So we ask: to which beliefs about objective values should we appeal when we make a decision? Those you have at the time of the decision? Or your past values? Future ones? An amalgation of them all?

## 5.1 The Reflection Principle

Now, once we have formulated the problem in this way, it seems that we have a ready-made solution to it. After all, it is not just our credences about the objective values that change over the course of our lives—it is also our credences about the world. These worldly credences also play a central role in determining which action we should choose in a decision problem—when deciding whether or not to take an umbrella, we need to know how much I value staying dry, but we also need to know how likely I think it is to rain. But notice this: we never hear epistemologists puzzling over which credences about the world we should use when we make a decision. Why not? In the remainder of this chapter, we attempt to answer this question. It concerns us here primarily because it holds the key to the objectivist's solution to the problem of choosing for changing selves. But it clearly has broader interest. After all, even the subjectivist must identity the credences to which you should appeal when you calculate expected utility—you current credences? past or future ones? an amalgamation of them all? The only difference is that the subjectivist uses only credences about the states of the world, whereas the objectivist uses credences about the states of the world *and* credences about the objective utilities.

In the end, we will conclude that it is your current credences—those that record your current doxastic state—that should determine your expected subjective utilities for the subjectivist and your expected objective utilities for the objectivist. The reason is that we demand of our current credences that they already take our past and future credences into account in a satisfactory way. Our job in the coming pages is to identify how they should do this. We will consider three possibilities: two are excessively restrictive and we will abandon them; the last is acceptable and provides our answer.

Here's a first attempt to say how our current credences should incorporate our past and possible future credences. First, we assume that, since my current credences are better informed than my past ones, my current credences already incorporate my past ones by assimilating the information they stored and adding to it what I have learned since. Second, we assume that, since my future credences are better informed than my current ones, to the extent I am aware of those future credences, I should already have incorporated the information they will contain into my current credences. This is summed up formally in van Fraassen's Reflection Principle (van Fraassen, 1984, 1995, 1999):[25]

> **Reflection Principle (RP)** Suppose $P$ is my current credence function; and suppose $P_1, \ldots, P_n$ are all my possible future credence functions at some particular future time $t$. Then, for any act $a$ in $\mathcal{A}$,
>
> $$P^a(-) = \sum_{i=1}^{n} P^a(P_i \text{ gives credences at } t) P_i^a(-)$$

The Reflection Principle says that my current credences should be my current expectations of my future credences. For instance, if I'm certain that, come tomorrow, I will be 70% confident that it will rain on Saturday, I should be 70% confident now that it will rain on Saturday. Or if I am certain that, come tomorrow, I will either be 60% or 70% confident that it will rain on Saturday, and I think each equally likely, then I should be 65% confident now that it will rain on Saturday. And so on.

[25] Note: this may not be the formulation of the Reflection Principle with which you are most familiar. When van Fraassen first defended the principle, he formulated it as follows (van Fraassen, 1984):

> **Reflection Principle$^*$ (RP$^*$)** Suppose $P$ is my current credence function; and suppose $P_1, \ldots, P_n$ are all my possible future credence functions at some particular future time $t$. Then, for any act $a$ in $\mathcal{A}$, and each $1 \leq i \leq n$,
>
> $$P^a(-|P_i \text{ gives my credences at } t) = P_i^a(-)$$

RP$^*$ stands to RP as David Lewis' favoured formulation of the Principal Principle stands to Jenann Ismael's alternative (Ismael, 2008; Pettigrew, 2014; Ismael, 2013). However, RP$^*$ implies RP and, if we assume that, for each $1 \leq i \leq n$, $P_i(P_i \text{ gives my credences at } t) = 1$, then RP implies RP$^*$ as well and they are therefore are equivalent.

If the Reflection Principle is true, then, my current credences already incorporate my past and possible future credences in exactly the right way. This, in itself, already furnishes us with a solution to the objectivist's version of the problem of choosing for changing selves. But there's more! It turns out that, if we assume the Reflection Principle, together with a very plausible principle relating credences about objective utilities to credences about the world, then not only do we have a reason for appealing to our current credences when we make decisions, but we can in fact prove that it doesn't matter whether we appeal to our current credences or our anticipated future credences when we make decisions. Here is the second principle:

> **Independence Principle (IP)** For any act $a$ in $\mathcal{A}$, state $s$ in $\mathcal{S}$, and possible objective utility function $OU$,
>
> $$P^a(s \mid OU \text{ gives the objective utilities}) = P^a(s)$$

The Independence Principle says that, on the supposition of performing any act $a$, the state of the world and the identity of the objective utility function should be independent of one another. That is, I take the one to have no bearing on the other.

Given these two principles, the Reflection Principle and the Independence Principle, we have:[26]

**Theorem 5.1.1** *Suppose $P$ is my current credence function; and suppose $P_1, \ldots, P_n$ are all my possible future credence functions at some particular future time $t$. And suppose that I satisfy the Reflection Principle and the Independence Principle. Then, for any act $a$ in $\mathcal{A}$,*

$$V(a) = \sum_{i=1}^{n} P^a(P_i \textit{ gives my credences at } t) V_i(a)$$

*where*

- $V_i(a) = \sum_{s \in \mathcal{S}} P_i^a(s) U_i^a(s)$, *and*
- $U_i^a(s) = \sum_{OU} P_i^a(OU \textit{ gives the objective utilities}) OU^a(s)$.

That is, if you evaluate options in line with those current credences (and the objectivist's account of expected utility), then you will get the same result as if you take your current expectation of the evaluations you will come to give in the future (using the objectivist's account of expected utility).

Before we move on to scrutinise this solution, let's see it in action. Those who wish to can skip this example without loss:

[26] For a proof, see the Appendix to this chapter.

**Army** Effie is 20 years old, and she's deciding whether or not to become a soldier. At the moment, she's pretty indifferent about it—she could take it or leave it; she doesn't have a strong preference in favour of a career in the Army, nor a strong preference against it. But she knows that, in ten year's time, when she's 30 years old, she'll have a very strong preference one way or the other—she'll either value that career very greatly or extremely little. She's just seen that this is how people's values develop through their twenties. But she doesn't know which of these two directions her values will take—indeed, she thinks each is equally likely. She does know, however, that whether or not she becomes a soldier will make no difference at all to how her values will evolve. On the objectivist view we are considering here, Effie's situation might be modelled as follows:

- The set $\mathcal{A}$ of acts contains *Enlist* and *Don't Enlist*.
- The set $\mathcal{S}$ of states of the world contains *Soldier* and *Civilian*.
- Effie knows that, if she enlists she'll become a solider and if she doesn't she'll remain a civilian. Thus, if $P$ is her current credence function:

  $$P(\mathit{Soldier}||\mathit{Enlist}) = 1 = P(\mathit{Civilian}||\mathit{Don't\ Enlist})$$

- One of the possible objective utility functions is $OU$, where:

  $$OU(\mathit{Soldier}\ \&\ \mathit{Enlist}) = 9 \quad \text{and} \quad OU(\mathit{Civilian}\ \&\ \mathit{Don't\ Enlist}) = 5$$

- The other possible objective utility function is $OU'$, where:

  $$OU'(\mathit{Soldier}\ \&\ \mathit{Enlist}) = 1 \quad \text{and} \quad OU'(\mathit{Civilian}\ \&\ \mathit{Don't\ Enlist}) = 5.$$

- Currently, Effie thinks that it is just as likely that $OU$ is the objective utility function as that $OU'$ is. So:

  $$P(OU \text{ gives objective utilities}) = 0.5 = P(OU' \text{ gives objective utilities})$$

Now, according to the objectivist account, an individual's subjective utility for a situation, such as *Soldier & Enlist*, is her expectation of its objective utility. Thus, for instance,

$$U(\mathit{Soldier}\ \&\ \mathit{Enlist}) =$$

$$\begin{aligned}&P(OU \text{ gives objective utilities})OU(\mathit{Soldier} \,\&\, \mathit{Enlist}) + \\ &\quad P(OU' \text{ gives objective utilities})OU'(\mathit{Soldier} \,\&\, \mathit{Enlist}) = \\ &\qquad 0.5 \times 9 + 0.5 \times 1 = 5\end{aligned}$$

And similarly for *Civilian* & *Don't Enlist*:

$$\begin{aligned}&U(\mathit{Civilian} \,\&\, \mathit{Don't\ Enlist}) = \\ &\quad P(OU \text{ gives objective utilities})OU(\mathit{Civilian} \,\&\, \mathit{Don't\ Enlist}) + \\ &\qquad P(OU' \text{ gives objective utilities})OU'(\mathit{Civilian} \,\&\, \mathit{Don't\ Enlist}) = \\ &\qquad\quad 0.5 \times 5 + 0.5 \times 5 = 5\end{aligned}$$

So, $V(\mathit{Enlist}) = 5 = V(\mathit{Don't\ Enlist})$.

This gives the credences and utilities and the resulting evaluation endorsed by Effie's current self. Let's now consider the credences and utilities of her two possible 30 year old selves: the first has come to value an army career, the second to disvalue it. For the objectivist, this means that her credences concerning which of $OU$ and $OU'$ is the one true objective utility function have changed. Her first future self admires soldiering, and so must have increased her credence that $OU$ is the objective utility function—let's say that the credence function of that future self is $P_1$:

$$\begin{aligned}P_1(OU \text{ gives objective utilities}) &= 0.9 \\ P_1(OU' \text{ gives objective utilities}) &= 0.1\end{aligned}$$

Her second future self disapproves of soldiering, and so must have increased her credence that $OU'$ is the objective utility function—we might assume that the credence function of that future self is $P_2$:

$$\begin{aligned}P_2(OU \text{ gives objective utilities}) &= 0.1 \\ P_2(OU' \text{ gives objective utilities}) &= 0.9\end{aligned}$$

Then, if $U_1$ is the subjective utility function of her first future self, and $U_2$ the subjective utility function of her second, we have:

$$\begin{array}{lclclcl} V_1(\mathit{Enlist}) & = & U_1(\mathit{Soldier} \,\&\, \mathit{Enlist}) & = & 0.9 \times 9 + 0.1 \times 1 & = & 8.2 \\ V_1(\mathit{Don't\ Enlist}) & = & U_1(\mathit{Civilian} \,\&\, \mathit{Don't\ Enlist}) & = & 0.9 \times 1 + 0.1 \times 9 & = & 1.8 \end{array}$$

And vice versa for $U_2$ and $V_2$:

$$\begin{array}{lclcl} V_2(\mathit{Enlist}) & = & U_2(\mathit{Soldier}\ \&\ \mathit{Enlist}) & = & 0.9 \times 1 + 0.1 \times 9 & = & 1.8 \\ V_2(\mathit{Don't\ Enlist}) & = & U_2(\mathit{Civilian}\ \&\ \mathit{Don't\ Enlist}) & = & 0.9 \times 9 + 0.1 \times 1 & = & 8.2 \end{array}$$

And recall: currently, Effie thinks it's just as likely that she'll end up as her first future self as it is that she will end up as her second future self; and she thinks that whether she chooses *Enlist* or *Don't Enlist* will make no difference to which she will be. So, for $i = 1, 2$:

$$P(P_i \text{ gives credences at } 30 || \mathit{Enlist}) = 0.5$$

and

$$P(P_i \text{ gives credences at } 30 || \mathit{Don't\ Enlist}) = 0.5$$

Notice, therefore, that $P$ satisfies the Reflection Principle. For instance:

$$\begin{aligned} & P(OU \text{ gives the objective utilities} || \mathit{Enlist}) = \\ & \quad 0.5 = 0.5 \times 0.9 + 0.5 \times 0.1 = \\ & \quad\quad P(P_1 \text{ gives credences at } 30 || \mathit{Enlist}) P_1(OU \text{ gives the objective utilities}) + \\ & \quad\quad\quad P(P_2 \text{ gives credences at } 30 || \mathit{Enlist}) P_2(OU \text{ gives the objective utilities}) \end{aligned}$$

Now, to which evaluation function should Effie appeal when she chooses between *Enlist* and *Don't Enlist*? Should she appeal to $V$, the value function of her current self? Or should she appeal to some aggregate of the evaluations functions $V_1$ and $V_2$ of her two possible future selves? In fact, it turns out that it doesn't matter. Effie's current value for each option—*Enlist* and *Don't Enlist*—is equal to her expectation of her future value for that option:

$$\begin{aligned} & V(\mathit{Enlist}) = 5 = 0.5 \times 1.8 + 0.5 \times 8.2 = \\ & \quad\quad P(P_1 \text{ gives credences at } 30 || \mathit{Enlist}) V_1(\mathit{Enlist}) + \\ & \quad\quad\quad\quad P(P_2 \text{ gives credences at } 30 || \mathit{Enlist}) V_2(\mathit{Enlist}) \end{aligned}$$

And similarly for $V(\mathit{Don't\ Enlist})$. And of course, the expectation of the two possible future evaluation functions is the natural way to aggregate them.

Thus, there is no dilemma: Effie doesn't need to choose between using her current or future evaluations of the two options to make her decision; she can use either and the resulting evaluations will agree.

So, this version of the objectivist solution solves the problem of choosing for changing selves in a certain class of cases—a class to which Effie's decision problem in Army belongs. What are these cases? There are two sorts: in the first, the individual is rationally *required* to satisfy the Reflection Principle; in the second, they are rationally *permitted* to satisfy the Reflection Principle, and they do *in fact* satisfy the Reflection Principle. In either of those cases, there is no question whether to use the individual's current subjective utilities, their past subjective utilities, or some amalgamation of their past, current, and future subjective utilities. Their evaluation of the options from the vantage point of their current subjective utilities is the same as their expectation of their evaluation of those options from the vantage point of their future subjective utilities.

## 5.2 The Weak Reflection Principle

However, there are cases that lie outside this class; and, in those cases, the objectivist cannot rely on the Reflection Principle. Recall, for instance, Fernando from the introduction. Fernando's pension scheme presents him with an offer. Opt in, and they will donate 10% of his pension payments to effective charities when he retires; opt out, and they will pay the full pension directly to him. Fernando knows that, when he retires, his values will have changed. Now, he values giving money to those in need; when he retires, he'll prefer to give it to his children. To which set of values should he appeal when he's making his decision? His current values? His values upon retirement? Or some compromise between them?

On the objectivist's account, what happens in Fernando's case is that he knows that his credences will change over time, and what's more he knows the direction in which they will move. His credence will shift away from the hypothesis ($H_1$) that it is objectively more valuable to give money to charity than to give it to his children; and it will shift toward the reverse hypothesis ($H_2$) that it is objectively more valuable to give money to his children. And he knows that this will happen. The Reflection Principle, therefore, deems Fernando irrational. His current credences are not his expectation of his future credences—his current credence in $H_1$ is high and in $H_2$ is low, while his known future credence in $H_1$ is low and in $H_2$ is high. So his current self will choose to opt in to the pension scheme offer, whereas his future self would choose to opt out.

What should Fernando do? We might say that he is irrational to violate the Reflection Principle. Thus, he should amend his current credences by bringing them into line with his known future credences. And he should choose to opt out on the basis of those. Or, alternatively, we might say that the Reflection Principle is not a universal principle of rational credence, and allow that this is a case in which it is rationally permissible for Fernando to violate it. In that case, we must still say whether Fernando should choose how to respond to his pension scheme offer based on his current credences about the objective utilities, or based on his future credences when he retires. Intuitively, we should go for the second option and reject the Reflection Principle; and when that leaves us with the question which of his credences about the objective utilities he should use in decision-making, we should say that Fernando should choose based on his current credences.

This reaction is analogous to our intuitive reaction to certain standard putative counterexamples to the Reflection Principle. Suppose, for instance, that I have just ingested psilocybin ('magic mushrooms'), but the effects have yet to kick in. Based on past experiences with the drug, I know that, in ten minutes, I will have very high credence that I am floating in outer space. Surely this knowledge does not oblige me to set my current credences in line with my known future ones; surely I am not obliged to set my current credences to my current expectations of my future credences and thereby set a high current credence that I am floating in outer space. But that is what the Reflection Principle demands.[27]

When we recall the original motivation for the Reflection Principle, we can see why I am rationally permitted to violate it in the psilocybin case. We motivate the Reflection Principle by saying that future credences are better informed than past or present credences, and thus a better guide to the truth. But of course, in the psilocybin case, this is not true. While psilocybin might throw open the doors of perception and reveal to me certain truths that I have failed to appreciate hitherto, it is rarely a reliable source of credences about the location of my body in relation to the Earth. And so I have no rational obligation to defer to my future, psilocybin-induced credences on that topic—indeed, I have a rational obligation to not defer to those credences.

This suggests the following weakening of the Reflection Principle:

> **Weak Reflection Principle (WRP)** Suppose $P$ is my current credence function; and suppose $P_1, \ldots, P_n$ are all my possible future

[27] See (Talbott, 1991; Briggs, 2009) for similar objections.

> credence functions at some particular future time $t$. Suppose, furthermore, that I know that, at $t$, my credence functions will be rational, and that I will have acquired them from my current ones by a rational process. Then, for any act $a$ in $\mathcal{A}$,
>
> $$P^a(-) = \sum_{i=1}^{n} P^a(P_i \text{ gives credences at } t) P_i^a(-)$$

The idea is this: first, deferring to my future credences in the psilocybin case seems wrong because I take my future credences to be irrational credences for me to have at that time, given the evidence I know I'll have at that time, and moreover I take them to be obtained from my current credences by an irrational process; second, when I instead think that my future credences will be rationally formed from my current ones, I should defer to my future credences because they will be at least as well informed as my current ones. So the Weak Reflection Principle captures what is right about the Reflection Principle, but jettisons what is wrong.

Let's apply this now to Fernando's case. He is obliged to defer to his future credences insofar as he thinks they will be better informed than his. But our natural reaction to his case is to say that his future credences differ from his current credences not because he becomes better informed between now and then, but rather because his credence evolve in some purely irrational way during this time—this shift not only proceeds in an irrational manner, but also lands Fernando with credences that are irrational for him to have given his evidence. In such cases, the Weak Reflection Principle does not demand that you bring your current credences in line with your known future credences. Since those future credences about $H_1$ and $H_2$ are irrational, Fernando should not choose on the basis of them. So, he must choose instead on the basis of his current credences. That is, he should choose to opt in. And that verdict matches our intuitive reaction to Fernando's case.

## 5.3 Permissivism and learning from others

So far, then, we have two sorts of case: Effie's and Fernando's. In Effie's case, she is permitted to satisfy the Reflection Principle, for she does take her future credences to be better informed than hers and to evolve from hers in a rational manner. And indeed she does satisfy that principle. The objectivist's version of the problem of choosing for changing selves then dissolves, because Effie's current credences incorporate her past and future credences; what's more, as we saw in Theorem 5.1.1, her evaluation of an option from the point of view of her current credences is the same as her expectation of its evaluation from the point of view of her future credences. In Fernando's case, on the other hand, he is permitted to violate the Reflection Principle because he judges his future credences to be irrational. And indeed, for the same reason, he is mandated not to take his future credences into consideration, and to make his decision on the basis of his current credences only. The problem of choosing for changing selves again dissolves, but this time because there is only one rational set of credences Fernando might use to evaluate the options, namely, his current credences, and so he appeals to them. Now, if the Weak Reflection Principle is true, these are the only sorts of case: either your future credences are rational, in which case, the Weak Reflection Principle makes the same demands as the Reflection Principle and we can appeal to the solution we gave for Effie's case; or your future credences are irrational, in which case, there is only one rational vantage point from which to make your choice, which is your current credences, and so you should use those. If these were the only sorts of case, we would have our solution to the objectivist's version of the problem of choosing for changing selves. But there is a third sort of case.

Consider Fernando's partner, Franklin. Franklin's pension scheme is offering something slightly different. Fernando's choice is this: (i) donate 10% of his pension to effective charities; or (ii) donate none of his pension to effective charities. Franklin's choice, on the other hand, is this: (a) donate 10% of his pension to effective health charities, (b) donate 10% of his pension to effective educational charities, and (c) donate none of his pension to any charity. Currently, Franklin values (a) more than (b) more than (c). But he knows that, come retirement, he will value (b) more than (a) more than (c). He has watched many of his family and acquaintances shift their priorities from health to education as they age. Thus, like Fernando, Franklin violates the Reflection Principle. However, unlike Fernando, he also violates the Weak Reflection Principle. This is because, while Franklin currently values (a) more than (b), and has the credences concerning the objective utilities that reflect that, he does not think it is irrational to value (b) more than (a), and thus thinks that the corresponding credences concerning the objective utilities are rationally permissible too. What's more, Franklin doesn't just think that his future subjective utilities and corresponding credences at retirement age are rational; he also thinks that he will shift towards them from his current subjective utilities and corresponding credences in a rational manner in the intervening period. Typically, in such a situation, we don't think that we ought to simply adopt those rationally permissible future credences, which is what the Weak Reflection Principle would have us do. Thus, such cases provide a counterexample to the Weak Reflection Principle, and thus a counterexample to the proposed solution to the problem of choosing for changing selves proposed above, which is based on that principle.

To see this more clearly, consider a case that does not involve credences concerning objective utilities. Consider a possible major financial event in the future—a crash on the stock market, for instance. I collect extensive data that is relevant to predicting whether the crash will occur; I analyse the data; I set my credence in the proposition that it will occur. It seems that there may well be two or more different credences I could rationally assign to that proposition; two equally rational responses to this complex body of evidence. The view that there could be such bodies of evidence is known as *permissivism*; its negation, which says that, to each body of evidence, there is a unique rational response, is known as the *uniqueness thesis*.[28] Suppose we accept permissivism; and suppose we currently have a credence function $c$; but, in a few hours, during which we'll learn nothing new, we will have shifted to a different credence function $c'$. Both $c$ and $c'$ are rationally permissible responses to the evidence. And suppose I know all this upfront. Then, according to the weakened version of the Reflection Principle, I am irrational—my current credences are not my expectations of my future credences, which I know will be given by $c'$. However, it seems intuitively that I am perfectly rational. Of course, you might worry if I simply oscillated repeatedly between the different possible rational responses to the evidence.[29] But we might assume that I haven't done this. Rather, I've moved gradually, over the course of the two hours, from $c$ to $c'$, always passing through other credence functions that are rationally permissible. To make it concrete, suppose my evidence warrants any credence between 0.44 and 0.47 in the proposition that the stock market will crash in the coming month. And suppose I start with credence 0.45, and I move continuously through the intermediate credences over the space of two hours, always getting more confident, until I have credence 0.46 by the end. Then I think we would allow that I am rational. But, if I knew that I would do that, and nonetheless retained my credence of 0.45 at the start, then I would violate even the Weak Reflection Principle.

The problem is this: the Reflection Principle relies on the assumption that future credences are better informed than current credences, but the

[28] For a survey of the literature on these two positions, see (Kopec & Titelbaum, 2016).

[29] This would be an epistemic version of what Richard Kraut calls 'brute shuffling' in the case of intentions (Paul, 2014b, 344).

psilocybin case refutes that; the Weak Reflection Principle relies on the assumption that *rational* future credences are better informed than current credences, but the stock market case refutes that—in that case, the rational credences I will come to have in the future are neither better nor worse informed than the rational credences I currently have; they are just different rational credences. So neither the Reflection Principle nor the Weak Reflection Principle is a universal principle of credal rationality. And thus the objectivist cannot hope to give a comprehensive solution to the problem of choosing for changing selves by appealing to either of them.

How should the objectivist treat such cases? To answer that, first note that the (Weak) Reflection Principle is really the conjunction of two claims:

(TTP) My current credence in a proposition should be my expectation of the credence I would have were I to learn my future credence in that proposition.

That is,

$$P^a(-) = \sum_{i=1}^{n} P^a(P_i \text{ gives my credences at } t) P^a(-|P_i \text{ gives my credences at } t)$$

(RP$^*$) Were I to learn my future credence in a proposition, I ought to set my credence in that proposition in line with that future credence (providing that future credence is rational given my future evidence and the move from my current credence to my future credence is rational).

That is,

$$P^a(-|P_i \text{ gives my credences at } t) = P^a_i(-)$$

Now (i) is simply a theorem of the probability calculus. It is a particular case of what is often called *the theorem of total probability*, which says that, if $E_1, \ldots, E_n$ are mutually exclusive and exhaustive propositions, then

$$P(-) = \sum_{i=1}^{n} P(E_i)P(-|E_i)$$

Thus, the controversial claim is (ii). Now, as Ray Briggs (2009) points out, (ii) is itself just a theorem of the probability calculus if you are certain that your future credence function will be obtained from your current credence function by conditionalizing on some proposition that you learn with certainty. After all, if that's the case, learning your future credence function is tantamount to learning the strongest proposition you will learn between now and that future time. Hence, you should conditionalize on that proposition and thereby obtain a credence function that matches the future credence function, which, by hypothesis, was also obtained by conditionalizing on that proposition. (ii) then follows. However, in other situations, this isn't the case. Indeed, the example of the stock market crash from above is a counterexample to (ii). As we saw, upon learning that my future credence in the crash will be 0.46, I am under no rational obligation to set my credence to that. This raises the question: am I under any rational obligation to change my credence at all, and in any particular way? That brings us to the debate about peer disagreement.

In that debate, we are interested in whether I should change my credences or beliefs when I encounter an epistemic peer who disagrees with me—that is, someone with the same evidence as me and who is rational to the same degree, but who has different beliefs or credences. Traditionally, in that debate, there has been a stand-off between the *steadfasters*, who say that you ought to stick with your credences in the face of disagreement, and the *conciliationists*, who say that you should move at least some way towards the beliefs and credences of your disagreeing peer. But, lately, there has been a growing recognition that there is no single answer that covers all cases (Easwaran et al., 2016; Titelbaum & Kopec, ta)—different cases call for different responses.

The Reasoning Room case due to Titelbaum & Kopec (ta) illustrates the point well:

> **Reasoning Room** "You are standing in a room with nine other people. Over time the group will be given a sequence of hypotheses to evaluate. Each person in the room currently possesses the same total evidence relevant to those hypotheses. But each person has a different method of reasoning about that evidence.
>
> "When you are given a hypothesis, you will apply your methods to reason about it in light of your evidence, and your reasoning will suggest either that the evidence supports belief in the hypothesis, or that the evidence supports belief in its negation. Each other person in the room will also engage in reasoning that will yield exactly one of these two results.
>
> "This group has a well-established track record, and its judgments always fall in a very particular pattern: For each hypothesis, 9 people reach the same conclusion about which belief the evidence supports, while the remaining person concludes the opposite. Moreover, the majority opinion is always accurate, in

> the sense that whatever belief the majority takes to be supported always turns out to be true.
>
> "Despite this precise coordination, it's unpredictable who will be the odd person out for any given hypothesis. The identity of the outlier jumps around the room, so that in the long run each agent is odd-person-out exactly 10% of the time. This means that each person in the room takes the evidence to support a belief that turns out to be true 90% of the time." (Titelbaum & Kopec, ta, 16)

Now suppose a hypothesis comes in. You engage in your method of reasoning, and the others in the room engage in theirs. Your methods indicate that the hypothesis is true. You thus adopt a credence of 90% that it is true. Now, suppose that you pick someone at random from the room and ask for their credence. They report that they are only 10% confident that the hypothesis is true—that is, they disagree with you. How should you respond? One thing their report might teach you is that it is rational to be 10% confident in the hypothesis based on your evidence—that is, there is some rational method of reasoning that leads from the evidence to that credence in the hypothesis. But in fact you already knew that because you knew about the set up of the room—for any hypothesis that comes in, it is rationally permissible to be 10% confident in it. And, what's more, you also knew before you encountered this individual that there would be someone in the Reasoning Room who has that credence. You might think, therefore, that your encounter teaches you nothing, and so you should not shift your credence from 90%. And indeed Stew Cohen (2013) has defended that view. He claims that, if you already know what the rationally permissible possible credences are, and thus know that there are possible rational individuals who hold them, then learning that an actual rational individual holds one of them should not affect you. But that's wrong, as the Reasoning Room example illustrates—indeed, Titelbaum and Kopec introduce the example to make precisely this point.

To see this why Cohen is wrong, suppose you are standing in front of an urn. You know that it is one of 10 urns—9 of these are red-majority urns, containing 9 red balls and 1 blue ball, and 1 is a blue-majority urn, containing 9 blue balls and 1 red ball. You have equal credence in each of the ten possibilities. So you are 90% confident that the urn is red-majority. Now you draw a ball at random—it is blue. Now, of course, you knew before you drew the ball that the urn contains a blue ball—either the only one or one of nine. So you haven't learned anything new there. But nonetheless you should shift your credence that the urn is red-majority. After all, while you already knew that the urn contained a blue ball, you didn't know that the first ball you drew at random would be blue. And that is much more likely if the urn if blue-majority than if it is red-majority. Indeed, you should shift your credence that the urn is red-majority from 90% to 50%, as the following calculation using Bayes' Theorem shows:[30]

$$\begin{aligned}
& P(\textit{Red urn}|\textit{Blue ball}) \\
= & \frac{P(\textit{Blue ball}|\textit{Red urn})P(\textit{Red urn})}{P(\textit{Blue ball}|\textit{Red urn})P(\textit{Red urn}) + P(\textit{Blue ball}|\textit{Blue urn})P(\textit{Blue urn})} \\
= & \frac{\frac{1}{10} \times \frac{9}{10}}{\left(\frac{1}{10} \times \frac{9}{10}\right) + \left(\frac{9}{10} \times \frac{1}{10}\right)} = \frac{1}{2}
\end{aligned}$$

For the same reason, in the Reasoning Room example, you should shift your credence in the hypothesis from 90% to 50% upon learning that the person you picked at random has credence 10% in the hypothesis. After all, just as drawing a blue ball at random is much more likely if the urn is blue-majority, so picking at random someone who is 10% confident in the hypothesis is much more likely if the hypothesis is false. The calculation that shows you should shift from 90% to 50% confidence exactly matches the calculation in the urn case.

$$\begin{aligned}
& P(\textit{Hypothesis}|\textit{10\%}) \\
= & \frac{P(\textit{10\%}|\textit{Hypothesis})P(\textit{Hypothesis})}{P(\textit{10\%}|\textit{Hypothesis})P(\textit{Hypothesis}) + P(\textit{10\%}|\overline{\textit{Hypothesis}})P(\overline{\textit{Hypothesis}})} \\
= & \frac{\frac{1}{10} \times \frac{9}{10}}{\left(\frac{1}{10} \times \frac{9}{10}\right) + \left(\frac{9}{10} \times \frac{1}{10}\right)} = \frac{1}{2}
\end{aligned}$$

Next, suppose that, the person you pick at random in the Reasoning Room has credence 90% in the hypothesis, rather than 10%. In that case, you become certain that you're in the majority and your credence in the hypothesis should jump from 90% to 100%. After all, in the set up of the case, the majority is always right.

What these examples illustrate is that different situations mandate different responses. In the first case, we responded to disagreement by adopting

[30] **Bayes' Theorem** If $P$ is a probability function,

$$P(X|E) = P(E|X)\frac{P(X)}{P(E)} = P(E|X)\frac{P(X)}{P(E|X)P(X) + P(E|\overline{X})P(\overline{X})}$$

a compromise between our credence and the disagreeing peer's. In the second, we responded to agreement by increasing our credence beyond the credence we share with our peer. What they also illustrate is that the correct response to learning the beliefs and credences of others is often dictated by what we believe about those beliefs and credences as indicators of the truth. Essentially, our response in the Reasoning Room cases turn not primarily on principles that govern peer disagreement, but rather on principles that govern how credences concerning the objective chances relate to credences in other propositions. We call these chance-credence norms, and the most well known is the Principal Principle, which says (Lewis, 1980; Hall, 1994, 2004; Ismael, 2008; Pettigrew, 2014):

> **Principal Principle** Suppose $P$ is my credence function, and $ch$ is a probability function. Then:
>
> $$P(-|ch \text{ gives the current objective chances}) = ch(-)$$

Notice that it is this principle that was at play above when we calculated $P(\textit{Red urn} \,|\, \textit{Blue ball})$. There, we assumed that

- $P(\textit{Blue ball} \,|\, \textit{Red urn}) = \frac{1}{10}$
- $P(\textit{Blue ball} \,|\, \textit{Blue urn}) = \frac{9}{10}$.

And that is because (i) we know that, if the urn is red-majority, then the chance of a blue ball is 10%, while it is 90% if the urn is blue-majority, and (ii) we assume that $P$ satisfies the Principal Principle. Similarly, when we calculate that, upon hearing that your random pick in the Reasoning Room has credence 10%, you should shift your credence in the hypothesis to 50%, we assume that

- $P(\textit{10\%} \,|\, \textit{Hypothesis}) = \frac{1}{10}$
- $P(\textit{10\%} \,|\, \overline{\textit{Hypothesis}}) = \frac{9}{10}$.

And, again, behind both of those assumptions is the assumption that $P$ satisfies the Principal Principle.

The point is this: By the theorem of total probability, my current credence must be my expectation of the credence I would have were I to learn my future credence. That much is uncontroversial. The (Weak) Reflection Principle adds to that the claim that, when I take my future credence to be rational, if I learn what it is, I should set my updated credence in line with it. As we saw above, if I know that my future credence will be obtained from my current credence by conditionalization, the Reflection Principle is a theorem, not an additional assumption. But, as we also saw above, in other situations, different responses are appropriate. For instance, we might imagine a version of the Reasoning Room in which the nine other people are nine of your future selves, and you pick one at random to learn their credence in the hypothesis. In this situation:

- $P(\mathit{Hypothesis} \,|\, \mathit{90\%}) = 1 \neq 0.9$
- $P(\mathit{Hypothesis} \,|\, \mathit{10\%}) = \frac{1}{2} \neq 0.1$

In these other cases, where my future credences are rational, but not obtained from my current credences by conditionalizing on a proposition, the way in which I incorporate my future credences into my current credences is dictated by my beliefs about those future credences as indicators of the truth. Thus, the appropriate response is often dictated by my knowledge of the objective chances and something like the Principal Principle.

## 5.4 The objectivist's version of our problem

Let's return to where we started. Above, I noted that, while decision theorists might puzzle over which of my values—past, present, or possible future—determine the utility function I should use to make decisions, epistemologists never puzzle over which of my beliefs and credences—past, present, or possible future—determine the credence function I should use. If the Reflection Principle were true, it would explain this phenomenon. After all, it explicitly describes how your anticipated possible future credences should be incorporated into your current ones. So there is no need to choose between them. But the Reflection Principle is false.

If the Weak Reflection Principle were true, then it would also explain the phenomenon. If I take my future credences to be rational, the Reflection Principle holds of me and I can appeal to the explanation it provides; if I do not, then I have a principled reason to choose on the basis of my current credences alone—after all, my future credences are irrational. But the Weak Reflection Principle is also false. However, as we saw in the case of the Reasoning Room, we can still explain why we need not worry about whether we should look to our current credences, or to our past or future ones. After all, it is simply a theorem of the probability calculus that our current credences should be our expectations of the credences we would have were we to learn our future credences. So those other anticipated credences are incorporated into our current credences, just not always in the way that the Reflection Principle envisages. Sometimes, upon learning of them, I will compromise with them; sometimes I will defer to them; and sometimes I will move away from them, as the Reasoning Room case illustrates.

Finally, then, we return to the objectivist's proposed solution to the problem of choosing for changing selves. For the objectivist, as for the subjectivist, when I face a decision problem, I should pick one of the options that maximises expected subjective utility, where my expected subjective utility for $a$ is

$$V(a) = \sum_s P^a(s)U^a(s)$$

However, for the objectivist, and not for the subjectivist, my subjective utility for a situation is my expected objective utility for it, where this is

$$U^a(s) = \sum_{OU} P^a(OU \text{ gives objective utilities})OU^a(s)$$

Thus, if we assume the Independence Principle from above, the objectivist takes my expected subjective utility for $a$ to be:

$$V(a) = \sum_{s,OU} P^a(s \ \& \ OU \text{ gives objective utilities})OU^a(s)$$

For the objectivist, then, the problem of choosing for changing selves is this: which doxastic attitudes should be recorded in $P$? My current, past, or possible future credences? Our answer is this: $P$ should record my current credences. Why? First, note that these credences already incorporate the information that any knowledge I have of my future credences can provide for me, since the theorem of total probability demands that they are the expectations of the credences I would have were I to learn my future credences. Of course, since this is simply a consequence of my current credences' adherence to the probability axioms, the same may be said of my past and future credences. So the mere fact that my current credences incorporate what information I can glean from my past and future credences does not itself pick out my current credences uniquely. So why not use one of those instead?

The answer, I think, lies in a crucial difference between credences, on the one hand, and the subjectivist conception of values or utilities, on the other. While we all have different credences, we are all trying to do the same thing by having them; our goal is the same. We are all trying to get at the truth about the world; it's just that we have different evidence about how the world is, and sometimes we also have different rational responses to the evidence we have.[31] With utilities, it's different. These characterise our different views of what matters in the world. At least for the subjectivist, this isn't an attempt to get at the truth about anything; it isn't an attempt to get closer to a set of objective utilities. As a subjectivist, I don't think that your penchant for conformity is further away from some objective truth about what is valuable than my desire for self-direction, for instance.

Suppose that we all share the same utilities, but have different credences. I have to make a decision on behalf of us all. I ask for everyone's credences, I update my credences on the basis of this information, and I make the decision on the basis of those updated credences and our shared utilities. Notice that you might say to me that I would have done better to use your credences, updated on the information about everyone else's credences. But you would not accuse me of unfairness. I haven't failed to treat you justly. The reason is that there is no value of yours that I've ignored, nor to which I've accorded insufficient weight. I have simply used the information about your credences, together with my beliefs about the ways in which your credences are indicators of the truth, to arrive at new credences on the basis of which to make our collective decision. My new credences are my best efforts to arrive at a doxastic ingredient to use when I make our collective decision; and my efforts in this area are guided by trying to best achieve our shared doxastic goal of getting at the truth about the world.

Next, suppose that we all have different utilities. And suppose I'm a subjectivist about those utilities—I don't think that they aim at some objectively correct utilities. Again, I have to make a decision on behalf of us all. I ask for everyone's utilities, I update my credences on the basis of this information, and I make the decision on the basis of those updated credences and my own subjective utilities. In this case, you might very well accuse me of unfairness. I have failed to treat you justly. And the reason is that I have not given appropriate weight to your values—updating my credences in the light of information about your utilities is not sufficient.

What is true of such group decision making is also true when I make decisions on behalf of my past, present, and possible future selves. My cur-

[31] In fact, there is a subtlety here: many epistemologists think that there are various different permissible ways to ascribe value to doxastic states. Even amongst those who agree that it is only accuracy, or fit with the world, that matters, many think that there are various different ways you might measure that accuracy (Joyce, 1998, 2009; Levinstein, 2017). There's an interesting question about how you should set your credences in the light of this fact. Should you try to achieve the goal of accuracy as measured by the lights of your own current way of measuring it, or should you use some amalgamation of the measures endorsed by your past, present, and possible future selves? To discuss this would take us too far astray.

rent credences are my best attempt to formulate credences that achieve the doxastic goal I share with my past and possible future selves. They give sufficient consideration to the doxastic states of those other selves, since they use the information they have about those other states in order to try to make themselves as good as possible from the point of view of our shared doxastic goal. My past and possible future selves have no cause for complaint.

Recall, then, the case of Franklin from above. He knows that his future subjective utilities will change between now and his retirement—now, he values health charities more than educational ones; by the time he retires, his values will have switched. Thus, on the objectivist account, his credences concerning the true objective utilities will change—he will become more confident that educational charities have greater objective utility than health charities. And we suppose that both his current values and his retirement values are permissible, and so his current credences concerning the objective values, and his retirement credences are rational. Now, when Franklin learns what his future credences concerning the objective values will be, he should incorporate that information by updating on it. But once he has incorporated it, he should choose on the basis of his new current credences. Thus, if he is a steadfaster in this situation, he should stick with his original current credences and support the health charity; if he defers entirely to his future self, he should bring his current credences in line with his retirement self and support the educational charity; if he is a conciliationist, he should move some way towards his retirement self, and choose whatever maximises expected utility from that perspective; and so on.

## 5.5 The implausible strength of objectivism

For proponents of the One True Utility Solution, then, there is a solution to the problem of choosing for changing selves. What's more, it is beguiling in its simplicity. After all, it renders the problem epistemic. And indeed, on the objectivist view, the problem becomes a particular case of a general problem in epistemology: how should I learn from others? how should I update my credences in the light of information about the doxastic states of other people or of my other selves?

Nonetheless, I don't think we should adopt the objectivist solution. The problem is not that it claims that there are objective facts about what is valuable. Most people agree that there are such facts. Most agree that the destruction of the world has less objective value than the scratching of my little finger. The problem is rather that the objectivist solution posits too many such facts. It claims not only that the world's annihilation is less value than damage to my finger, but also that there is an objective fact whether self-determination is better than conformity, whether a life spent amongst nature is better than one spent amongst books, whether studying particle physics is more valuable than playing in a band, whether health is more important than education, and so on. And indeed, not only must it say that there are these objective facts about the ordering of situations by value, but also that the extent to which one is better than the other is a matter of objective fact. After all, in order for the objectivist solution to work, there must be a single true objective utility function. If we wish to use standard expected utility theory and represent your subjective values at a time by a single utility function (up to positive linear transformation), and we take the objectivist view that subjective utilities are just the shadows cast by your credences about objective utilities and, in particular, your expectations of those objective utilities, then we must take there to be just one objective utility function. But, as we have seen, that is extremely implausible.

Of course, if we lift the restriction that your subjective values are represented by a single utility function, then we can similarly give a more liberal version of objectivism. But that must wait until Chapter 16.

## 5.6 Appendix: Proof of Theorem 5.1.1

**Theorem 5.1.1** *Suppose $P$ is my current credence function; and suppose $P_1, \ldots, P_n$ are all my possible future credence functions at some particular future time $t$. And suppose that I satisfy the Reflection Principle and the Independence Principle. Then, for any act $a$ in $\mathcal{A}$,*

$$V(a) = \sum_{i=1}^{n} P^a(P_i \textit{ gives my credences at } t)V_i(a)$$

*where*

- $V_i(a) = \sum_{s \in \mathcal{S}} P_i^a(s) U_i^a(s)$, *and*

- $U_i^a(s) = \sum_{OU} P_i^a(OU \textit{ gives the objective utilities}) OU^a(s)$.

*Proof.*

$$
\begin{aligned}
& V(a) \\
= \quad & \sum_s P^a(s)U(s) \\
= \quad & \sum_s P^a(s) \sum_{OU} P(OU \text{ gives obj utilities}) OU^a(s) \\
= \quad & \sum_{s,OU} P^a(s) P(OU \text{ gives obj utilities}) OU^a(s) \\
= \quad & \sum_{s,OU} P^a(s \text{ \& } OU \text{ gives obj utilities}) OU^a(s) \quad \text{(by IP)} \\
= \quad & \sum_{s,OU,i} P^a(P_i \text{ gives creds at } t) P_i^a(s \text{ \& } OU \text{ gives obj utilities}) OU^a(s) \quad \text{(by RP)} \\
= \quad & \sum_i P^a(P_i \text{ gives creds at } t) \sum_s P_i^a(s) \sum_{OU} P_i^a(OU \text{ gives obj utilities}) OU^a(s) \\
= \quad & \sum_i P^a(P_i \text{ gives creds at } t) \sum_s P_i^a(s) U_i^a(s) \\
= \quad & \sum_i P^a(P_i \text{ gives creds at } t) V_i(a)
\end{aligned}
$$

as required. □

# Chapter 6

# The Aggregate Utility Solution I: which attitudes to aggregate?

Over the past three chapters, we have seen three putative solutions to the problem of choosing for changing selves: the Unchanging Utility Solution, the Utility of Utility Solution, and the One True Utility Solution. The first two treat some cases correctly—Erik's and Giang's, respectively—and we will incorporate their insights into our own favoured solution. The third has the potential to treat all correctly, but it relies upon an extremely strong version of objectivism.

In this chapter, I formulate my own favoured solution. In fact, I will describe a particular species of solution. When I describe my solution in this part of the book, I will leave unspecified a number of parameters; we obtain a different particular instance of that species for each way we might set those parameters. It is the purpose of the second part of the book to discuss how we might set those parameters. We will call our solution the *Aggregate Utility Solution* for reasons that will quickly become obvious.

The problem of choosing for changing selves arises when my past, current, and future selves do not all share the same values. Or, at least, it arises when this might be the case. So, perhaps better: the problem arises when my past selves, my current self, and my *possible future selves* do not all share the same values. The solution I wish to propose begins with the observation that, presented in this way, our problem can be viewed as a judgment aggregation problem. In a judgment aggregation problem, we take the attitudes of each member of a group of individuals and ask what the aggregate attitude of the whole group taken together is. This is precisely our problem here—the individuals in question are my different selves, past, present, and future; the group of them is me, the corporate entity that comprises them. In this sense, our problem is analogous to a variety of other judgment aggregation problems. For instance, we face a judgment aggregation problem when we wish to combine the probabilistic beliefs of individual climate scientists to give the probabilistic beliefs of the whole climate science community taken together concerning, say, sea level rise in the coming twenty years, or global mean surface temperatures in 2100.[32] And we face another one when we try to aggregate the preferences of the citizens of a democratic country in order to determine the government they will have.[33] And we encounter yet another judgment aggregation problem when we are uncertain which moral theory is correct, and we need to aggregate the judgments of each of the competing theories concerning the morally permissible actions in order to decide what to do ourselves.[34] And so on. What is clear from these examples is that, in judgment aggregation problems, the sorts of judgments we wish to aggregate might be quite varied—from probabilistic beliefs to preferences to judgments of moral permissibility, and in our case subjective utilities or values—and the sorts of entities making those judgments might be quite diverse—from individual scientists to individual citizens to moral theories, and in our case different selves belonging to the same person.

I propose that we treat the problem of choosing for changing selves as a judgment aggregation problem. Viewed in this way, our question is as follows: how should we aggregate the attitudes of my past selves, my current self, and my possible future selves to give the collective attitudes of the group of those selves when taken together? That is, how should we aggregate the attitudes of my various selves to give *my* attitudes as the corporate entity that comprises those selves? In this chapter and the next, I explore this proposal by exploring the different ways in which we might aggregate the judgments of a group of individuals; I conclude with a particular detailed thesis that constitutes the main normative claim of this book.

Now, there are three natural ways to aggregate the judgments of our past, present, and possible future selves. These correspond to the three different levels of attitude that we ascribe to those various selves. As good, card-carrying realists, credences ($P$) and utilities ($U$) are on the lowest, most fundamental level; the evaluation function ($V$), which records the individ-

[32] For surveys of the techniques used for such aggregation in general, see (Genest & Zidek, 1986; Dietrich & List, 2015; Russell et al., 2015). In the particular case of climate science, see (Moss & Schneider, 2000).

[33] See, for instance, (Arrow, 1951; Gaertner, 2009; Sen, 2017).

[34] See, for instance, (Lockhart, 2000; Ross, 2006; Sepielli, 2009; Hedden, 2015; MacAskill, 2016; Hicks, ta).

ual's subjective expected utilities, is on the next level, determined by the credences and utilities; and the preference ordering ($\preceq$) is on the highest level, determined by the evaluation function. So we might aggregate our individuals' attitudes by aggregating their preference orderings (section 6.1). Or we might aggregate our individuals' evaluation functions and determine their aggregate preference ordering on the basis of that (section 6.2)—this is sometimes known as the *ex ante* approach. Or we might aggregate our individuals' credences and aggregate our individuals' utilities, separately, and combine them to give their aggregate evaluation function and determine their aggregate preference ordering on the basis of that (section 6.3)—this is sometimes called the *ex post* approach.[35] We consider them in turn. At each level, we will work initially in the standard framework for social choice theory, where we have a fixed set of individuals and we wish to aggregate their attitudes. After that, we will consider what happens when we then move to our slightly different setting, where we do not have a fixed set of individuals, but instead different possible sets of individuals. Each of these sets contains my past selves and my current self; but each contains a different collection of possible future selves.

## 6.1 Aggregating preferences

We begin with the method of aggregating preference orderings. First, then, the standard case, where we have a group of $n$ individuals with preference orderings over a set $\mathcal{A}$ of possible actions. We want to find a preference ordering that represents the preferences of the whole group of those individuals. More precisely: we want a method that takes a sequence $\langle \preceq_1, \ldots, \preceq_n \rangle$ of preference orderings, which we call a *preference profile*, and returns a single preference ordering $\preceq_G$. And we would like that method to have certain features. In his pioneering work in social choice theory, Kenneth Arrow considered three such features, which we will consider below—Weak Pareto, No Dictator, and Independence of Irrelevant Alternatives. He argued that each is required for a reasonable aggregation method. And then he showed that no aggregation method can have all three features. This is the so-called

[35] Note: the *ex ante*/*ex post* terminology might seem the opposite way round to what the names would suggest. This is because the names come out of a constructivist approach to decision theory on which preferences are most fundamental and credences and utilities are extracted from those afterwards. Thus, from that perspective the *ex ante* method combines preferences *before* the credences and utilities are extracted, while the *ex post* method combines the credences and utilities *after* they have been extracted.

*Arrow Impossibility Theorem*.[36]

Let's meet Arrow's conditions (Arrow, 1951; Gaertner, 2009):

- **Weak Pareto** This says that, if all individuals agree that one option is strictly better than another, the aggregate must agree as well.

  Formally: For any acts $a, b$ in $\mathcal{A}$ and any profile $\langle \preceq_i \rangle$: if $a \prec_i b$, for all $i$, then $a \prec_G b$.

- **No Dictator** This says that there is no individual whose preference ordering is guaranteed to be identical to the aggregate ordering. Such an individual would be a dictator, and so this condition says that there should be no dictator.

  Formally: There is no individual $k$ such that, for any $a, b$ in $\mathcal{A}$ and any profile $\langle \preceq_i \rangle$, $a \preceq_k b$ iff $a \preceq_G b$.

- **Independence of Irrelevant Alternatives** This says that the aggregate ordering of two options depends only on the individual orderings of those two options and not on the ordering of any other options.

  Formally: For any acts $a, b$ in $\mathcal{A}$ and any two profiles $\langle \preceq_i \rangle$, $\langle \preceq'_i \rangle$: if $a \preceq_i b$ iff $a \preceq'_i b$ for all $i$, then $a \preceq_G b$ iff $a \preceq_{G'} b$.

Thus, suppose we are considering the preferences of a group of individuals over three candidates in an election: Marine, Emmanuel, and Jean-Luc. Weak Pareto says, for instance, that if all voters prefer Emmanuel to Marine, then the final ranking of candidates should place Emmanuel above Marine. No Dictator says that there should be no voter such that, however they rank the candidates, the final ranking agrees. And the Independence of Irrelevant Alternatives says, for instance, that the order of Jean-Luc and Marine in the final ranking should depend only on the individual rankings of those two candidates, and not on the order in which any voter ranks Jean-Luc and Emmanuel, nor on the order in which any voter ranks Marine and Emmanuel.

Conditions in social choice theory often fall into one of two categories: they are usually either *unanimity preservation principles* or *dependence principles*. A unanimity preservation principle tells us, for some particular feature

[36] Oddly, Arrow himself called the result his "general possibility theorem", but it is really an impossibility theorem, or, as such results are sometimes called, a *no-go theorem*. It tells us that a certain set of conditions cannot jointly be satisfied. Arrow proved the result originally in his doctoral dissertation and published it in his paper 'A Difficulty in the Concept of Social Welfare' (Arrow, 1950), but it gained widespread influence through his 1951 book, *Social Choice and Individual Values* (Arrow, 1951).

that a preference ordering may or may not have—the feature of preferring *b* to *a*, for instance—that, if each individual has a preference ordering with that feature, then the group preference ordering should also have that feature. The Weak Pareto principle is a unanimity preservation principle, where the feature in question is indeed preferring *b* to *a*. There are two sorts of dependence principle: positive and negative dependence principles. A positive dependence principle tells us that some feature of the group preference ordering—the order in which it ranks *a* and *b*, for instance—should depend only on certain features of the individual preference orderings—how they order *a* and *b*, for instance. Independence of Irrelevant Alternatives is a positive dependence principle, which says that the group ordering of *a* and *b* should depend only on the individual orderings of *a* and *b*. A negative dependence principle, on the other hand, tells us that some feature of the group preference ordering—its ordering of the acts in $\mathcal{A}$, for instance—should *not* depend only on certain features of the individual preference orderings—the orderings of the acts in $\mathcal{A}$ by individual *k*, for instance. No Dictator is a negative dependence principle, which says that the group ordering should not depend only on some particular individual ordering.

As mentioned above, Arrow proved that no method for aggregating the preference orderings of a group of individuals adheres to Weak Pareto, No Dictatorship, and the Independence of Irrelevant Alternatives. Any method that satisfies Weak Pareto and No Dictatorship will make the aggregate ordering of two possible acts depend on the individual orderings of other acts; any method that satisfies Weak Pareto and the Independence of Irrelevant Alternatives will give rise to a dictator; and any method that satisfies No Dictator and Independence of Irrelevant Alternatives will sometimes fail to preserve a unanimous consensus that one act is better than another.

Let's see what these conditions amount to and how plausible they are in our setting. We will have a great deal to say about Weak Pareto in the next section, so I'll leave its treatment until then. So, first, No Dictator. In our setting this says that there should be no single self that always calls the shots. So suppose we have a range of possible acts $\mathcal{A}$. And my various selves, past, present, and possible future, each have their own preference ordering over those acts. Then, if we have a method that will take any profile of such preference orderings—that is, any combination of preference orderings that my various selves might have—and produces an aggregate ordering, No Dictator says that there shouldn't be a single self—my current self, say, or my self at the beginning of my epistemic life, or my 'best' self, the one at the pinnacle of my cognitive and moral life—such that the aggregate preferences are just the preferences of that self. This seems plausible. Interestingly, a dictatorship in which my current self is the tyrant errs in exactly the opposite way to the Unchanging Utility Solution. Recall: our objection to the Unchanging Utility Solution was that it pays too little attention to our current selves, and will always exhort me to change my utilities if by doing so I will obtain more of what I will then come to value. A tyranny of the current self pays too much attention to my current values. It will never let me break free from my current utilities unless doing so somehow serves those utilities. And a dictatorship led by any of my other selves suffers from a similar problem. In cases like Aneri's, the correct solution will surely involve weighing the fit between her current values and the career as a conservation officer, together with the fact that they're her current values, and the fit between her future values and the law enforcement career, together with the fact that they would be her future values. If Aneri were to make one of those selves a dictator, she could not weigh the two considerations in this way.

Next, consider the Independence of Irrelevant Alternatives. Recall: this says that the order in which the group ranks two acts depends only on the orders in which the individuals rank those two acts. Thus, whether the UK Conservative Party ranks Theresa May above or below Andrea Leadsom in their leadership election should depend only on how the individual citizens of that country rank those candidates; it should not depend on how they rank Michael Gove or Stephen Crabb or Liam Fox. Thus:

**Tory Leaders** Consider the two preference profiles below. Voter $i$ ranks May and Leadsom exactly as Voter $i^*$ does, for each $i = 1, 2, 3$. Thus, according to the Independence of Irrelevant Alternatives, the two groups of voters both collectively rank May and Leadsom in the same way—either May above Leadsom or Leadsom above May:

| Voter 1 | Voter 2 | Voter 3 |
|---|---|---|
| **Leadsom** | **May** | **May** |
| **May** | **Leadsom** | Crabb |
| Fox | Gove | **Leadsom** |
| Gove | Fox | Fox |
| Crabb | Crabb | Gove |

| Voter 1* | Voter 2* | Voter 3* |
|---|---|---|
| **Leadsom** | **May** | Gove |
| Fox | Gove | Fox |
| Crabb | **Leadsom** | Crabb |
| Gove | Fox | **May** |
| **May** | Crabb | **Leadsom** |

Or, in our framework:

> **Careers** You and I both know, let us suppose, that our values will change over the next two years. Each of us must now decide whether to take a job as a librarian, a park ranger, a carpenter, an actor, or a police officer. At the moment, I value learning from books more than creating things myself; I value creating things more than spending time in the natural world; I value that more than performing in front of others; and I value performing more than following rules. Thus, I currently rank the career choices before me as follows: first, librarian, then carpenter, then ranger, then actor, then police officer. You rank them differently: librarian, police officer, ranger, actor, carpenter. And, for each of us, our two possible future selves have different rankings again. Here they are—they are structurally identical to those in the Tory Leader example:

| Current me | Future me 1 | Future me 2 |
|---|---|---|
| **Police** | **Librarian** | **Librarian** |
| **Librarian** | **Police** | Ranger |
| Carpenter | Actor | **Police** |
| Actor | Carpenter | Carpenter |
| Ranger | Ranger | Actor |

| Current you | Future you 1 | Future you 2 |
|---|---|---|
| **Police** | **Librarian** | Actor |
| Carpenter | Actor | Carpenter |
| Ranger | **Police** | Ranger |
| Actor | Carpenter | **Librarian** |
| **Librarian** | Ranger | **Police** |

Now, notice that I always order librarian and police officer in the same way that you do. So Independence of Irrelevant Alternatives tells us that our aggregated preferences should rank

> those two careers in the same way. Thus, you should choose to become a librarian iff I should, and similarly for becoming a police officer.

How plausible is the Independence of Irrelevant Alternatives? Not, I think, very plausible. To see why, consider an aggregation method, known as the *Borda count*, that violates it.

> **Borda count method** Suppose $\mathcal{A}$ consists of $n$ different options, $a, \ldots, a_n$. And suppose $\langle \preceq_1, \ldots, \preceq_m \rangle$ is a sequence of preference orderings over $\mathcal{A}$. Then, in order to obtain the group preference ordering $\preceq_G$, we proceed as follows:
>
> - Take each act $a_i$ and each individual $j$ in turn.
> - Score $a_i$ relative to individual $j$ based on the position in the ordering $\preceq_j$ that $a_i$ occupies.
>
>   So, if $a_i$ is at the top of the ranking $\preceq_j$, it receives a score of $n$, if it is second in the ranking, it scores $n-1, \ldots,$ if it is bottom of the ranking, it scores 1.
> - Score an act $a_i$ by taking its (mean) average score relative to the individuals $j$, for $1 \leq j \leq m$.
> - Let $\preceq_G$ be the ordering of the acts by their scores.

Thus, in the Tory leadership examples above:

| Candidate | V1 | V2 | V3 | Borda | V1* | V2* | V3* | Borda* |
|---:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|
| Crabb | 1 | 1 | 4 | 2 | 3 | 3 | 3 | 3 |
| Fox | 3 | 2 | 2 | 2.333 | 4 | 2 | 4 | 3.333 |
| Gove | 2 | 3 | 1 | 2 | 2 | 4 | 5 | 3.666 |
| Leadsom | 5 | 4 | 3 | 4 | 5 | 3 | 1 | 3 |
| May | 4 | 5 | 5 | 4.666 | 1 | 5 | 2 | 2.666 |

Now, as we can see in the two examples above, the Borda count method violates the Independence of Irrelevant Alternatives. Relative to the first set of voters in the Tory leadership contest, May's Borda score is $\frac{4+5+5}{3} \approx 4.666$, while Leadsom's is $\frac{5+4+3}{3} = 4$. Relative to the second set, May scores $\frac{1+5+2}{3} \approx 2.666$, while Leadsom scores $\frac{5+3+1}{3} \approx 3$. Thus, the first group ranks May above Leadsom, while the second ranks Leadsom above May.

The Borda count is slightly more complicated in the second example concerning careers. In that example, as so often with cases of choosing for changing selves, we do not have a fixed set of selves to aggregate. Rather, I have two sets of possible selves, and you have two sets of possible selves: for me, the first contains my current self and my first possible future self, while the second contains my current self and my second possible future self; and similarly for yours. The natural thing to do to aggregate the attitudes of these various selves in my case is to take my expectation of the Borda count for my possible selves; and, again, similarly for you. The Borda count for being a police officer relative to the set containing my current self and my first possible future self is $\frac{5+4}{2} = 4.5$; for the other set, it is $\frac{5+3}{2} = 4$. If we assume that I am completely ignorant which of the two selves will come about, and thus think each is 50% likely, my expectation of the Borda count for being a police officer for my whole self is $0.5\frac{5+4}{2} + 0.5\frac{5+3}{2} = 4.25$. And for being a librarian it is $0.5\frac{4+5}{2} + 0.5\frac{4+5}{2} = 4.5$. So, in aggregate, I should prefer Librarian to Police Officer. In contrast, your expectation for the Borda count for a police officer is $0.5\frac{5+3}{2} + 0.5\frac{5+1}{2} = 3.5$, and for a librarian it is $0.5\frac{1+5}{2} + 0.5\frac{1+2}{2} = 2.25$. So, in aggregate, you should prefer Police Officer to Librarian. So I prefer being a librarian to being a police officer, while you prefer being a police officer to being a librarian, even though each of my selves ranks the two options exactly as your corresponding self does. So, expected Borda count violates the Independence of Irrelevant Alternatives just as the standard Borda count does.

Now, as a voting method, there are well-known and legitimate concerns about the Borda count method. In particular, if voters know that this is the method that will be used to aggregate their preferences to give the collective preferences, they can then engage in tactical voting: that is, they may have an incentive to present as their ranking an ordering that is different from their true ordering. For instance, recall the Tory leadership election from above. Suppose the first table gives the true preference ordering of Voters 1, 2, and 3. And suppose that Voter 1 knows the preference ordering of Voters 2 and 3, or is pretty confident of what they are. Then Voter 1 has an incentive to report the following preference ordering, which is not their true ordering: Leadsom, Fox, Gove, Crabb, May. If they do that, then Leadsom will receive a Borda score of $\frac{5+4+3}{3} = 4$, while May will receive $\frac{1+5+5}{3} \approx 3.666$. Thus, the group will rank Leadsom above May, and indeed will rank Leadsom above Fox, Gove, and Crabb as well. That is, Voter 1 has an incentive to vote tactically.

The worry about tactical voting does not seem to occur when we apply the method as part of a Aggregate Utility Solution to the problem of choosing for changing selves. There, the preference orderings we wish to aggregate are not the *reported* orderings of the various selves that I comprise; they are the *true* orderings of those selves. So there can be no tactical voting. Having said that, it may be possible to take steps to change the preference ordering of a particular self so that it contributes to the aggregate in a way that best serves that self's interests. For instance, by analogy with the instance of tactical voting in the Tory leadership example, in the careers example, my current self might take steps to change its ordering from Police, Librarian, Carpenter, Actor, Ranger to Police, Carpenter, Actor, Ranger, Librarian. After all, if my current self manages to pull off such a feat of preference-management, then the Borda count method will rank Police above everything else, which is just as that particular self wants it. By skilfully moving myself from less than full enthusiasm for the life of books to actively and vigorously disliking it, I secure my preferred career as a police officer, even though both of my possible future selves would prefer being a librarian. But such preference-management isn't obviously possible, and in any case doesn't give us a strong reason against Borda counting.

Thus, in our context, the usual arguments against Borda counting, and in favour of the Independence of Irrelevant Alternatives principle, do not apply. The insight of Borda counting is, roughly, this: How I rank options $c$, $d$, and $e$ relative to $a$ and $b$ gives a very rough indication of how strongly I prefer $a$ to $b$ or $b$ to $a$. If I rank $a \succ b \succ c \succ d \succ e$ then it seems that I prefer $a$ over $b$ less strongly than if I rank $a \succ c \succ d \succ e \succ b$. The Borda count reflects that and allows us to factor it in to our aggregation. Thus, the thought is that, while the group ordering of $a$ and $b$ really does just depend on the individual attitudes to $a$ and $b$, the individual orderings of $c$, $d$, and $e$ relative to $a$ and $b$ give information about the strength of the individual attitudes to $a$ and $b$—information that is left out if we look just at whether each individual prefers $a$ or $b$.

However, this leads us to a more fundamental problem with aggregating preference orderings directly, whether we use the Borda count method or something else. While the number of alternatives that you rank between $a$ and $b$ might give *some* indication of how much more you value $a$ and $b$, it is by no means a perfect guide. My current self might value being a police officer enormously, and a librarian, carpenter, actor, and ranger hardly at all, though in that order, while your current self might value all five careers enormously, with only tiny differences between them, but in the order police officer, carpenter, ranger, actor, librarian. Then the Borda count would say that your current self values being a police officer much more than being a librarian, while my current self only values it a little more. And that would be a mistake. Thus, the Borda count might be a good method to use when the *only* information you have at your disposal is the individuals' preference orderings, though even there it can go wrong. But when you have more information about the evaluation functions on which they are based—that is, when you have substantial cardinal information as well as ordinal information—it is not the best thing to do.

Here's another way to see this point: Suppose we have two sets of three voters with the following evaluation (or subjective expected utility) functions:

| | Voter 1 | Voter 2 | Voter 3 |
|---|---|---|---|
| *a* | 10 | 4 | 7 |
| *b* | 1 | 5 | 3 |

| | Voter 1* | Voter 2* | Voter 3* |
|---|---|---|---|
| *a* | 5 | 1 | 6 |
| *b* | 4 | 10 | 4 |

They have the same preference profile: first and third voters prefer *a* to *b*, while second voter prefers *b* to *a*. Thus, *any* aggregation method that pays attention only to the preference orderings of the individuals, and not to the evaluation functions on which they are based, must assign the same aggregate preference ordering to both groups. But it seems that we might well want their aggregate ordering to be different. For instance, if we just take straight averages of the expected utilities of the acts, the first group gives an average of 7 to *a* and an average of 3 to *b*, while the second group gives an average of 4 to *a* and 6 to *b*. That is, the average expected utilities order *a* and *b* differently in the two cases, even though any aggregation method that paid attention only to the preference orderings must, of necessity, order them in the same way in both cases.

The lesson is this: when we have the cardinal information from which the ordinal information in the preference ordering is extracted, and upon which it is based, we should use that cardinal information and aggregate that, rather than directly aggregating the preferences. And, on the realist interpretation of decision theory that we have adopted, we do have that information. So let us turn to that proposal now.

Before we move on, however, it is worth saying that the aggregation methods that we will consider in the coming two sections do satisfy their own versions of the Independence of Irrelevant Alternatives. Thus, the weighted average ex ante method that we consider in section 6.2 makes the group expected utility for a particular act depend only on the individual expected utilities for that act, and the weighted average ex post method that we consider in section 6.3 makes the group credences and the group utilities for a particular act (and thus the group expected utility for that act) depend only the individual credences and individual utilities for that act. Thus, we will preserve the spirit of the Independence of Irrelevant Alternatives, if not the letter. However, as we will see in chapter 9, this is not why we select those aggregation methods. Rather, we argue for these methods without appealing to this feature of them, and this feature is then just an attractive but unintended consequence.

## 6.2 Aggregating evaluation functions

So aggregating preference orderings won't work. Let's now consider how we might aggregate evaluation (or expected utility) functions instead. These are the attitudes that sit at the next level down from preference orderings. As noted above, this is sometimes called the *ex ante* method. Again, we start by considering such a method in the usual context of social choice theory, where we have a fixed collection of $n$ individuals whose judgments we wish to aggregate. Their credence functions are $P_1, \ldots, P_n$ and their utility functions $U_1, \ldots, U_n$. Each individual $i$ has a evaluation function $V_i$ that records her expected utility, so that $V_i(a) = \sum_{s \in \mathcal{S}} P_i(s||a)U_i(a \;\&\; s) = \sum_{s \in \mathcal{S}} P_i^a(s)U_i^a(s)$. And each individual $i$ also has a preference ordering $\preceq_i$ that is defined in the usual way, so that $a \preceq_i b$ iff $V_i(a) \preceq V_i(b)$.

On the particular version of the ex ante method I'll consider—a version that we might called the *weighted average ex ante method*—we take the group's aggregate evaluation function $V_G$ to be a weighted arithmetic average of the individuals' evaluation functions.[37] That is, we take a set of non-negative real numbers $\alpha_1, \ldots, \alpha_n$—one for each individual in the group—such that they sum to 1—that is, $\alpha + \ldots + \alpha_n = 1$. And we define $V_G$, the evaluation function of the group, as follows: for each act $a$ in $\mathcal{A}$,

$$V_G(a) = \sum_{i=1}^{n} \alpha_i V_i(a).$$

Having done that, we determine $\preceq_G$ in the usual way on the basis of $V_G$.

Note: in order for such an aggregation method to make sense, the utility functions of the various individuals in group $G$, which partially determine the evaluation (or expected utility) functions of those individuals, must all

[37] Given a set of non-negative real numbers $\alpha_1, \ldots, \alpha_n$ such that $\alpha + \ldots + \alpha_n = 1$, and a set of real numbers $r_1, \ldots, r_n$, the weighted arithmetic average of the $r_i$s by the $\alpha_i$s is $\alpha_1 r_1 + \ldots + \alpha_n r_n$.

measure value on the same scale. Indeed, it was precisely because Arrow thought it impossible to guarantee this that he believed that we should aggregate preference orderings, rather than expected utility functions. In Chapter 8, I will argue that it is in fact possible, at least in the case we are considering, when the individuals in the group are my past, present, and possible future selves, and group is me.

To evaluate this particular ex ante method it will help to compare it with the corresponding ex post method, so we set this out now. As we explained above, on an ex post method, we aggregate the individuals' credences, $P_1, \ldots, P_n$, to give the aggregate group credences, $P_G$, and we aggregate their utilities, $U_1, \ldots, U_n$, to give the aggregate group utilities, $U_G$, and then we use those aggregates to first determine the aggregate evaluation function $V_G$, and then to determine the preference ordering on the acts $\preceq_G$. On the weighted average version of the ex post method, the group's credence function, $P_G$, is a weighted sum of the individual credences: $P^a_G(-) = \sum^n_{i=1} \alpha_i P^a_i(-)$.[38] And the group's utility function, $U_G$, is a weighted sum of the individual utilities: $U^a_G(-) = \sum^n_{i=1} \beta_i U^a_i(-)$. We then determine $V_G$ in the usual way: $V_G(a) = \sum_{s \in \mathcal{S}} P^a_G(s) U^A_G(s)$. And we determine the preference ordering $\preceq_G$ on the basis of that.

Before we go any further, it is worth noting that the two methods we have just described—the weighted average ex ante method and the weighted average ex post method—are not compatible. That is, proceeding in one way often gives a result that cannot possibly be recovered by proceeding in the other way. Let's see this in an example:

[38] When we determine the credences of a group on the basis of the credences of its members in this way, it is known as *linear pooling*, and it is one of many different ways we might aggregate a number of different credence functions (Genest & Zidek, 1986; Dietrich & List, 2015; Pettigrew, taa). It has certain advantages and certain disadvantages. In its favour:

(i) it preserves probabilistic coherence, so that the linear pool of a group of coherent credence functions is itself guaranteed to be coherent;

(ii) the group's credence in a given proposition depends only on the individuals' credences in that proposition (Aczél & Wagner, 1980);

(iii) when all the individuals agree on a credence, the group agrees with them on that credence.

To its detriment:

(iv) it does not commute with conditionalization (Madansky, 1964);

(v) when all individuals take two propositions to be independent, the group usually does not take them to be independent.

We will consider these in greater detail in chapter 9, where we offer a direct argument in favour of linear pooling for credences and utilities.

**Date night** It is date night, and Isaak and Jeremy are deciding which restaurant they should book: Thai Garden (let's say that the act of booking this is $a$) or Silvio's (act $b$). There are two relevant states of the world: in the first, Silvio's is serving baked zitti and Thai Garden is serving green curry with chicken (state $s_1$); in the second, Silvio's is serving meatballs and Thai Garden is serving red curry with vegetables (state $s_2$). These two states partition the space of possibilities—that is, Isaak and Jeremy are both certain that one or other is true, but not both. They also take the states of the world to be independent of their choice of restaurant. Isaak is 10% confident in $s_1$ and 90% confident in $s_2$, while Jeremy is exactly the opposite—that is, he is 90% confident in $s_1$ and 10% confident in $s_2$. The utilities they assign to each situation—going to Silvio's when Silvio's serves zitti and Thai Garden serves chicken curry, going to Silvio's when Silvio's serves meatballs and Thai Garden serves vegetable curry, and so on—are given in the table below:

| | $a$ & $s_1$ | $a$ & $s_2$ | $b$ & $s_1$ | $b$ & $s_2$ |
|---|---|---|---|---|
| | Silvio's | Silvio's | Thai Garden | Thai Garden |
| | Zitti | Meatballs | Zitti | Meatballs |
| | Chicken | Veg | Chicken | Veg |
| Isaak | 3 | 2 | 10 | 1 |
| Jeremy | 4 | 1 | 3 | 2 |

Now, let's look first to the weighted average version of the ex post method. On this, we assign a weight, $\alpha$, to Isaak's credences and the rest, $1-\alpha$, to Jeremy's. And we assign a weight, $\beta$, to Isaak's utilities and the rest, $1-\beta$, to Jeremy's. Then their aggregate credence in $s_1$ is $0.1\alpha + 0.9(1-\alpha)$, while their aggregate credence in $s_2$ is $0.9\alpha + 0.1(1-\alpha)$. And their aggregate utility in $a$ & $s_1$ is $3\beta + 4(1-\beta)$; their aggregate utility in $a$ & $s_2$ is $2\beta + 1(1-\beta)$; and so on. And so, for instance, their aggregate evaluation for going to Silvio's (act $a$) is:

$$\begin{aligned}
&\text{ExPost}_{\alpha,\beta}(a)\\
&= \underbrace{[0.1\alpha + 0.9(1-\alpha)]}_{\text{Aggregate credence in } s_1} \times \underbrace{[3\beta + 4(1-\beta)]}_{\text{Aggregate utility in } a \text{ \& } s_1} + \\
&\qquad \underbrace{[0.9\alpha + 0.1(1-\alpha)]}_{\text{Aggregate credence in } s_2} \times \underbrace{[2\beta + 1(1-\beta)]}_{\text{Aggregate utility in } a \text{ \& } s_2}
\end{aligned}$$

Next, consider the weighted average version of the ex ante method. Here, we assign a weight, $\gamma$, to Isaak's evaluation function and the rest, $1-\gamma$, to Jeremy's. Thus, their aggregate evaluation for going to Silvio's (act $a$) is:

$$\begin{aligned}&\text{ExAnte}_{\gamma}(a) = \\ &\quad \gamma \underbrace{[0.1 \times 3 + 0.9 \times 2]}_{\text{I's evaluation of } a} + (1-\gamma) \underbrace{[0.9 \times 4 + 0.1 \times 1]}_{\text{J's evaluation of } a} = \\ &\qquad 2.1\gamma + 3.7(1-\gamma)\end{aligned}$$

And their aggregate evaluation for going to Thai Garden (act $b$) is:

$$\begin{aligned}&\text{ExAnte}_{\gamma}(b) = \\ &\quad \gamma \underbrace{[0.1 \times 10 + 0.9 \times 1]}_{\text{I's evaluation of } b} + (1-\gamma) \underbrace{[0.9 \times 3 + 0.1 \times 2]}_{\text{J's evaluation of } b} = \\ &\qquad 1.9\gamma + 2.9(1-\gamma)\end{aligned}$$

Now, notice: both Isaak and Jeremy evaluate $a$ as better than $b$—Isaak evaluates $a$ at 2.1 and $b$ at 1.9, while Jeremy evaluates $a$ at 3.7 and $b$ at 2.9. So every weighted average of their evaluations will rank $a$ above $b$. That is, the weighted average ex ante method is guaranteed to rank $a$ above $b$.

Now suppose we use the weighted average ex post method with $\alpha = \beta = 0.5$—that is, when we set the group credences and the group utilities, we weight Isaak's and Jeremy's credences and utilities equally. Then:

$$\begin{aligned}&\text{ExPost}_{\alpha,\beta}(a) = \\ &\quad [0.1\alpha + 0.9(1-\alpha)] \times [3\beta + 4(1-\beta)] + \\ &\qquad [0.9\alpha + 0.1(1-\alpha)] \times [2\beta + 1(1-\beta)] = \\ &\qquad\quad 0.5 \times 3.5 + 0.5 \times 1.5 = 2.5\end{aligned}$$

$$\begin{aligned}&\text{ExPost}_{\alpha,\beta}(b) = \\ &\quad [0.1\alpha + 0.9(1-\alpha)] \times [10\beta + 3(1-\beta)] + \\ &\qquad [0.9\alpha + 0.1(1-\alpha)] \times [1\beta + 2(1-\beta)] = \\ &\qquad\quad 0.5 \times 6.5 + 0.5 \times 1.5 = 4\end{aligned}$$

> So the instance of the ex post method that assigns equal weighting to Isaak and Jeremy ranks $b$ above $a$. Thus, the two methods are incompatible:
>
> - $\text{ExPost}_{0.5,0.5}(a) < \text{ExPost}_{0.5,0.5}(b)$
> - $\text{ExAnte}_{\gamma}(b) < \text{ExAnte}_{\gamma}(a)$, for all $0 \leq \gamma \leq 1$.

So the weighted average ex ante method and the weighted average ex post method are different and indeed incompatible. Given this, the question arises: which method should we adopt? A natural strategy is to seek an enumeration of the desirable features that one has while the other lacks. Fortunately, we don't have to look far. Recall the Weak Pareto condition from above: if all members of a group strictly prefer $b$ to $a$, then the group should prefer $b$ to $a$. Formally,

> **Weak Pareto** If $a \prec_i b$ for each individual $i$, then $a \prec_G b$.

Now, it is clear that the weighted average version of ex ante aggregation satisfies this—if each individual prefers $b$ to $a$, then each assigns higher expected utility to $b$ than to $a$, and thus any weighted average of their expected utilities for $b$ is greater than the corresponding weighted average of their utilities for $a$.[39] But one of the lessons of the Date Night example is that the weighted average ex post method does not always satisfy Weak Pareto. After all, in that example, Isaak and Jeremy both evaluate $a$ as better than $b$, and thus both prefer $a$ to $b$—that is, $b \prec_i a$, for each individual $i$. But if we assign them both equal weighting, the resulting ex post method ranks $b$ above $a$—that is, $a \prec_G b$.

Notice how the phenomenon arises in this example. Isaak and Jeremy agree that going to Silvio's is better than going to Thai Garden—that is, their expected utility for the former is greater than for the latter—but the reasons behind their judgments are different. Jeremy thinks that going to Silvio's or going to Thai Garden will be quite similar in their utility, whichever menu they are serving, but he puts more credence in the world at which he prefers Silvio's menu, and so he prefers that option. Isaak, by contrast, thinks that going to Thai Garden is much better than going to Silvio's in state $s_1$, where Thai Garden is serving chicken, but she's pretty confident that $s_1$ doesn't obtain; she's pretty confident that they're serving vegetable curry, and she gives that a much much lower utility; in the state in which she is much more

[39] In symbols: If $a \prec_i b$, for all $i$, then $V_i(a) < V_i(b)$, for all $i$, and thus, $\sum_{i=1}^{n} \gamma_i V_i(a) < \sum_{i=1}^{n} \gamma_i V_i(b)$, which gives $a \prec_G b$.

confident, namely, the state in which Thai Garden is serving vegetable curry and Silvio's is serving meatballs, Silvio's is slightly better for her. However, when we aggregate their credences, the aggregate is indifferent between the two states, $s_1$ and $s_2$, and when we aggregate the utilities, the aggregate thinks that going to Thai Garden is better than going to Silvio's in state $s_1$, and equally good in state $s_2$, and so it is to be preferred overall—in the jargon of decision theory, going to Thai Garden weakly dominates going to Silvio's, since it is at least as good in all states and better in some.

So the weighted average version of ex ante aggregation satisfies Arrow's Weak Pareto condition, while the corresponding version of ex post aggregation violates it. Surely this is a serious mark against ex post aggregation? After all, the so-called *unanimity preservation principles* are usually the most intuitively secure desiderata in judgment aggregation. As mentioned above, these are the principles that demand that, whenever all of the individuals to be aggregated agree on a particular judgment—that is, whenever they all make that judgment—the aggregate should agree with them too—that is, the aggregate should also make that judgment. Some putative unanimity preservation principles:

- **Indifference Preservation** If all individuals are indifferent between acts $a$ and $b$, then the group should be indifferent between $a$ and $b$.

  That is, if $a \sim_i b$, for all $i$, then $a \sim_G b$.

- **Strict Preference Preservation (or Weak Pareto)** If all individuals strictly prefer act $b$ to act $a$, then the group should strictly prefer $b$ to $a$.

  That is, if $a \prec_i b$, for all $i$, then $a \prec_G b$.

- **Independence Preservation** If all individuals have credences functions on which propositions $A$ and $B$ are probabilistically independent of one another, then, on the group credence function, $A$ and $B$ should be independent of one another.[40]

  That is, if $P_i(A \,\&\, B) = P_i(A)P_i(B)$, for all $i$, then $P_G(A \,\&\, B) = P_G(A)P_G(B)$.

[40] We say that $A$ and $B$ are probabilistically independent of one another relative to some probability function $P$ iff $P(A|B) = P(A)$, which is equivalent to $P(B|A) = P(B)$, and equivalent to $P(AB) = P(A)P(B)$. Thus, $A$ and $B$ are independent if conditioning on $B$ doesn't change the probability of $A$; or, equivalently, if conditioning on $A$ doesn't change the probability of $B$; or, equivalently, if the probability of the conjunction of $A$ and $B$ is the product of the probabilities of the conjuncts.

- **Credence Preservation** If all individuals have credence $r$ in proposition $A$, then the group credence in $A$ should be $r$.

  That is, if $P_i(A) = r$, for all $i$, then $P_G(A) = r$.

- **Equal Utility Preservation** If all individuals agree that two situations have equal utility, then the group should agree that those two situations have the same utility.

  That is, if $U_i(a \,\&\, s) = U_i(a' \,\&\, s')$ for all $i$, then $U_G(a \,\&\, s) = U_G(a' \,\&\, s')$.

But we have to be careful with these. In general, a unanimity preservation principle has the following form:

> **UPP** If $\Phi(i)$, for each $i$, then $\Phi(G)$,

where $\Phi$ is a property of an individual's judgments or attitudes—thus, $\Phi(i)$ could be $P_i(A \,\&\, B) = P_i(A)P_i(B)$, for instance, in which case the resulting unanimity preservation principle would be Independence Preservation from above. But we don't want to enforce this for every property $\Phi$. For instance, suppose I wish to aggregate the credences of Kacee and Lonnie in the proposition that the UK will leave in European Union. Kacee has credence $p$, and Lonnie has a different credence $q$. Now consider the instance of UPP that results from taking the following property:

$$\Psi(i) = (P_i(\mathit{Leave}) = p \text{ or } P_i(\mathit{Leave}) = q)$$

Then the antecedent of $\text{UPP}_\Psi$ is satisfied, since Kacee and Lonnie both have property $\Psi$. So, their aggregate must have that property. That is, $P_G(\mathit{Leave}) = p$ or $P_G(\mathit{Leave}) = q$. That is, the aggregate must agree with Kacee or with Lonnie; it cannot be any sort of compromise between the two. And this seems implausible. So we should not apply UPP indiscriminately, using just any property $\Phi$.

Thus, the question arises: for which properties $\Phi$ should we endorse the corresponding unanimity preservation principle? What properties of judgments should be preserved by aggregates when they are shared by all individuals in the group? Here is one sort of case where you might want to be careful: properties such that it's possible to share them with another individual but where such a shared property only reveals superficial agreement; that is, where is it possible for both individuals to have that property for very different reasons. An example:

> **Miners** Ten miners are trapped in a coal mine, which is filling with water. The miners are all in the same shaft: shaft A or shaft

> B. Kai and Leo are the first emergency responders on the scene. They have the means to block one or other shaft, but not both. If they block the shaft containing the miners, the flood will be stopped and all ten will survive. If they block the other shaft, the shaft will flood completely and all ten miners will die. If they block neither, the water will rise only far enough to kill exactly one miner, so that nine will survive. Kai is certain that the miners are in shaft A, and so prefers blocking that shaft over blocking neither, and prefers blocking neither over blocking shaft B. Leo, on the other hand, is certain they are in shaft B, and so prefers blocking that shaft over blocking neither, and prefers blocking neither over blocking shaft A.
>
> So, Kai and Leo share the following property: there is some shaft such that blocking that is preferred to blocking neither—that is, they share the disjunctive property *(Block neither* $\prec$ *Block A) or (Block neither* $\prec$ *Block B)*. However, this is not a property that we should demand of the aggregate. It seems perfectly legitimate to have an aggregate preference ordering on which the group prefers blocking neither shaft to blocking A and to blocking B. And, indeed, we obtain exactly such an ordering if, instead of aggregating the preference ordering in line with a unanimity preservation principle, we aggregate the evaluation functions or the credences and utilities using a weighted average ex ante or ex post method with equal weights for Kai and Leo: if we crudely take each life saved to add a utile, then the average expected utility of blocking neither is 9, while the average expected utility of blocking shaft A is 5 and similarly for shaft B; likewise, the expected utility of blocking neither from the point of view of the aggregate credences is 9, while he expected utility of blocking shaft A is 5, and similarly for shaft B.

What's going on here? A natural reading: while Kai's and Leo's preference orderings agree on the property in question, that agreement is only superficial. Kai and Leo agree that blocking neither shaft is not the best plan, but they agree on this for dramatically different reasons—Kai is certain that blocking A is better and that blocking B is catastrophically bad; Leo is certain that blocking B is better and that blocking A is catastrophically bad. This explains why we should not preserve this property in the aggregate ordering. Instead, we should look to the underlying reasons that determine their preference ordering. That is, we should look to their credences and utilities and the resulting expected utilities. We should aggregate those, the underlying reasons, rather than the preference ordering they determine. The problem for unanimity preservation principles is that, when individuals agree on judgments at one level but as a result of disagreement on their lower-level reasons for making these judgments, we should aggregate the lower-level reasons first and then use those to produce the higher-level aggregate judgment. Another example to illustrate our point:

> **In the Archives** Suppose two historians, Khalid and Lana, are researching the same question, but in two different archives. Both know that there may be a pair of documents, one in each archive, whose joint existence would establish a controversial theory beyond doubt. Khalid finds the relevant document in his archive, but doesn't know whether Lana has found hers; and Lana finds the relevant document in her archive, but doesn't know whether Khalid has found his. Indeed, each assigns a very low credence to the other finding their document; as a result, both have a very low credence in the controversial theory.
>
> Now, suppose we wish to aggregate Khalid's and Lana's doxastic attitudes. Since both assign a very low credence to the controversial theory, something like the weighted average approach will then say that their joint credence in that theory is very low. But this seems wrong. Together, their evidence establishes the controversial theory beyond doubt. So surely this should be reflected in their joint credences.

What is going on in this example? Your doxastic state is not exhausted by your credences. It also contains your evidence, which gives at least part of your reason for having the credences you do have. Thus, while Khalid and Lana agree that the controversial theory is very improbable, they agree on that for very different reasons. Khalid thinks it improbable because he knows that he has found his document, but thinks it's unlikely that Lana found hers. Lana thinks it improbable because she found her document, but think it's unlikely that Khalid found his. If, instead of aggregating the higher-level aspects of their doxastic state—that is, their credences—we aggregate the lower-level aspects—namely, their evidence—we include the discovery of both documents in their joint evidence, and therefore a high credence in the controversial theory, as required.

One final example to support our claim:

> **Badminton** Suppose you and I share the same evidence, and we

> agree that it is 60% likely that Ji Hyun Sung or Carolina Marin will win the badminton tournament ($X_1 \vee X_2$). But I think that because I think it is 20% likely that Sung will win ($X_1$) and 40% likely that Marin will ($X_2$), while you think it 50% likely that Sung will win ($X_1$) and 10% likely that Marin will ($X_2$). We are both agreed that it is 40% likely that neither will ($X_3 = \neg(X_1 \vee X_2)$). The question is this: should the aggregate of our credences agree that Sung or Marin will win ($X_1 \vee X_2$)?

There are many popular aggregation procedures that answer 'no'. For instance, one such procedure is so-called *geometric pooling*. Just as linear pooling takes the aggregate of two credence functions to be their weighted *arithmetic* average, geometric pooling takes it to be their weighted *geometric* average (which we then normalise).[41] Thus, if we give a weight of 0.5 to each of us, then geometric pooling gives the following joint credences in $X_1$, $X_2$, and $X_3$:

- $P_G(X_1) = \frac{\sqrt{0.2}\sqrt{0.5}}{\sqrt{0.2}\sqrt{0.5}+\sqrt{0.4}\sqrt{0.1}+\sqrt{0.4}\sqrt{0.4}} \approx 0.345$
- $P_G(X_2) = \frac{\sqrt{0.4}\sqrt{0.1}}{\sqrt{0.2}\sqrt{0.5}+\sqrt{0.4}\sqrt{0.1}+\sqrt{0.4}\sqrt{0.4}} \approx 0.218$
- $P_G(X_3) = \frac{\sqrt{0.4}\sqrt{0.4}}{\sqrt{0.2}\sqrt{0.5}+\sqrt{0.4}\sqrt{0.1}+\sqrt{0.4}\sqrt{0.4}} \approx 0.437$

We then determine the aggregate credence in $X_1 \vee X_2$ by summing the aggregate in $X_1$ and the aggregate in $X_2$. So:

$$P_G(X_1 \vee X_2) = P_G(X_1) + P_G(X_2) \approx 0.563 \neq 0.6$$

Thus, geometric pooling violates the unanimity preservation principle called Credence Preservation from above. Of course, you may take this to be a strike against geometric pooling—but, given the success of that method of aggregation, and the other arguments in its favour, you may very well also take it as a strike against certain unanimity preservation principles, such as Credence Preservation from above.[42] Again, the point is that, when we produce our aggregate credences in the propositions $X_1$, $X_2$, $X_3$, $X_1 \vee X_2$, we

[41] Suppose $0 \leq \alpha, \ldots, \alpha_n \leq 1$ is a set of weights that sum to 1. Then the weighted geometric average of a set of non-negative real numbers $0 \leq r_1, \ldots, r_n$ with weights $\alpha, \ldots, \alpha_n$ is $r_1^{\alpha} \times \ldots \times r_n^{\alpha_n}$.

[42] For a discussion of this problem with geometric pooling, as well as an exploration of the possibly ways of avoiding it, see (Pettigrew, 2017, Section 9). For arguments that highlight the desirable features of geometric pooling, see (Russell et al., 2015).

aggregate the lower-level credences—the credences in the more fine-grained propositions $X_1$, $X_2$, $X_3$—as opposed to the higher-level credences—the credences in the more coarse-grained proposition $X_1 \vee X_2$. This is because we take the credences in the more fine-grained propositions to provide the underlying reasons for the credences in the more coarse-grained propositions.

Let us return now to Weak Pareto, the unanimity preservation principle that the weighted average version of ex ante satisfies, but which the corresponding version of ex post violates. When we require that our aggregation rule satisfies Weak Pareto, we ignore the possibility that agreement on the preference ordering of two options may mask deeper disagreement on the reasons behind those preferences, by analogy with our examples, Miners, In the Archives, and Badminton. In such a case, the preference ordering occurs at the higher level, while the credences and utilities occur at the lower level. When we see that all individuals agree that one act is better than another, we do not yet know whether this is an agreement that we should preserve in the aggregate preference ordering or whether it is a merely superficial agreement that the individuals have come to for very different reasons.

Of course, constructivists will not be moved by this. For them, credences and utilities are merely shadows cast by the real thing, which is the preference ordering. So, for them, there is no sense in which an individual's credences and utilities give the reasons behind their preference ordering—the credences and utilities are simply useful mathematical tools for representing the preference ordering; they have no reality beyond this. But, as I have said, I adopt a realist line here. And it seems to me that the constructivist's failure to explain what goes wrong with the Weak Pareto principle in these cases reveals something of the bizarre behaviourism behind their view.

In sum: I don't think we should be tempted by ex ante aggregation. In general, when you have access to two levels of judgment and the lower gives your reasons for the higher, then you should first aggregate the judgments at the lower level—that is, your reasons—and then use those aggregates to determine the aggregate judgment at the higher level.

There is, in fact, another argument against ex ante aggregation, but it needn't detain us too long. It is based on a theorem by Philippe Mongin (1995), which says that, for many sets of preference orderings $\preceq_1, \ldots, \preceq_n$ over a set of acts, any aggregate preference ordering $\preceq_G$ that satisfies the Weak Pareto principle with respect to them cannot itself be represented as having been generated by expected utilities; that is, there is no credence function $P_G$ and utility function $U_G$ such that $a \preceq_G b$ iff $V_G(a) \leq V_G(b)$, where $V_G(a) = \sum_{s \in \mathcal{S}} P_G(s||a)U_G(a \,\&\, s)$.

Now, you might wonder whether this is really a mark against the ex ante approach. After all, while some argue that groups can be thought of as individuals in their own right, and thus should have preferences that are representable as having been generated by a credence function and utility function in the usual way, no-one thinks that we *must* think of every aggregation process in this way.[43] That is, there's no obligation to consider your group as an individual, and so it doesn't seem a very strong reason for dismissing the ex ante method that it is impossible to represent the group preference ordering using credences and utilities in the way you'd expect from a group individual. However, this misses the point. The concern is not so much that $\preceq_G$ is not representable, but rather that, by being unrepresentable, $\preceq_G$ must therefore violate one of the standard axioms for preference orderings—the Savage, or Jeffrey, or Joyce axioms, for instance (Savage, 1954; Jeffrey, 1983; Joyce, 1999). These axioms are thought to lay down necessary rationality conditions on a preference ordering regardless of whether that preference ordering is taken to be held by an individual or not—they are simply coherence constraints on preferences.

## 6.3 Aggregating credences and utilities

Our discussion above of the problems that arise when we aggregate attitudes at higher levels rather than lower levels tells against the ex ante method. Better to aggregate individuals' reasons for their higher-level attitudes rather than to aggregate the judgments themselves. And that points to ex post aggregation.

However, as Matthias Hild (2001) has shown, ex post aggregation is not without its problems. It can give rise to what Hild calls *unstable preferences*. The idea is this: suppose we must decide between two options, *a* and *b*. Our standard approach is this: we set up a decision problem in which we represent *a* and *b* as acts. As we have seen above, a decision problem includes a set of states of the world, and a set of possible acts. Each individual is then equipped with a credence function over pairs of acts and states, and a utility function over the same pairs. How should we specify the states of the world? For instance, suppose I am trying to decide whether or not to take an umbrella when I go outside. I start to construct my decision problem: I specify the set of acts so that it includes *Umbrella* and *No Umbrella*. Then I turn to the states of the world. At what level of grain should I specify these? Should I simply divide the possibilities into two, *Rain* and *No Rain* ? Or into three, *Heavy Rain*, *Light Rain*, and *No Rain*? Or into more, *1mm of*

[43] See (Pettit & List, 2011; Tollesfen, 2015).

*Rain*, *2mm of Rain*, …, *10mm of Rain*, and *No Rain*? There is an enormous range of possible levels. It is important that the evaluation that I give for a particular act, and my preference ordering over the acts, do not depend on the level of grain at which I choose to specify the states of the world in my decision problem. If they did so depend, my preferences would be unstable, in Hild's sense. Without a privileged level of grain, the decision theory would be rendered useless: relative to the decision problem in which the worlds are specified at one grain, I'd prefer *a* to *b*; specified at a different grain, I'd prefer *b* to *a*; and there would be no way to tell which I should follow.[44]

As we noted in Chapter 2 (Footnote 14), many decision theories ensure that no such instability can arise by demanding that an individual's utility at a coarser level of grain is just the expectation of their utility at the finer level of grain (Jeffrey, 1983; Joyce, 1999). Thus, for instance, if we write *b* for the act of taking the umbrella, we might demand:

$$U^b(Rain) = P^b(Heavy\ Rain|Rain)U(Heavy\ Rain) + \\ P^b(Light\ Rain|Rain)U^b(Light\ Rain)$$

where $P^b(-) = P(-||b)$ and $U^b(-) = U(- \ \&\ b)$, as usual. And, in general:[45]

> **Inter-Grain Coherence** If $s_1, \ldots, s_n$ is a fine-graining of the state $s$, and $a$ is an act, then
>
> $$U^a(s) = \sum_{i=1}^{n} P^a(s_i|s)U^a(s_i)$$

It is then straightforward to show that the value of an act will be the same whether it is calculated relative to one level of grain or another. In the jargon of decision theory, this says that our decision theory is *partition invariant*—its recommendations in any decision problem do not vary with the level of grain of the partition of the ways the world might be that is used to specify the states; they are not sensitive to the level of grain.

However, as Hild shows, if we aggregate using the weighted average ex post method, the following is possible: we have two individuals with credences and utilities defined on two levels of grain such that

(i) Individually, their credences and utilities over the two levels of grain satisfy Inter-Grain Coherence.

[44] This is sometimes known as the *problem of partition sensitivity* in decision theory.

[45] This is our version of Joyce's formula from (Joyce, 1999, 178).

(ii) Collectively, relative to the more coarse-grained level, they prefer $a$ to $b$.

(iii) Collectively, relative to the more fine-grained level, they prefer $b$ to $a$.

Indeed, Hild provides an infinite descending change of levels of grain, together with credences and utilities at all of them for a pair of individuals, such that both individuals satisfy Inter-Grain Coherence and such that the group preference between $a$ and $b$ generated by the weighted average ex post method switches back and forth at each new level: at the first level, $a$ is preferred to $b$; at the second, $b$ is preferred to $a$; at the third, $a$ is preferred to $b$ again; and so on. I won't describe that entire hierarchy, but I will lay out the first two levels and illustrate them in an example:

> **Cinema** Maura and Noni are trying to decide whether to go to see a film or just stay at home. They both agree that the utility of staying at home is 0. There's only one film showing at their local cinema. It's called *Washington*, and they don't know for sure whether it's a biopic of the president or a modern political drama set in the city. If it's a biopic, Maura will enjoy it, giving it a utility of 3, while Noni will hate it, giving it a utility of -5. If it's a modern political drama, their utilities will be reversed—Maura will give it -5 and Noni will give it 3. Maura is pretty confident that it is a biopic (75% confident, to be precise), and Noni is pretty confident that it is a modern political drama (75% confident, to be precise).
>
> There are two levels of grain: On the first, there is just one state of the world—the film showing at the cinema is called *Washington*—and both are certain of this. On the second, there are two states of the world—the film showing at the cinema is called *Washington* and it is a biopic; the film showing at the cinema is called *Washington* and it is a modern political drama. Since we assume Inter-Grain Coherence, we have:
>
> - $U_M(\textit{Watch}\ \text{Washington}) =$ $0.75 \times U_M(\textit{Watch biopic}) + 0.25 \times U_M(\textit{Watch drama}) =$ $\frac{9}{4} - \frac{5}{4} = 1$
> - $U_N(\textit{Watch}\ \text{Washington}) =$ $0.25 \times U_N(\textit{Watch biopic}) + 0.75 \times U_N(\textit{Watch drama}) =$ $-\frac{5}{4} + \frac{9}{4} = 1$

> Thus, Maura and Noni agree on the utility of going to the cinema. It is 1, and thus at this coarse-grained level, going to the cinema gets higher utility than staying home, which receives 0 for sure.
>
> However, now look to the more fine-grained level. There, Maura and Noni disagree on both utilities and credences. Let's suppose that we aggregate the utilities by taking a straight average. Then the group utility for *Watch biopic* will be $\frac{3-5}{2} = -1$, and the group utility for *Watch drama* will be $\frac{-5+3}{2} = -1$. Thus, however the world turns out, the group assigns a lower utility to going to the cinema than to staying home.
>
> Thus, at the coarse-grained level, our group prefers going to the cinema to staying at home; but at the fine-grained level, it prefers staying at home to going to the cinema.

And note that the example of Date Night from the previous section will furnish another example. And indeed any case in which weighted average ex post aggregation is incompatible with weighted average ex ante aggregation will furnish another example. After all, weighted average ex ante aggregation is equivalent to weighted average ex post aggregation at a higher level (given Inter-Grain Coherence). Thus, in the Date Night case, instead of comparing ex post aggregates at the level that specifies the four restaurant-and-menu combinations with ex ante aggregates at that same level and noting that they disagree on the order of the options, we compare ex post aggregates at that level with ex post aggregates at the level that specifies only the two possible restaurants and we note that they disagree on the order of the options.

Thus, the weighted average ex post method fails to adhere to what we might called the Independence of Grain condition:

> **Independence of Grain** Suppose we have two decision problems. They share the same set of acts, and the states of one are a fine-graining of the states of the other. And suppose that the credences and utilities at the different levels of grain are related by Inter-Grain Coherence. Then an aggregation method should give the same result for both problems.

In the terminology we introduced above, Independence of Grain is a positive dependence condition. Just as Independence of Irrelevant Alternatives says that the group preferences over *a* and *b* should depend only on the individual preferences over *a* and *b* and should not, for instance, depend on the individual preferences over $b$ and $c$, so Inter-Grain Coherence says that the group preferences over the acts in $\mathcal{A}$ should depend only on the credences and utilities in the states of the world and the acts, not on the level of grain at which the states of the world are described.

In this chapter, then, we have considered methods of judgment aggregation that work at the three different levels represented in our decision problems: the preference orderings, the evaluation (or expected utility) functions, and the credences and utilities. As we saw, all face problems. If we aggregate preference orderings, we run up against Arrow's Impossibility Theorem, of course. But, more importantly for realists, we are forced to aggregate in the same way preference orderings that are generated by very different credences and utilities. We face a similar problem if we aggregate evaluation (or expected utility) functions. Again, we end up sometimes preserving agreement that is only superficial. These problems suggest that we should aggregate credences and utilities first, as the underlying reasons behind the expected utilities and preference orderings. However, as we saw in the final section, this can give rise to unstable preferences that violate the Independence of Grain. In the next chapter, we will endorse one of these aggregation methods in the context of choosing for changing selves, and we will say how we intend to deal with the problem that it faces.

# Chapter 7

# The Aggregate Utility Solution II: the solution itself

The version of the Aggregate Utility Solution that I favour is, in the language of the previous chapter, an ex post method. We aggregate the credences of my past, present, and possible future selves; and we aggregate the utilities of my past, present, and possible future selves; and then we combine these in the usual way to give the aggregate evaluation function and thus the aggregate preference ordering that I will use to make my decision. Indeed, not only is my version of this solution an ex post method, it is a weighted average ex post method. As a result, we might worry that it will fall foul of Hild's objection; we might think that our favoured method will violate Independence of Grain in the same way—and indeed for the same reasons—that it violates Weak Pareto. And, as we will see, it does. But I will argue that we can solve this problem.

## 7.1 The framework

To begin, we must look at how we specify the states of the world in our decision problem. In our overview of expected utility theory, we mentioned the set of states of the world, but we didn't say much about what these specify. How fine-grained are they? What information do they supply? Are they Lewisian possible worlds, specifying a truth value for every proposition—whether Cleopatra was right-handed or left-handed, whether Shakespeare liked apples better than oranges or oranges better than apples—or are they something more coarse-grained than that? If more coarse-grained, how coarse-grained can we permit? Now that we are considering cases in which our utilities might change over time, we must ensure that our states are grained finely enough that they specify such changes. Thus, each state must specify not only your current utilities, but also your past and future utilities. Also, while it will not in fact play a role in our decision theory, it will be useful to insist that each state specifies not only your utilities at each time, but also your credences at each time.

Given a state $s$ in $\mathcal{S}$:

- $w^s$ is the possible world that is actual in state $s$.

  A possible world is a way that the world might be beyond you. Thus, in the example of Aneri, there is a possible world at which she is a police officer and she has to follow such-and-such a protocol, there is a possible world at which she is a police officer and she is tasked with enforcing a law that she thinks is immoral, there is a possible world at which she is a conservation officer and she has to shut down a particular wildlife reserve, there is a possible world at which she is a conservation officer and she is producing a report on the biodiversity in a given area, and so on.

  Let $\mathcal{W}$ be the set of possible worlds—that is, $\mathcal{W} = \{w^s : s \in \mathcal{S}\}$.

- $P_{s,i}$ is your credence function at time $t_i$ in state $s$.

  $P_{s,i}$ is defined on pairs of acts from $\mathcal{A}$ and possible worlds from $\mathcal{W}$, so that $P^a_{s,i}(w)$ is your credence function at $t_i$ in $s$ that world $w$ is actual on the supposition that act $a$ is performed.

- $U_{s,i}$ is your utility function at time $t_i$ in state $s$.

  $U_{s,i}$ is defined on pairs of acts from $\mathcal{A}$ and possible worlds from $\mathcal{W}$, so that $U^a_{s,i}(w)$ is your utility at $t_i$ in $s$ for the outcome in which $w$ is actual and $a$ is performed.

- These three components—the possible world and the sequences of credence and utility functions belonging to your successive selves—determine the state. Thus, we might represent a state of the world $s$ as follows:

  $$s = \langle w^s, U_{s,1}, \ldots, U_{s,n}, P_{s,1}, \ldots, P_{s,n} \rangle$$

  Of course, we know from Chapter 2 that it doesn't make a lot of sense to say that $U_{s,i}$ is your utility function without specifying a scale on which it measures value. After all, your conative state at a particular time is equally well represented by one utility function as by another that is a positive linear transformation of it. Thus, what we really mean when we say that $U_{s,i}$ is you utility function at $t_i$ in $s$ is that we've picked a particular scale on which to measure our your utilities at all times and in all states and $U_{s,i}$ gives your utilities measured on that scale. Now, as we have noted before, there are potential problems here, since it is often claimed that we cannot compare the scales on which the utilities of different individuals are measured, and it has been argued that the same is true even for different selves of the same individual (Briggs, 2015). We'll address that concern in Chapter 8.

- $V_{s,i}$ is your evaluation function at time $t_i$ in state $s$.

  $V_{s,i}$ is determined from $P_{s,i}$ and $U_{s,i}$ as follows:

$$V_{s,i}(a) := \sum_{w \in \mathcal{W}} P_{s,i}(w||a) U_{s,i}(a \,\&\, w) = \sum_{w \in \mathcal{W}} P^a_{s,i}(w) U^a_{s,i}(w)$$

- $\preceq_{s,i}$ is determined from $V_{s,i}$ in the usual way:

$$a \preceq_{s,i} b \quad \text{iff} \quad V_{s,i}(a) \leq V_{s,i}(b)$$

## 7.2 The decision rule

This, then, furnishes us with the possible states that we include in the decision problem; and it specifies the credences, utilities, values, and preferences that belong to your various selves within these states. How, then, do we propose to aggregate these attitudes to give your overall attitude at a particular time—that is, the attitude that you will use to make her decisions?

The first thing to note is that, as we argued at length in Chapter 5 in the context of the One True Utility Solution, we do not require a special method for aggregating our past, present, and possible future credences. If our current credences satisfy certain standard principles of rationality, such as Probabilism and the Principal Principle, they will incorporate their past and future credences in a satisfactory way. We do not need a further principle, like the Reflection Principle. In the cases in which I must satisfy that principle, it follows from more basic principles. Thus, if $P_p$ is your current credence function, then you should use $P_p$ to make your decisions.

In contrast, there is nothing analogous for utilities. At least on the subjectivist view, it is not sufficient for me to incorporate my information about my past and possible future subjective utilities by updating my credences on the evidence I have about them. This was the point at the end of Chapter 5. If I am making a decision on behalf of a group, no member of the group has a complaint against me if I use my own credences updated on my information about theirs—I have used my best attempt to achieve the shared goal of credences, namely, getting at the truth about the world. But a member of the group would have a complaint against me if I were to use my own utilities.Thus, unlike our past, present, and possible future credences, we do need to aggregate our past, present, and possible future utilities. And, indeed, as in the weighted average ex post approach from above, we will take our aggregate of them to be their weighted average. Thus, suppose $s$ is a state in $\mathcal{S}$. We then take a set of weights $0 \leq \alpha_{s,1}, \ldots, \alpha_{s,n} \leq 1$ that sum to 1 and we let:

$$U^a_G(s) = \sum_{i=1}^{n} \alpha_{s,i} U^a_{s,i}(w^s)$$

We define $V_G$ as usual: So:

$$\begin{aligned} V_G(a) &= \sum_{s \in \mathcal{S}} P^a_G(s) U^a_G(s) \\ &= \sum_{s \in \mathcal{S}} P^a_p(s) \sum_{i=1}^{n} \alpha_{s,i} U^a_{s,i}(w^s) \end{aligned}$$

And we define $\preceq_G$ in terms of $V_G$ as before.

Thus, when we choose, we ought to maximise $V_G$. That is, we ought to maximise the subjective expected utility from the point of view of the aggregate of our various selves, where our aggregate credences are given by our current credences, $P_p(-)$, and our aggregate utilities for a given state are given by a weighted average of our past, present, and future utilities within that state, $\alpha_{s,i} U_{s,i}(-)$.

## 7.3 The instability of preferences

Let us turn now to see how Hild's instability objection plays out in our context and for my favoured version of the Aggregate Utility Solution to the problem of choosing for changing selves. We illustrate it first with an example that is structurally similar to the Cinema example of Maura and Noni from above. It's a fairly involved example and it will take a while to lay it out and work through its features. For those who wish to skip forward, the upshot is that my favoured solution is vulnerable to Hild's objection just as standard weighted average ex post aggregation methods are.

> **Introvert or extrovert?** I am deciding whether or not to adopt a child called Sophie. At one level of description, the outcomes are

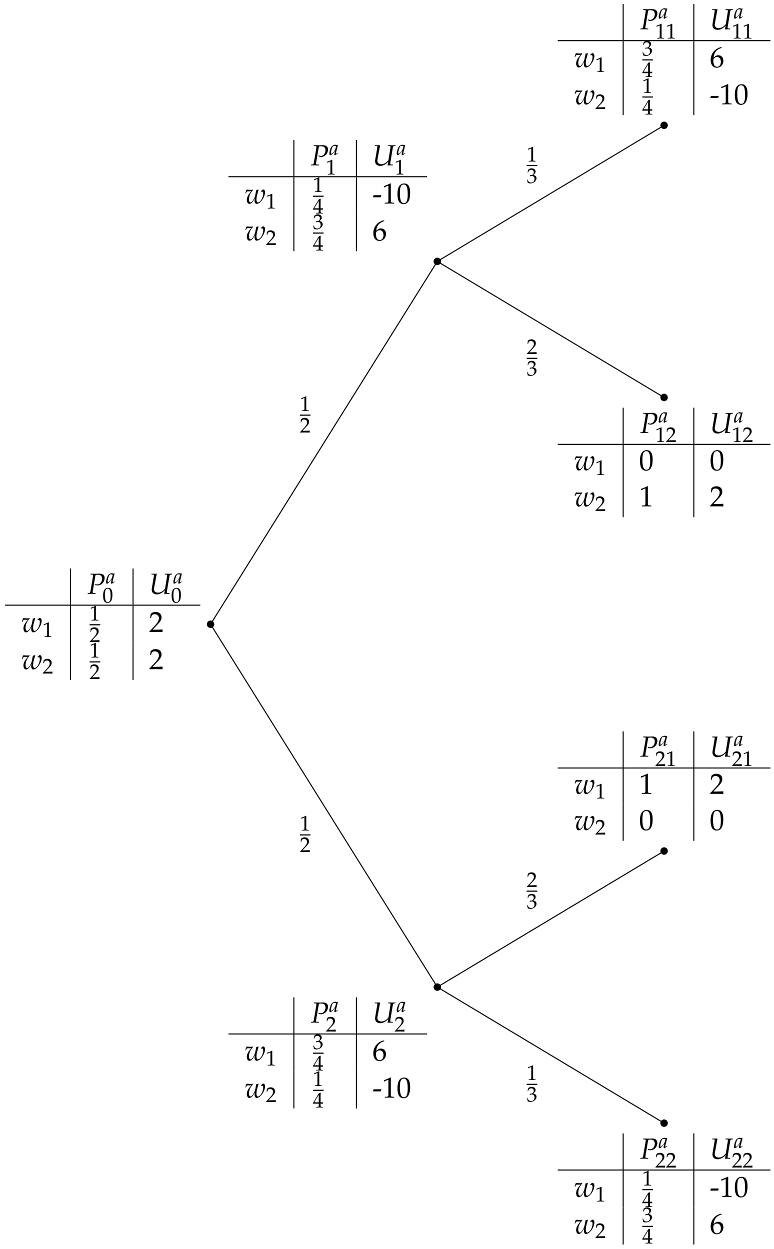
simple: if I adopt ($a$), I become a parent to Sophie ($W_1$); if I don't adopt ($b$), I don't ($W_2$). At a very slightly more detailed level of description, however, the outcomes are slightly less simple: if I adopt ($a$), either I become a parent to Sophie and she turns out to be an introvert ($w_1$), or I become a parent to Sophie and she turns out to be an extrovert ($w_2$); if I don't adopt ($b$), I don't become a parent to Sophie ($W_2$).

We will consider three times in these worlds. The present time (time $t_0$), a time five years in the future ($t_1$), and a time ten years in the future ($t_2$). As we will see, if I adopt, my credences and utilities will change between $t_0$ and $t_1$ and again between $t_1$ and $t_2$, as I learn more about my daughter. At each point, I'll learn one of two pieces of evidence. We'll write $P_0$ and $U_0$ for my initial credence and utility functions, $P_1$ and $U_1$ for my functions at $t_1$ should I learn the first piece of evidence and $P_2$ and $U_2$ should I learn the second, $P_{11}$ and $U_{11}$ for my functions at $t_2$ should I learned the first piece of evidence at $t_1$ and then the first piece at $t_2$, and so on. There are thus eight states of the world: the state in which Sophie is an introvert ($w_1$), I learn the first piece of evidence at $t_1$ (and thus have credence function $P_1$ and utility function $U_1$ at that time), and I learn the first piece of evidence at $t_2$ (and thus have credence function $P_{11}$ and utility function $U_{11}$ at that time); and so on. I will begin by setting out these different states in the form of a series of branching paths in Figure 7.1, and then I will walk us through it.

Currently:

- I think it's 50% likely that Sophie is an introvert and 50% likely that she's an extrovert (i.e. $P^a_0(w_1) = 0.5 = P^a_0(w_2)$).
- I have utility 2 for becoming Sophie's parent, whether she is an introvert or an extrovert (i.e. $U^a_0(w_1) = 2 = U^a_0(w_2)$).
- Therefore, by Inter-Grain Coherence, I have utility 2 for becoming Sophie's parent (i.e. $U^a_0(W_1) = 2$).
- I have utility 1 for not becoming Sophie's parent (i.e. $U^b_0(W_2) = 1$)

In five years' time, however, if I do adopt Sophie, I will have gained some evidence about whether she is an introvert or an extrovert. Let's say that I will either come to think it's 25% likely



|  | $P^a_0$ | $U^a_0$ |
|---|---|---|
| $w_1$ | $\frac{1}{2}$ | 2 |
| $w_2$ | $\frac{1}{2}$ | 2 |

|  | $P^a_1$ | $U^a_1$ |
|---|---|---|
| $w_1$ | $\frac{1}{4}$ | -10 |
| $w_2$ | $\frac{3}{4}$ | 6 |

|  | $P^a_{11}$ | $U^a_{11}$ |
|---|---|---|
| $w_1$ | $\frac{3}{4}$ | 6 |
| $w_2$ | $\frac{1}{4}$ | -10 |

|  | $P^a_{12}$ | $U^a_{12}$ |
|---|---|---|
| $w_1$ | 0 | 0 |
| $w_2$ | 1 | 2 |

|  | $P^a_2$ | $U^a_2$ |
|---|---|---|
| $w_1$ | $\frac{3}{4}$ | 6 |
| $w_2$ | $\frac{1}{4}$ | -10 |

|  | $P^a_{21}$ | $U^a_{21}$ |
|---|---|---|
| $w_1$ | 1 | 2 |
| $w_2$ | 0 | 0 |

|  | $P^a_{22}$ | $U^a_{22}$ |
|---|---|---|
| $w_1$ | $\frac{1}{4}$ | -10 |
| $w_2$ | $\frac{3}{4}$ | 6 |

Figure 7.1: The probabilities on the edges are my credences at the earlier time for the state at the later time.

that she's an introvert and 75% likely that she's an extrovert ($P_1$), or 75% likely that she's an introvert and 25% likely that she's an extrovert ($P_2$). And let's suppose that, currently:

- I think it's 50% likely that my credences will evolve in the first way and 50% likely that they'll evolve in the second way (i.e. $P_0(\textit{cred at } t_1 \textit{ is } P_1) = 0.5$ and $P_0(\textit{cred at } t_1 \textit{ is } P_2) = 0.5$).

So my current credences (given $P$) are my expectations of my credences at the five year point (given by $P_1$ and $P_2$)—that is, they satisfy the Reflection Principle.

What's more, as I become more confident that she's an introvert, I'll come to value being the parent of an introvert more, and if I become more confident that she's an extrovert, I'll come to value that more. So, if in five years' time, at $t_1$, my credence that she's an extrovert goes to 75% (i.e. $P_1$), then:

- my utility for being Sophie's parent and her being an introvert ($w_1$) will increase to 6 (i.e. $U^a_1(w_1) = -10$);
- my utility for being Sophie's parent and her being an extrovert ($w_2$) will decrease to $-10$ (i.e. $U^a_1(w_2) = 6$).

And, mutatis mutandis, if my credence that she's an introvert goes to 75% (i.e. $P_2$), then:

- my utility for being Sophie's parent and her being an introvert ($w_1$) will decrease to -10 (i.e. $U^a_2(w_1) = 6$);
- my utility for being Sophie's parent and her being an extrovert ($w_2$) will increase to 6 (i.e. $U^a_2(w_2) = -10$).

And therefore, by Inter-Grain Coherence, at the five year point:

- my utility for being Sophie's parent will be $U^a_i(W_1) = 2$, for each $i = 1, 2$.[46]

[46] After all:

$$\begin{aligned} U^a_1(W_1) &= P^a_1(w_1)U^a_1(w_1) + P^a_1(w_2)U^a_1(w_2) \\ &= \left(\frac{1}{4} \times -10\right) + \left(\frac{3}{4} \times 6\right) = 2 \end{aligned}$$

and similarly for $U^a_2(W_1)$.

Furthermore:

- my utility for not becoming Sophie's parent remains unchanged at 1 (i.e. $U_i^b(W_2) = 1$).

Finally, let's suppose that, in ten years' time, at $t_2$, if I adopt Sophie, I will have gained yet more evidence about whether she is an introvert or an extrovert. If, after five years, I received evidence that moved me to credence function $P_1$, then at ten years either I will think it's 75% likely that she's an introvert and 25% likely that she's an extrovert ($P_{11}$), or I will think it's certain (i.e. 100% likely) that she's an extrovert ($P_{12}$). On the other hand, if, after five years, I received evidence that moved me to credence function $P_2$, then at ten years either I will think it's 25% likely that she's an introvert and 75% likely that she's an extrovert ($P_{21}$), or I will think it's certain (i.e. 100% likely) that she's an introvert ($P_{12}$). And let's suppose that, at the five year point:

- $P_1^a(\textit{cred at } t_2 \textit{ is } P_{11}) = \frac{1}{3}$ and $P_1^a(\textit{cred at } t_2 \textit{ is } P_{12}) = \frac{2}{3}$.
- $P_2^a(\textit{cred at } t_2 \textit{ is } P_{21}) = \frac{2}{3}$ and $P_2^a(\textit{cred at } t_2 \textit{ is } P_{22}) = \frac{1}{3}$.

So, again, my current credences (given $P$) are my expectations of my credences at the ten year point (given by $P_{11}$, $P_{12}$, $P_{21}$, or $P_{22}$), and my five year credences (given by $P_1$ or $P_2$) are my expectations of my ten year credences (given by $P_{11}$, $P_{12}$, $P_{21}$, or $P_{22}$)—that is, my current credences and my credences in five years' time both satisfy the Reflection Principle.[47]

Again, as my credence that Sophie is an introvert change, so does my utility for that outcome. Here's how my credences at the ten year point match with my utilities at that time:

- If my credence function is $P_{11}$, then my utility function is $U_{11}$, where $U_{11}^a(w_1) = 6$ and $U_{11}^a(w_2) = -10$;
- If my credence function is $P_{11}$, then my utility function is $U_{11}$, where $U_{12}^a(w_1) = 0$ and $U_{12}^a(w_2) = 2$;

[47] After all:

$$\begin{aligned} P_1^a(w_1) &= \frac{1}{4} = \left(\frac{1}{3} \times \frac{3}{4}\right) + \left(\frac{2}{3} \times 0\right) \\ &= P_1^a(\textit{cred at } t_2 \textit{ is } P_{11})P_{11}(w_1) + P_1^a(\textit{cred at } t_2 \textit{ is } P_{12})P_{12}(w_1) \end{aligned}$$

And similarly for $P_2^a(w_1)$.

- If my credence function is $P_{11}$, then my utility function is $U_{11}$, where $U^a_{21}(w_1) = 2$ and $U^a_{21}(w_2) = 0$;
- If my credence function is $P_{11}$, then my utility function is $U_{11}$, where $U^a_{22}(w_1) = -10$ and $U^a_{22}(w_2) = 6$.

And therefore, by Inter-Grain Coherence:

- my utility for being Sophie's parent will be $U^a_{ij}(W_1) = 2$, for each $i, j = 1, 2$.

Furthermore:

- my utility for not becoming Sophie's parent remains unchanged at 1 (i.e. $U^b_{ij}(W_2) = 1$).

Now, let's evaluate adopting Sophie and not adopting from the coarse-grained point of view. On that view, there are two states:

- $S_1 = \langle W_1, U_0, U_1/U_2, U_{11}/U_{12}/U_{21}/U_{22}, P_0, P_1/P_2, P_{11}/P_{12}/P_{21}/P_{22}\rangle$
- $S_2 = \langle W_2, U_0, U_1/U_2, U_{11}/U_{12}/U_{21}/U_{22}, P_0, P_1/P_2, P_{11}/P_{12}/P_{21}/P_{22}\rangle$

If I adopt, the coarse-grained state is determined—it is $S_1$. And, in that state, both now and in the future, my utility for $W_1$ is 2. That is,

$$U^a_0(W_1) = U^a_i(W_1) = U^a_{ij}(W_1) = 2.$$

So, writing $P$ for my aggregate credence function, and $U$ for my aggregate utility function, my aggregate value for adopting is

$$\begin{aligned} V(a) &= P^a(S_1)U^a(S_1) + P^a(S_2)U^a(S_2) \\ &= U^a(S_1) \\ &= \alpha_0 U^a_0(W_1) + \alpha_1 U^a_i(W_1) + \alpha_2 U^a_{ij}(W_1) \\ &= \alpha_0 2 + \alpha_1 2 + \alpha_2 2 = 2 \end{aligned}$$

On the other hand, if I don't adopt, the coarse-grained state is determined as well—it is $S_2$. And both now and in the future, my utility for $W_2$ is 1. That is,

$$U^a_0(W_2) = U^a_i(W_2) = U^a_{ij}(W_2) = 1.$$

So my aggregate value for not adopting is

$$\begin{aligned} V(b) &= P^b(S_1)U^b(S_1) + P^b(S_2)U(S_2) \\ &= U^b(S_2) \\ &= \alpha_0 U^b_0(W_2) + \alpha_1 U^b_i(W_2) + \alpha_2 U^b_{ij}(W_2) \\ &= \alpha_0 1 + \alpha_1 1 + \alpha_2 1 = 1 \end{aligned}$$

So, on the coarse-grained version, $V(a) > V(b)$, and I prefer adopting to not adopting, so I should choose in accordance with that.

Now, let's evaluate adopting Sophie and not adopting from the fine-grained point of view. On that view, there are eight states: for each $i, j, k = 1, 2$, we have

$$s_{ijk} = \langle w_k, U_0, U_i, U_{ij}, P_i, P_{ij} \rangle$$

Then

$$V(a) = \sum_{i,j,k=1,2} P^a(s_{ijk}) U^a(s_{ijk})$$

Now, suppose we define $P^a(s_{ijk})$ as follows (this corresponds to the probabilities on the edges of the tree depicted in Figure 7.1):

| $s_{111}$ | $s_{112}$ | $s_{121}$ | $s_{122}$ | $s_{211}$ | $s_{212}$ | $s_{221}$ | $s_{222}$ |
|---|---|---|---|---|---|---|---|
| $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{2}{12}$ | $\frac{2}{12}$ | $\frac{2}{12}$ | $\frac{2}{12}$ | $\frac{1}{12}$ | $\frac{1}{12}$ |

Now, if we let $\alpha_0 = \alpha_1 = \alpha_2 = \frac{1}{3}$, then

$$\begin{aligned}
U^a(s_{111}) &= \alpha_0 U_0^a(w_1) + \alpha_1 U_1^a(w_1) + \alpha_1 U_{11}^a(w_1) \\
&= \frac{1}{3}(2) + \frac{1}{3}(-10) + \frac{1}{3}(6) = -\frac{2}{3} \\
U^a(s_{112}) &= \alpha_0 U_0^a(w_2) + \alpha_1 U_1^a(w_2) + \alpha_1 U_{11}^a(w_2) \\
&= \frac{1}{3}(2) + \frac{1}{3}(6) + \frac{1}{3}(-10) = -\frac{2}{3} \\
U^a(s_{121}) &= \alpha_0 U_0^a(w_1) + \alpha_1 U_1^a(w_1) + \alpha_1 U_{12}^a(w_1) \\
&= \frac{1}{3}(2) + \frac{1}{3}(-10) + \frac{1}{3}(0) = -\frac{8}{3} \\
U^a(s_{122}) &= \alpha_0 U_0^a(w_2) + \alpha_1 U_1^a(w_2) + \alpha_1 U_{12}^a(w_2) \\
&= \frac{1}{3}(2) + \frac{1}{3}(6) + \frac{1}{3}(2) = \frac{10}{3} \\
U^a(s_{211}) &= \alpha_0 U_0^a(w_1) + \alpha_1 U_2^a(w_1) + \alpha_1 U_{21}^a(w_1) \\
&= \frac{1}{3}(2) + \frac{1}{3}(6) + \frac{1}{3}(2) = \frac{10}{3} \\
U^a(s_{212}) &= \alpha_0 U_0^a(w_2) + \alpha_1 U_2^a(w_2) + \alpha_1 U_{21}^a(w_2) \\
&= \frac{1}{3}(2) + \frac{1}{3}(-10) + \frac{1}{3}(0) = -\frac{8}{3} \\
U^a(s_{221}) &= \alpha_0 U_0^a(w_1) + \alpha_1 U_2^a(w_1) + \alpha_1 U_{22}^a(w_1) \\
&= \frac{1}{3}(2) + \frac{1}{3}(6) + \frac{1}{3}(-10) = -\frac{2}{3} \\
U^a(s_{222}) &= \alpha_0 U_0^a(w_2) + \alpha_1 U_2^a(w_2) + \alpha_1 U_{22}^a(w_2) \\
&= \frac{1}{3}(2) + \frac{1}{3}(-10) + \frac{1}{3}(6) = -\frac{2}{3}
\end{aligned}$$

Thus,

$$
\begin{aligned}
V(a) = \\
&\tfrac{1}{12}\left(-\tfrac{2}{3}\right) + \tfrac{1}{12}\left(-\tfrac{2}{3}\right) + \tfrac{2}{12}\left(-\tfrac{8}{3}\right) + \tfrac{2}{12}\left(\tfrac{10}{3}\right) + \\
&\quad \tfrac{2}{12}\left(\tfrac{10}{3}\right) + \tfrac{2}{12}\left(-\tfrac{8}{3}\right) + \tfrac{1}{12}\left(-\tfrac{2}{3}\right) + \tfrac{1}{12}\left(-\tfrac{2}{3}\right) \\
&\qquad = 0
\end{aligned}
$$

Thus, my aggregate value for adopting is 0.

And both now and in the future, my utility for $W_2$ is 1. That is,

$$U^a_0(W_2) = U^a_i(W_2) = U^a_{ij}(W_2) = 1.$$

So my aggregate value for not adopting is $V(b) = 1$. And so, on the coarse-grained version, $V(a) < V(b)$, so I prefer not adopting to adopting, and I should choose in accordance with that.

The upshot: when I consider whether or not to adopt Sophie, I know that my credences and my utilities will change if I choose to. So I face a particular instance of the problem of choosing for changing selves. However, as we have seen, if I turn to my favoured solution to that problem, the recommendation that decision theory makes depends on the level of grain at which the decision problem is formulated—in the jargon of decision theory, my favoured solution is not partition invariant; rather, it is partition sensitive.

Of course, this is just a single case. Perhaps it is an anomaly? Let's see how Hild's instability objection plays out in the general case. We will assume that your credence function $P_{s,i}$ that they have in state $s$ at time $t_i$ is obtained from their present credence function $P_p$ by conditionalizing on the total evidence they have at that time in that state. So

(i) $P_{s,i}(-) = P_p(-|E_{s,i})$ and $P_{S,i}(-) = P_p(-|E_{S,i})$.

Note that this entails that $P_p$ satisfies the Reflection Principle. Now, let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two different sets of states of the world, where $\mathcal{S}_2$ is a fine-graining of $\mathcal{S}_1$—that is, each state in $\mathcal{S}_1$ is partitioned by some set of states in $\mathcal{S}_2$; each state in $\mathcal{S}_2$ belongs to one and only one state in $\mathcal{S}_1$. We will distinguish between these levels by using upper case variables—such as $S$—for states in $\mathcal{S}_1$ and lower case variables—such as $s$—for states in $\mathcal{S}_2$. Now, the fine-graining only applies to the possible worlds, not to the possible utility functions nor to the propositions learned as evidence. Thus, if $s \in \mathcal{S}_2$ and $S \in \mathcal{S}_1$, and $s$ is in $S$, then

(ii) $U_{s,i} = U_{S,i}$, for all $i$.

(iii) $E_{s,i} = E_{S,i}$, for all $i$.

(iv) $\alpha_{s,i} = \alpha_{S,i}$, for all $i$.

(v) $w^s$ is in $w^S$.

And we assume Inter-Grain Coherence: that is, the utilities at the coarse-grained level are just the expectations of the utilities at the fine-grained level.

(vi) $U^a_{S,i}(w^S) = \sum_{s \in \mathcal{S}_2} P^a_{s,i}(w^s|w^S)U^a_{s,i}(w^s)$.

Now, consider an act $a$ in $\mathcal{A}$. Let's first of all consider your evaluation of $a$ relative to the coarse-grained set of states, $\mathcal{S}_1$, which we'll write $V_1$:

$$\begin{aligned}
V_1(a) &= \sum_{S \in \mathcal{S}_1} P^a_p(S) \sum_{i=1}^{n} \alpha_{S,i} U^a_{S,i}(w^S) \\
&= \sum_{S \in \mathcal{S}_1} P^a_p(S) \sum_{i=1}^{n} \alpha_{S,i} \sum_{s \in \mathcal{S}_2} P^a_{s,i}(w^s|w^S)U^a_{s,i}(w^s) \quad \text{by (vi)} \\
&= \sum_{S \in \mathcal{S}_1} P^a_p(S) \sum_{i=1}^{n} \alpha_{s,i} \sum_{s \in \mathcal{S}_2} P^a_p(w^s|w^S \;\&\; E_{s,i})U^a_{s,i}(w^s) \quad \text{by (i)} \\
&= \sum_{S \in \mathcal{S}_1} P^a_p(S) \sum_{i=1}^{n} \sum_{s \in S} P^a_p(w^s|w^S \;\&\; E_{s,i})\alpha_{s,i} U^a_{s,i}(w^s)
\end{aligned}$$

Next, consider its value relative to the fine-grained set of states, $\mathcal{S}_2$, which we'll write $V_2$:

$$\begin{aligned}
V_2(a) &= \sum_{s \in \mathcal{S}_2} P^a_p(s) \sum_{i=1}^{n} \alpha_{s,i} U^a_{s,i}(w^s) \\
&= \sum_{S \in \mathcal{S}_1} P^a_p(S) \sum_{i=1}^{n} \sum_{s \in S} P^a_p(s|S)\alpha_{s,i} U^a_{s,i}(w^s)
\end{aligned}$$

Now, there are outcomes in which $V_1$ and $V_2$ are guaranteed to agree on act $a$. For instance:

**Proposition 7.3.1** *Suppose*

$$P^a_p(w^s|S) = P^a_p(w^s|w^S \;\&\; E_{s,i})$$

*for each $S$ and $s$ in $S$ and each $1 \leq i \leq n$. Then $V_1(a) = V_2(a)$.*

This condition will not always be satisfied. After all, $S$ is stronger than $w^S$ & $E_{S,i}$. $S$ specifies $w^S$, of course, and $E_{S,i}$ as well. But it also specifies the utilities that you will have at other times throughout state $S$ and the other evidence that you will obtain. However, if you are certain ahead of time how your utilities will develop if you choose $a$, and if you are sure that you will not obtain any new evidence, then Proposition 7.3.1 tells us that $V_1(a) = V_2(a)$—that is, your evaluation of $a$ is partition invariant. But usually that won't be the case and there will be acts $a$ for which $V_1(a)$ and $V_2(a)$ are different—that is, acts for which your evaluation is sensitive to the partition used. And, if they are different, then there are values that lie strictly between them—e.g. $m = \frac{V_1(a)+V_2(a)}{2}$. Let's suppose $V_1(a) < m < V_2(a)$. Then, if we specify an alternative act $b$ such that $U^b_{s,i}(w^s) = m$ for all fine-grained states $s$ in $\mathcal{S}_1$, then $V_1(b) = V_2(b) = m$. So

$$V_1(a) < V_1(b) = V_2(b) < V_2(a)$$

So, relative to the fine-grained version of the decision problem, $b$ is better than $a$, while relative to the coarse-grained version, $a$ is better than $b$. So, we see that the partition sensitivity of our favoured solution is widespread.

## 7.4 Responding to instability

How might we respond to this concern? I think the natural move is to insist that there is a privileged level of description of the world, and it is our credences and utilities concerning the states of the world in that graining that we should aggregate using the method I propose.[48] Indeed, this move fits well with the theme of this section that we should begin by aggregating your reasons for having the attitudes you have, and not the attitudes that you base on those reasons. After all, we might see the problem for ex post aggregation as similar to the problem for ex ante aggregation. There is a hierarchy of levels at which our attitudes sit: there are our preferences, which are determined by our credences and utilities defined over a particular grain of description; but then those credences and utilities in turn are determined by credences and utilities at a finer grain of description; and so on. Now, if we can identify the finest grain of description that we have—the

[48] Note that this is the same solution proposed by Lara Buchak (2013) to the partition sensitivity of her own decision theory, *risk-weighted expected utility theory*, which we'll meet again in Chapter 16. See (Thoma & Weisberg, 2017) for a discussion of this feature of Buchak's theory.

description on the basis of which all others are determined—then we can simply aggregate those. In this way, we might avoid Hild's problem.

What are the candidates for this role? How might we pick out that finest-grained level such that the credences and utilities at that level determine all the others? There may well be principled ways to do this. For instance, there should come a level of description so fine-grained that fine-graining any further won't change the utilities. That is, at this level, everything that determines my utilities has been specified. Specifying anything further changes nothing. Then, as a consequence of Proposition 7.3.1(i), if we calculate our expected utility relative to that level, fine-graining any further won't make any difference—that is, our evaluation function and thus preference ordering will be stable relative to all levels below that level. For instance, if your utility is determined solely by how much pleasure and pain you experience in the state of the world in question, then the most coarse-grained level of description in which that is fully specified is the privileged level we seek. Specify anything further—the pleasure or pain of others, the number of stars in the universe, or whether Mozart was taller or shorter than 5ft4in or exactly that height—and the utilities won't change. To avoid Hild's problem of the instability of preferences—that is, in order to comply with the Independence of Grain principle—we must stipulate that the decisions are made relative to a level of grain at which everything that you care about is specified.

Of course, the problem of choosing for changing selves arises precisely because what you care about changes from one time to another. Thus, we must not only choose the level of grain at which everything that you *currently* care about is specified, but the level of grain at which everything you *ever* care about is specified. That is, the level of grain at which you must make my decision is the one such that, for each of your past, current, and possible future selves, fine-graining any further will not change the utilities that it assigns—that is, the utility assigned to each more fine-grained possibility will be the same as the utility assigned to the possibility of which it is a fine-graining.[49]

This, then, is my proposed solution to the problem of choosing for changing selves. It treats this problem as a judgment aggregation problem. We have a population of individuals—my various selves at different times and in different possible futures—and they may have different credences and different utilities. I adopt an ex post aggregation method to solve the problem. I argued for this in Chapter 6 on the grounds that we should always

[49] Thanks to Laurie Paul for urging me to address this.

aggregate the reasons for your judgments, whenever they are available to us, rather than the judgments themselves. Thus, we should aggregate credences and utilities separately, rather than aggregating the evaluations to which they give rise or the preference orderings based on those evaluations. On the credal side of the ex post method, I propose that we simply use your current credal state, providing that it satisfies the principles of rationality—Probabilism, Conditionalization, and the Principal Principle—that underpin our current response to information about our credences at other times. On the utility side, I propose that we use a weighted average approach. I will argue for this in more detail in Chapter 9. And, to address Hild's problem of unstable preferences, I have argued that we should define the decision problem so that the set of states is grained finely enough that every past, present, and possible future self agrees that everything they care about is specified at that level of grain—below that level, none of their utilities change.

# Chapter 8

# Can we compare utilities between different selves?

Recall Aneri from the beginning of the book:

> **Aneri** is deciding between two career prospects: she has been offered a place on a training programme for new police officers; and she has been offered a position as an conservation officer for her local council. She is trying to decide which offer to accept. Aneri currently values conformity more than she values self-direction, but not much more. She knows that the conservation job provides some scope for self-direction, though not too much. A police officer, on the other hand, has very little room for self-direction. If Aneri's values stay as they are, the conservation role will suit her well, while she will find the role of police officer frustrating. But she also knows that a person's values tend to become 'socialised', at least to some extent. In particular, she knows that she will likely come to value conformity more than she does now if she trains for the police. And, if that's the case, she will not find it frustrating. Indeed, we might suppose that being a police officer will fit to her socialised values very slightly better than her current values fit with the conservation role.

Bearing all of this in mind, I asked, what career should Aneri choose? The answer I sketched in the previous chapter proceeds as follows.

Aneri formulates a particular decision problem in which the two options between which she must choose are: becoming a police officer (*Police*) and becoming a conservation officer (*Conservation*). To do this, she starts by specifying possible states of the world; then she specifies how likely each of these states is under the supposition that she chooses a particular option; then she specifies the utilities of the outcomes of performing each option when each state obtains; then she calculates the expected utility of each option and picks one with maximal expected utility.

Recall: each state must specify three things: (i) a possible world, which details how things are outside her; (ii) the times within that world; (iii) the utility functions that record her values at each of these times.[50]

(i) To specify (i), Aneri must pick the grain at which she is going to formulate her decision problem. Given our purposes in this chapter, we can be quite crude about this. We might in fact just specify two possible worlds: in the first, $w_1$, Aneri becomes a police officer; in the second, $w_2$, she becomes a conservation officer.

(ii) To specify (ii), she might also be quite crude and specify just three times: one in the past, $t_0$, the present moment, $t_1$, and one in the future, $t_2$.

(iii) Then, to specify (iii), she must specify her utility function at each time—$t_0$, $t_1$, and $t_2$—which will assign numerical utilities to the two possible outcomes *Police* & $w_1$ and *Conservation* & $w_2$.[51] More precisely, she must specify her past, present, and future utilities in these two outcomes *on the same scale*. For recall from Chapter 2 that, if one utility function specifies a legitimate numerical representation of an individual's values, so does any positive linear transformation of it. Thus, as with temperature, there are many different possible scales on which to measure value. We must ensure that we measure values on the same scale for an individual at the different times in the different states in our decision problem. After all, just as we wouldn't calculate the mean surface temperature on Earth by taking the surface temperatures at points in the northern hemisphere in celsius and the points in the southern hemisphere in fahrenheit and averaging, so when we wish to aggregate the utilities of different selves to give the aggregate utility of a state, and then compare that to the aggregate utility of some other

[50] In the previous chapter, we also required a state to specify how Aneri's credences evolve throughout that state. But that was only for the purpose of considering Hild's instability objection. Henceforth, we drop that requirement.

[51] We ignore the other two outcomes because we are certain that by choosing to become a police officer, Aneri will become a police officer, and by choosing to become a conservation officer, she will become a conservation officer.

> state, we want to ensure that all of the utilities in play are measured on the same scale.

There will be two such states: one with world $w_1$, which we'll call state $s_1$, where Aneri becomes a police officer; the other with world $w_2$, which we'll call $s_2$, where Aneri becomes a conservation officer. In both possible states, Aneri's values in the past and at the present are the same: she values conformity most, but not much more than self-direction. So, we might say that, at $t_0$ and $t_1$, Aneri's utility for the outcome *Police* & $w_1$ is 3, while her utility for the other outcome *Conservation* & $w_2$ is 5.

In the state where she becomes a conservation officer—that is, $s_2$—her values don't change and so she retains these same utilities at $t_2$ in the state, *when measured on the same scale*. In contrast, in the state in which she becomes a police officer—that is, $s_1$—her values change: she comes to value conformity more and self-direction less. So, perhaps, in that state, her utility at $t_2$ for *Police* & $w_1$ is 10, while her utility for *Conservation* & $w_2$ is 1. Again, more precisely, we say that these are her utilities *on the same scale* we used to specify her utilities at $t_0$ and $t_1$.

With this in hand, we have specified the two possible states of the world, $s_1$ and $s_2$. Then, to specify our aggregate utility in each state, we take a weighted average of the utilities that Aneri assigns at the different times in that state to the world that is actual in that state. Thus, her aggregate utility in $s_1$ is

$$(\alpha_{s_1,0} \times 3) + (\alpha_{s_1,1} \times 3) + (\alpha_{s_1,2} \times 10)$$

while her aggregate utility in $s_2$ is

$$(\alpha_{s_2,0} \times 5) + (\alpha_{s_2,1} \times 5) + (\alpha_{s_2,2} \times 5)$$

While we must wait until the second half of the book to discuss how we should set these weights, let's assume for the moment that Aneri completely discounts her past selves, so that $\alpha_{s_1,0} = \alpha_{s_2,0} = 0$. And let's suppose that her weightings don't depend on the state, so that $\alpha_{s_1,1} = \alpha_{s_2,1} = \alpha$ and $\alpha_{s_1,2} = \alpha_{s_2,2} = 1 - \alpha$. So her aggregate utility in $s_1$ is $3\alpha + 10(1 - \alpha)$, while her aggregate utility in $s_2$ is $5\alpha + 5(1 - \alpha) = 5$. Thus, since choosing to be a police officer necessitates $s_1$, and choosing to be a conservation officer necessitates $s_2$, Aneri should choose the former if $3\alpha + 10(1 - \alpha) > 5$; that is, if $\alpha < \frac{5}{7}$. She should choose the latter if $3\alpha + 10(1 - \alpha) < 5$; that is, if $\alpha > \frac{5}{7}$. And she may choose either if the two are equal.

Our topic in this chapter is how we accomplish the latter steps in the process above, where we give the numerical values that specify the utilities that

Aneri assigns to the different outcomes at different times and in different states. I take there to be two tasks in this area: first, I need to explain what we are doing when we give numerical representations of values, which is after all what utilities are, and how we achieve this; and second, I need to explain what it means to say that the utilities of two different selves are measured on the same scale, and how we achieve such measurements.

I trust that the discussion above of Aneri's decision establishes the need for numerical representations of her values, and the need to give them on the same scale. Without this, we cannot say, for instance, how much more value Aneri must obtain from being a police officer, from her future point of view as a police officer, and how much weight that future police officer self must receive in order to make it rational for her to choose to change her values by training to become a police officer. And indeed, something like this is always the reason we want such numerical representations of well-being or value or happiness, and the reason we want them on the same scale. We need them in order to adjudicate trade-offs. In order to allocate scarce resources earmarked for healthcare, we need to know how much worse it is to suffer a year with kidney failure than a year with medically-managed diabetes (Bognar & Hirose, 2014); we need to know how much better £1, 000 is for me than £300 in order to decide whether or not I should take the latter for sure or a 50-50 bet on the former (von Neumann & Morgenstern, 1947); and we need to know how much more Caro values free access to wilderness spaces than Don values free access to libraries in order to know which should be our priority for government funding if Caro and Don are both residents in our country whose opinions we wish to include (Sen, 2017).

## 8.1 Representing values with numbers

So, first, the numerical representation itself. What do the numbers represent that I attach to outcomes like *Police* & $w_1$ and *Conservation* & $w_2$ as my utilities? Like the numbers we use to represent temperature, they measure a particular quantity; they say how much there is of some thing that comes in different amounts. The quantity in question is the strength or intensity of my desire for the outcome in question; it is the degree to which I value it.

We talk of such a quantity often. Sometimes we make categorical statements about it: I want to be a musician; I value the life of the mind. Sometimes we make ordinal statements: I prefer being a musician to being a stonemason; I value being a conservation officer more than being a police officer. Sometimes we make cardinal statements about it: I value a walk in the woods much more than being stuck in rush-hour traffic; I value walking to work a little more than cycling; for me, the difference in value between watching musical theatre and watching rugby is greater than the difference in value between eating a cupcake and eating tree bark.

Together, these suggest that there is a quantity here that we might represent. But we can say more. We can specify its functional role in the workings of the mind; and we can say how we might measure it. And surely this is sufficient to establish its legitimacy. We ask no more of other mental items about which we theorise: for beliefs, we note that we talk about them in our folk psychology, we specify a functional role for them in our mental life, and we give a reliable but fallible method for attributing such states to subjects.

The functional role of these strengths or intensities of desires is just a more nuanced version of the functional role of categorical desires. Thus, for instance, just as desiring something causes emotions such as disappointment when you learn that it won't be yours, so different strengths of desire cause different strengths of disappointment. The more I value my friend finding happiness, the more I am disappointed when they are sad. Desiring is also related to hope: if I desire something and I discover it might happen, I hope for it. Strengths of desire cause strengths of hope. The more I desire a Labour government, the more strongly I will hope that they will win at the next election. Just as desires interact with beliefs to give rise to action, so strengths of desires interact with beliefs and with strengths of belief to give rise to action. Thus, if I desire one thing more strongly than another, and I believe that I can obtain the first by one action and the second by another, I will choose the first action. And the more strongly I desire something the more of a risk I will take to obtain it—that is, the less confident I have to be that an action will obtain it for me in order to take that action all the same. As the Hungarian-American economist, John Harsanyi, puts it in this moving autobiographical example:

> if a person is known to have risked his life in order to obtain a university education (e.g., by escaping from a despotic government which had tried to exclude him from all higher education), then we can take this as a reasonably sure sign of his attaching very high personal importance (very high utility) to such an education. (Harsanyi, 1977a, 643-4)

This latter functional role is crucial for us here, for it forms the basis of the method by which we seek to measure these strengths of utility. The idea is that we will measure how intensely an individual values an outcome by the risks they are willing to take to acquire it, just as Harsanyi says. The method is due to Ramsey (1931) and von Neumann & Morgenstern (1947). Here it is, set out more precisely:

- Let $o_{\text{worst}}$ be the outcome this individual values least;
- Let $o_{\text{best}}$ be the outcome she values most;
- Pick two real numbers, $a < b$;
- Let her utility for $o_{\text{worst}}$ be $a$ and her utility for $o_{\text{best}}$ be $b$.

  That is, $U(o_{\text{worst}}) = a$ and $U(o_{\text{best}}) = b$.
- Given any outcome $o$, her utility for that outcome will be determined by finding the gamble between $o_{\text{worst}}$ and $o_{\text{best}}$ that she considers exactly as good as getting $o$ for sure.

  In particular:

  - Let $p_o$ be the probability such that our individual is indifferent between outcome $o$ for sure, on the one hand, and $p_o$ chance of $o_{\text{best}}$ and $1 - p_o$ chance of $o_{\text{worst}}$, on the other.
  - Her utility for $o$ is $(1 - p_o)a + p_o b$.
    That is, $U(o) = (1 - p_o)U(o_{\text{worst}}) + p_o U(o_{\text{best}})$.

In this way, we might measure the utility this individual assigns to each outcome. And notice that all we appealed to at any point was the ordering of outcomes and gambles on outcomes by their value. That is, we moved from ordinal information concerning the desirability of gambles on outcomes to cardinal information about the desirability of those outcomes. Few doubt the psychological reality of the former; and we leverage that to bolster our case in favour of the psychological reality of the latter.

Now notice: we placed no restrictions on the numbers that represent the top and bottom of the scales other than that they should be ordered in the correct way, so that $o_1$ has lower utility than $o_2$. Thus, we might set $a = 0$ and $b = 1$, or $a = -100$ and $b = 100$, or anything else. This is why any positive linear transformation of a utility function is as good as a representation of an individual's values as the original utility function. It simply results from choosing different values for the utilities of the worst and best outcomes.

Sometimes, this method is taken not as a technique by which to *measure* the strength of an individual's desires, but rather as a definition of what it *means* to have a certain utility in an outcome. But I have no truck with that sort of behaviourist interpretation. I take a realist view here. Strengths of desires are defined by their functional role, and we have sketched some aspects of that above. It is perfectly possible that a mental item doesn't always play its functional role well. Sometimes something goes wrong and I don't feel disappointment when I realise I won't get something I desire strongly. And similarly, sometimes something goes wrong and I value an outcome very highly but won't take much of a risk to obtain it. So this method of measurement can sometimes misfire, just as using a thermometer to measure temperature can sometimes misfire. But it will usually work just fine. And it serves its purpose here, which is to bolster our case that there is a quantity, the strength or intensity of our desires, that we can represent numerically.

Granted the method described above, given two selves—Aneri before she becomes a police officer and Aneri afterwards—what really represents the values of each self is not a single utility function, but rather a set of utilities functions, any two of them positive linear transformations of one another. In order to aggregate these, we need to pick just one utility function from each; and, as we emphasised above, we must ensure that both measure utility on the same scale. How are we to do that?

In the social choice literature, where we wish to pick a utility function for each individual in society and aggregate them, and where again we must ensure that the utility functions we pick all measure utility on the same scale, this is known as the *problem of interpersonal utility comparisons*. I can't see inside your mind and you can't see inside mine. So how can we tell whether the utility function we use to represent the strength of your desires measures those strengths on the same scale as the one we use to represent the strength of my desires? How can we tell whether, when your utility function assigns 4 to an outcome and mine assigns 6, we can conclude that I value that outcome more than you do? In the remainder of this chapter, I will spell out a number of standard solutions to the problem of interpersonal utility comparisons that have been proposed in the social choice case, and we will ask whether any of them might solve the analogous problem for different selves; what we might call the *problem of interself/intrapersonal utility comparisons*, which was first raised by R. A. Briggs (2015). As we will see, none of the standard proposals will work in our case. But something closely related will. We conclude by describing that proposal.

## 8.2 Empathetic preferences

The most famous proposal in this area is Harsanyi's appeal to so-called *empathetic preferences* (Harsanyi, 1977b). When we specified our method for measuring the utility Aneri assigns to an outcome, we began by assuming that she put those outcomes in order from worst to best—these are, of course, her preferences over those outcomes. Then we assumed that she made judgments of indifference between different gambles on those outcomes. Harsanyi thinks she can do more.

Suppose Aneri has two friends, Ben and Camille. She knows them well. Then not only does Aneri have her own preference ordering over the outcomes, and her own judgments of indifference between gambles over them; she also has what Harsanyi would call an empathetic preference ordering over outcome-friend pairs, and judgments of indifference between gambles over these pairs. An outcome-friend pair is a pair $(o, i)$, where $o$ is an outcome and $i$ is a friend—for Aneri, either Ben or Camille. Thus, Aneri can judge not only whether she prefers being Aneri with outcome $o$ over being Aneri with outcome $o'$, but also whether she prefers being Ben with outcome $o$ over being Camille with outcome $o'$. And she can make judgments of indifference between gambles over such outcome-friend pairs. So she can judge whether being Ben with outcome $o$ for sure is exactly as good as a gamble that makes you Ben with outcome $o^*$ with chance $p$ and Camille with outcome $o^\dagger$ with chance $1 - p$. She does this by empathetically inhabiting their perspective. A necessary condition on her doing this successfully, of course, is that, when restricted just to the outcome-friend pairs of the form $(-, \text{Ben})$—that is, where Ben is always the second argument—the ordering or indifference judgments agree with Ben's personal ordering; and similarly for Camille. That is, Aneri ranks $(o, \text{Ben})$ above $(o', \text{Ben})$ iff Ben ranks $o$ above $o'$, and she ranks $(o, \text{Camille})$ above $(o', \text{Camille})$ iff Camille ranks $o$ over $o'$.

If she does manage that, then we can construct measures of utility for Ben and Camille that are guaranteed to be on the same scale. Thus, let $(o_{\text{worst}}, i_{\text{worst}})$ be the outcome-friend pair at the bottom of Aneri's empathetic preference ordering, and let $(o_{\text{best}}, i_{\text{best}})$ be the pair at the top. Then, as before, we pick $a < b$ and let $U_A(o_{\text{worst}}, i_{\text{worst}}) = a$ and $U_A(o_{\text{best}}, i_{\text{best}}) = b$ be Aneri's utilities for those pairs. Next, suppose Aneri judges a gamble in which she is friend $i$ with outcome $o$ for sure to be exactly as good as a gamble on which she is $i_{\text{best}}$ with outcome $o_{\text{best}}$ with chance $p_o$ and $i_{\text{worst}}$ with outcome $o_{\text{worst}}$ with chance $1 - p_o$. Then $U_A(o, i) = (1 - p_o)a + p_o b$. That is, we set Aneri's utilities for the various outcome-friend pairs exactly as we set her utilities for outcomes alone in the standard case. And we pick utility functions for Ben and Camille on the same scale by letting Ben's utility in outcome $o$ be Aneri's utility in being Ben with outcome $o$ and letting Camille's utility in outcome $o$ be Aneri's utility in being Camille with outcome $o$—that is, $U_B(o) = U_A(o, \text{Ben})$ and $U_C(o) = U_A(o, \text{Camille})$.

Might Aneri use the same trick to ensure that when she specifies her past, present, and future utilities in the states of the world in her decision problem, she specifies them all on the same scale? You might think that the accurate empathy required in such a case will be more easily achievable than in the standard case: empathising with other selves is easier than empathising with other people. And perhaps it is on average. But some of the cases that exercise us most when we consider choosing for changing selves are precisely those in which empathy with future selves is so difficult. Can I empathise with my possible future parent-self sufficiently that I can judge accurately whether being me currently and in an outcome in which I lose my job, say, is better or worse than being my future parent-self and being in an outcome in which my child loses their job? It's not obvious that I can. And indeed it is a central thesis of L. A. Paul's *Transformative Experience* that we cannot (Paul, 2014a).[52]

## 8.3 The Zero-One Rule

Let's turn, then, to another attempt to ensure that Aneri measures her current utilities on the same scale as her past and future utilities. It begins with the following insight. Suppose we have two individuals, or two selves. The values of the first are represented by the set of utility functions $\mathcal{U}_1$, while the values of the second are represented by $\mathcal{U}_2$. Next, pick a utility function $U_1$ from $\mathcal{U}_1$. Then we can fix the utility function $U_2$ in $\mathcal{U}_2$ that measures utility on the same scale if we can find just four outcomes, $o_1, o_1', o_2, o_2'$, for which we want to say the following:

(i) neither individual is indifferent between $o_1$ and $o_2$;

(ii) neither individual is indifferent between $o_1'$ and $o_2'$;

(iii) $o_1$ is exactly as good for the first individual as $o_2$ is for the second;

(iv) $o_1'$ is exactly as good for the first individual as $o_2'$ is for the second.

[52] For more on this, see Chapter 10.

Suppose we can do that. Then we just pick $U_2$ from $\mathcal{U}_2$ so that $U_1(o_1) = U_2(o_2)$ and $U_1(o'_1) = U_2(o'_2)$. And it turns out that there is just one utility function in $\mathcal{U}_2$ for which that holds—there are infinitely many for which $U_1(o_1) = U_2(o_2)$, and infinitely many for which $U_1(o'_1) = U_2(o'_2)$, but only one for which both hold.

How might we discover these anchor points, $o_1$, $o_2$ and $o'_1$, $o'_2$? One suggestion is that they can be determined by attending only to the meaning of utility. We might think that it is an analytic or conceptual truth, for instance, that the outcome I consider worst must be exactly as bad for me as the outcome that you consider worst is for you. And similarly for our best outcomes. Thus is sometimes called the *zero-one rule*, since we might dictate that an individual's worst outcome always receives utility 0 from them, whilst their best outcome receives utility 1. We thereby obtain a single scale on which to measure utility for all individuals.

In the interpersonal case, in which the individuals whose utility we wish to compare are different people, the problems are well known (Griffin, 1986; Hammond, 1991; Sen, 2017). Consider, for instance, an individual, call him Michael, with a vivid and dark imagination. He is forever dreaming up more and more nightmarish scenarios, more and more horrifying forms of torture inflicted on his loved ones, more and more monstrous ways a human life can go. As he does this, each of these worse and worse outcomes takes the place of the previous worst outcome he considered. In line with the zero-one rule, they are successively assigned utility 0. But this then pushes up the utilities assigned to everything above them. So Michael's utility for having a lemon ice cream increases as he imagines worse and worse possible outcomes. Similarly, consider Kimmy, Michael's mirror image, who imagines better and better outcomes, filling her outcome space with wonderfully imagined ways in which everyone might be happy and content and fulfilled. Then, as each of these replaces the previous best outcome in Kimmy's set, it is assigned utility 1, in line with the zero-one rule, and those below it receive less utility. This doesn't seem right. You can't make things worse for yourself simply by imagining hypothetical good things nor make things better for yourself by imagining hypothetical bad things.

Of course, there are some people who feel happier the further their situation is from the worst possible situation and those who feel sadder the further their situation is from the best. For some people who suffer a major calamity, they report lower levels of well-being immediately after the calamity but soon start reporting well-being at much the same level as they did beforehand. One putative explanation is that, while they now inhabit a situation that they value less, in some sense, they have seen truly awful possibilities they hadn't imagined before the calamity, and because they apply the zero-one rule to their well-being reports, they report the worse situation as having the same numerical well-being as the previous better situation. But these are reported well-being levels. They are not measures of how much an individual values an outcome. And it seems much less plausible to say that a person genuinely changes how much they value an outcome when they imagine a new worst possible outcome.

## 8.4 The Fixed Points Solution

So the zero-one rule won't do. Another suggestion that also attempts to fix the utility of two outcomes comes from the literature on moral uncertainty—I'll call it the *fixed points solution*. Suppose I don't know which is the correct moral theory. I know that, whatever it is, it is a rights-based theory, but I don't know which of the many such theories is correct. Or I know that it is a utilitarian theory, but I don't know whether the utility to be maximised is Bentham's hedonic utility or the satisfaction of preferences. Suppose I wish to make a decision with morally relevant consequences. Then, as I mentioned in the opening paragraphs of Chapter 6, I face a judgment aggregation problem. I wish to make the decision in line with the true moral theory, but I don't know which that is. So I want to aggregate the judgments of the possible theories in some way. Perhaps I wish to take some measure of the values those theories assign to each outcome and then set my utilities for those outcomes to be a weighted average of theirs—perhaps the weights reflect my credences in the various theories, so that the weighted average is just my expectation of the true moral value of the outcome. If I am to do this, I must ensure that I am measuring the value that each moral theory assigns on the same scale. How might I do that?

Jacob Ross (2006) makes the following suggestion. While the different possible moral theories between which I'm uncertain disagree on many things, they also agree on a great deal. For instance, the different rights-based theories might disagree on the relative values of an outcome in which ten people are tortured and another in which one person is killed. But if an outcome involves no violations of rights, those theories must agree on its value. Thus, to ensure that we are measuring the outcomes on the same scale, we need only find two outcomes $o_1$, $o_2$ such that (i) neither $o_1$ nor $o_2$ involves any violations of rights, and (ii) none of the candidate moral theories is indifferent between $o_1$ and $o_2$. Then we pick a utility function that measures the values assigned by the first and call it $U$. And we choose the utility function $U'$ that measures the values assigned by another of the theories such that $U(o_1) = U'(o_1)$ and $U(o_2) = U'(o_2)$. And this ensures that $U$ and $U'$ measure value on the same scale.

Note that Ross' solution is unlikely to help in the case of interpersonal utility comparisons. In the case of the moral theories, we know that there are outcomes to which each of the theories assign the same value because we know everything about those theories—we constructed them ourselves. But in the interpersonal case, it is not possible to specify outcomes such that we can be sure two individuals agree on the value of those two outcomes. Even if I know that Aneri and Blandine agree that it is unimportant whether the tree outside their front door is an ash or an oak, I can't assume they assign equal utility to that outcome. I can't assume that what is unimportant for Aneri has the same utility for her as what is unimportant for Blandine.

You might think, however, that Ross' solution would work well for the intrapersonal/interself case. After all, my future self has a good deal more insight into the mind of my past selves than I have into your mind or the minds of even my close friends and family. And surely we can expect my future self to be able to judge when they value something to the same extent as they used to—that is, when they and some past self assign the same utility to an outcome. What's more, however great the change in our values, presumably there are always some things we continue to value to the same extent. When Aneri becomes a police officer and her values have become socialised so that she is now more conformist, this surely changes only her utilities in situations which demand more or less conformity. It won't change, for instance, the utility assigns to eating chocolate ice cream, nor the utility she assigns to eating lemon sorbet. So, if her utilities for those two outcomes are different from one another—she prefers chocolate to lemon, for instance—we can use those as our anchor points to ensure that we measure her present and future utilities on the same scale.

Unfortunately, it's not quite so simple. Recall from above: utilities are primarily defined on the finest-grained outcomes; they are then defined on coarser-grained outcomes in line with Inter-Grain Coherence from Chapter 6. Let's consider the finest-grained outcomes first. Since these outcomes specify everything the individual cares about, it is actually rather unlikely that we'll be able to identify some finest-grained outcome that we value to the same extent before and after our values change. For instance, Aneri's present self values self-determination more than her future officer-self does. But each finest-grained outcome will specify exactly how much she will be able to determine for herself, and thus exactly how much this particular desire of hers will be satisfied. So, given an outcome that involves conformity, Aneri will value it more in the future, once she's a police officer and her values have become socialised; and given any outcome that does not involve conformity, Aneri will value it less in the future. There will be no outcome that she values to the same extent before and after she becomes a police officer. And the same holds for my decision whether or not to adopt Sophie from Chapter 7. Each finest-grained outcome specifies whether or not I adopt Sophie. Thus, I will value any outcome in which I do adopt her more in the future, when I am Sophie's parent, than now, when I am not. And I will value any outcome in which I am not Sophie's parent less in the future when I am her parent than now when I am not. So we cannot hope to apply Ross' technique to finest-grained outcomes, since the extent to which we value those will change whenever almost any of our values change.

But surely we can apply it to coarser-grained outcomes? And that would be sufficient. Again, I think not, and for related reasons to before. Let's focus on Aneri's present self, where she values conformity less than her future police-self will. Aneri's future police-self might say: 'I've come to value conformity more; but the value I assign to eating chocolate ice cream hasn't changed, nor the value I assign to eating vanilla ice cream, and I value the former more than the latter'. Can we not then pick $U_1$ and then fix $U_2$ by demanding that $U_1(\textit{Eat chocolate ice-cream}) = U_2(\textit{Eat chocolate ice-cream})$ and $U_1(\textit{Eat vanilla ice-cream}) = U_2(\textit{Eat vanilla ice-cream})$? And would that not be sufficient? Well, it would be sufficient if we could do it; but unfortunately we can't. The problem is that there are two ways to hear Aneri's assertion. On the first, it concerns the values she assigns to the coarse-grained outcomes *Eat chocolate ice-cream* and *Eat vanilla ice-cream*. On the second, it concerns the contribution that her consumption of chocolate or vanilla ice cream makes to the overall value of a given finest-grained outcome. In order to apply Ross' solution, we must interpret her in the first way. But on that reading what she says is false.

To see why, let's look again at how the utility of a coarse-grained outcome is related to the utility of the finest-grained outcomes compatible with it. For simplicity, let's suppose that all that Aneri cares about are her career and the ice cream flavour she eats. So the finest-grained outcomes compatible with eating chocolate ice cream are: *Eat chocolate ice cream* & *Police Officer*, and *Eat chocolate ice cream* & *Conservation Officer*. Now, her present self then sets her utility in coarse-grained outcomes in which she eats chocolate ice cream as follows:

$$\begin{aligned} U^a_0(\textit{Choc}) = P^a_0(\textit{Choc \& PO}|\textit{Choc})U^a_0(\textit{Choc \& PO}) + \\ P^a_0(\textit{Choc \& CO}|\textit{Choc})U^a_0(\textit{Choc \& CO}) \end{aligned}$$

And her future police-self sets that utility similarly:

$$U^a_1(Choc) = P^a_1(Choc\ \&\ PO|Choc)U^a_1(Choc\ \&\ PO) + P^a_1(Choc\ \&\ CO|Choc)U^a_1(Choc\ \&\ CO)$$

Now, Aneri's future police-self values being a police officer more than her current self does, and values being a conservation officer less. So:

$$\begin{array}{rcl} U^a_1(Choc\ \&\ PO) & > & U^a_0(Choc\ \&\ PO) \\ U^a_1(Choc\ \&\ CO) & < & U^a_0(Choc\ \&\ CO) \end{array}$$

What's more, an important feature of Aneri's story is that, if she becomes a police officer, her values will become socialised to such an extent that her future police-self values being a police officer a little bit more than her current self values being a conservation officer:

$$U^a_1(Choc\ \&\ PO) > U^a_0(Choc\ \&\ CO)$$

But Aneri's future police-self is also certain that she chose to be a police officer, and not a conservation officer, so $P^a_1(PO) = 1$ and $P^a_1(CO) = 0$. So

$$\begin{array}{rcl} P^a_1(Choc\ \&\ PO|Choc) & = & 1 \\ P^a_1(Choc\ \&\ CO|Choc) & = & 0 \end{array}$$

And thus,

$$U^a_1(Choc) > U^a_0(Choc)$$

So it just isn't true that Aneri assigns the same utility to the coarse-grained outcome on which she eats chocolate ice cream before and after her values socialise to being a police officer; and similarly for vanilla ice cream.

## 8.5 The Fixed Interval Solution

So, when Aneri says that, while she's come to value conformity more since becoming a police officer, the values she assigns to eating chocolate or vanilla ice cream haven't changed, she cannot mean that her values for the two coarse-grained outcomes *Eat chocolate ice cream* and *Eat vanilla ice cream* are the same now as they were before she became an officer. Rather, she must be talking of the value that eating them adds to a particular fine-grained outcome. Thus, at most, what she might mean is this: if $o$ is a finest-grained outcome in which Aneri eats chocolate ice cream and $o'$ is one that is identical in all respects except that she eats vanilla ice cream instead, then

$$U_0(o) - U_0(o') = U_1(o) - U_1(o')$$

Now this helps us a little. When we fix a scale on which to measure utility or temperature we fix a *zero* and we fix a *unit*. The zero of the Celsius scale, for instance, sits at the freezing point of water at sea level, while the unit is the change of temperature such that a hundred increases of this size reaches the boiling point of water at sea level. And, were we to follow the doomed zero-one rule proposed above, the zero of your utility function would be your worst outcome and the unit would be the change of utility between that and your best outcome. If two utility functions $U, U'$ that both represent your utilities share the same zero, one is a scaling of the other: that is, there is $\alpha$ such that $U' = \alpha U$. If they share the same unit, one is a translation of the other: that is, there is $\beta$ such that $U' = U + \beta$. If we could say that the difference between $o$ and $o'$ is the same for Aneri's present and future selves, then we could ensure that we are measuring her utilities on scales with the same unit. But it would not help to fix the zero.[53]

Now, you might be forgiven for wondering whether it is in fact necessary to fix the zero as well as the unit; you might be forgiven for thinking that it is sufficient to fix the unit only. After all, suppose we wish to choose between two options $a$ and $b$. And suppose we want to do so on the basis of our own credences and the values of two individuals. We will aggregate the utility functions that record the values of those two individuals by taking a weighted average of them, with weights $\alpha_1$ and $\alpha_2$. We have been able to fix the unit for the second individual's utility function relative to the first individual's, but not the zero. Thus, we begin by picking a utility function $U_1$ to represent the values of the first individual. This sets the unit for the utility function that represents the values of the second individual, but not the zero for that utility function. So we pick one of the utility functions with the correct unit, say, $U_2$.Then the aggregate utility for an outcome $a$ & $s$ is:

$$\alpha_1 U_1^a(s) + \alpha_2 U_2^a(s)$$

And the expected aggregate utility for option $a$ is:

$$\sum_s P^a(s)\left(\alpha_1 U_1^a(s) + \alpha_2 U_2^a(s)\right)$$

Now, suppose we pick a different utility function to represent the values of second individual. Since we have been able to fix the unit but not the

[53] Note that, if $U' = U + \beta$, then

$$U'(o) - U'(o') = (U(o) + \beta) - \left(U(o') + \beta\right) = U(o) - U(o')$$

for any two outcomes $o, o'$.

zero, this must have the form $U_2 + \beta$, where $\beta$ is any real number. Thus, the aggregate utility of $a$ & $s$ is:

$$\alpha_1 U_1^a(s) + \alpha_2 \left(U_2^a(s) + \beta\right) = \alpha_1 U_1^a(s) + \alpha_2 U_2^a(s) + \alpha_2 \beta$$

And the expected aggregate utility for option $a$ is:

$$\sum_s P^a(s) \left(\alpha_1 U_1^a(s) + \alpha_2 \left(U_2^a(s) + \beta\right)\right) = \sum_s P^a(s) \left(\alpha_1 U_1^a(s) + \alpha_2 U_2^a(s)\right) + \alpha_2 \beta$$

So,

$$\begin{aligned}
\sum_s P^a(s) \left(\alpha_1 U_1^a(s) + \alpha_2 U_2^a(s)\right) &< \sum_s P^a(s) \left(\alpha_1 U_1^b(s) + \alpha_2 U_2^b(s)\right) \\
&\text{iff} \\
\sum_s P^a(s) \left(\alpha_1 U_1^a(s) + \alpha_2 U_2^a(s)\right) + \alpha_2 \beta &< \sum_s P^a(s) \left(\alpha_1 U_1^b(s) + \alpha_2 U_2^b(s)\right) + \alpha_2 \beta \\
&\text{iff} \\
\sum_s P^a(s) \left(\alpha_1 U_1^a(s) + \alpha_2 (U_2^a(s) + \beta)\right) &< \sum_s P^a(s) \left(\alpha_1 U_1^b(s) + \alpha_2 (U_2^b(s) + \beta)\right)
\end{aligned}$$

Thus, one option is evaluated as better than the other relative to the first representation of the second individual's values iff they are also evaluated in that order relative to the second representation. That is, fixing the unit but not the zero of the second individual's utility function relative to the first individual's is sufficient if you wish to compare the expectations of their weighted average aggregate utilities—whichever representation you use that respects the fixed unit will give you the same answer.

However, the same is not true when we move into our framework. Here's the first problem: in our framework, but not in the example just given, different states of the world can include different selves with different utility functions. For instance, take the state in which Aneri becomes a police officer and the state in which she becomes a conservation officer. Then her self a particular time in the first has a different utility function from her self at that same time in the second. Thus, suppose that each state I am considering contains just two times, $t_1$ and $t_2$, and thus two selves. Now, suppose that, in state $s$, we have fixed the unit of the second self's utility function relative to the first, but not the zero; and similarly for state $s'$. Then we pick the utility functions of my first self in $s$ to be $U_{s,1}$, and we pick the utility function of my second self to be one of those that respects the fixed unit, say, $U_{s,2}$; and similarly for $s'$, we pick $U_{s',1}$ to be the utility function of my first self in that state and $U_{s',2}$ to be the utility function of my second self. Then the aggregate value of $a$ & $s$ on the first representation is:

$$\alpha_{s,1}U^a_{s,1}(w^s) + \alpha_{s,2}U^a_{s,2}(w^s)$$

while their aggregate value for $a$ & $s'$ is:

$$\alpha_{s',1}U^a_{s',1}(w^{s'}) + \alpha_{s',2}U^a_{s',2}(w^{s'})$$

Now, suppose we pick a different utility function to represent the values of my second self in $s$. Then it must have the form $U_{s,2} + \beta$, for some real number $\beta$. And suppose we do likewise for my second self in $s'$. So it has the form $U_{s',2} + \beta'$, for some $\beta'$. So, relative to this representation, the aggregate value for $a$ & $s$ is

$$\alpha_{s,1}U^a_{s,1}(w^s) + \alpha_{s,2}\left(U^a_{s,2}(w^s) + \beta\right) = \alpha_{s,1}U^a_{s,1}(w^s) + \alpha_{s,2}U^a_{s,2}(w^s) + \alpha_{s,2}\beta$$

while their aggregate value for $a$ & $s'$ is:

$$\alpha_{s',1}U^a_{s',1}(w^{s'}) + \alpha_{s',2}\left(U^a_{s',2}(w^{s'}) + \beta'\right) = \alpha_{s',1}U^a_{s',1}(w^{s'}) + \alpha_{s',2}U^a_{2}(w^{s'}) + \alpha_{s',2}\beta'$$

And thus, there is no guarantee that

$$\begin{array}{rcl}
\alpha_{s,1}U^a_{s,1}(w^s) + \alpha_{s,2}U^a_{s,2}(w^s) & < & \alpha_{s',1}U^a_{s',1}(w^{s'}) + \alpha_{s',2}U^a_{s',2}(w^{s'}) \\
 & \text{iff} & \\
\alpha_{s,1}U^a_{s,1}(w^s) + \alpha_{s,2}U^a_{s,2}(w^s) + \alpha_{s,2}\beta & < & \alpha_{s',1}U^a_{s',1}(w^{s'}) + \alpha_{s',2}U^a_{2}(w^{s'}) + \alpha_{s',2}\beta' \\
 & \text{iff} & \\
\alpha_{s,1}U^a_{s,1}(w^s) + \alpha_{s,2}\left(U^a_{s,2}(w^s) + \beta\right) & < & \alpha_{s',1}U^a_{s',1}(w^{s'}) + \alpha_{s',2}\left(U^a_{s',2}(w^{s'}) + \beta'\right)
\end{array}$$

Indeed, providing the individual's values at time $t_1$ or time $t_2$ are different in $s$ and in $s'$—so that $\beta$ and $\beta'$ can be chosen independently of one another—it will always be possible to pick them so that these equivalences do not hold. The upshot: in our framework, we must fix not only the unit, but also the zero. Nothing short of that will do even if we wish only to compare the aggregate utilities for two outcomes $a$ & $s$ and $a$ & $s'$, let alone the expected aggregate utilities of two options.[54]

[54] Note that, for the same reason, in an interpersonal rather than intrapersonal case, it does not suffice to fix the unit and not the zero if the options between which you are picking might affect the people whose values you wish to aggregate, or the values of the people you wish to aggregate.

## 8.6 The Matching Intervals Solution

Unfortunately, then, Ross' solution won't work for us. But it does point towards an alternative solution. As we saw in Aneri's case, when an individual's values change, this usually gives rise to changes in the utility they assign to *every* finest-grained outcome. If I adopt Sophie and I come to value being her parent more, my values in all finest-grained outcomes in which I am her parent increase and my values in all finest-grained outcomes in which I am not her parent decrease. But just as I might think a future self could judge when its utility for an outcome is the same as that of some recent past self, which was the proposal in our first attempt to transpose Ross' solution to our case, and just as I might think it could judge when the difference between its utilities for two outcomes hasn't changed, so we might think that it could judge when the difference between its past and present utilities for an outcome is the same as the difference between its past (or present) utilities for two outcomes. Suppose I can say, for four outcomes $o_1, o_2$ and $o'_1, o'_2$, that the difference between my current utilities in $o_1$ and $o_2$ is the same as the difference between my current utility in $o_2$ and my future utility in $o_2$, and likewise for $o'_1$ and $o'_2$, then:

$$\begin{array}{rcl} U_0(o_1) - U_0(o_2) & = & U_0(o_2) - U_1(o_2) \\ U_0(o'_1) - U_0(o'_2) & = & U_0(o'_2) - U_1(o'_2) \end{array}$$

This then fixes the utilities of $U_1(o_2)$ and $U_1(o'_2)$ once I've fixed $U_0$. And, as we noted in our discussion of the zero-one solution and Ross' solution, this is sufficient to ensure that $U_0$ and $U_1$ measure utility on the same scale; it is sufficient to fix both zero and unit.

Thus, perhaps Aneri can judge that the difference between her past and future utilities in the finest-grained outcome in which she eats chocolate ice cream and she is a police officer ($o_2$) is equal to the difference between her past utility in that outcome and her past utility in the finest-grained outcome in which she eats vanilla ice cream and she is a police officer ($o_1$). This, then, is our solution to the problem of intrapersonal/interself utility comparisons—I call it the *matching intervals solution*. An important feature of it is that it does not require individuals to be able to introspect the absolute degrees of intensity in their desires. Rather, it requires individuals only to be able to compare differences between the utilities of outcomes, either the same outcome at different times or different outcomes at the same time. This, I submit, is a reasonable requirement to place upon them.

# Chapter 9

# Why aggregate using weighted averages?

In Chapter 6, we considered a number of different ways in which we might aggregate the attitudes and judgments of a group of individuals who boast credences, utilities, evaluation functions, and preference orderings. We considered how we might aggregate preferences directly, such as through the Borda count method; we considered the weighted average ex ante method, on which we combine each individual's credence and utility function first to give their evaluation function, and then aggregate the resulting evaluation functions to give the aggregate evaluation function of the whole group; and we considered the weighted average ex post method, on which we aggregate the individuals' credences and utilities separately first to give the group credences and utilities, and then combine those aggregates to give the aggregate evaluation function. In each case, whenever the judgments were represented numerically, we aggregated them by taking weighted averages of the numerical representations, a species of aggregation method that is known as *linear pooling*.

Thus, the Borda count method, which provides a numerical representation of the ordinal preference ordering, takes the group Borda count for a particular option to be the average of the Borda counts of the individuals in the group—and of course straight averages are just weighted averages in which each individual is assigned the same weight. The weighted average ex ante method takes the group evaluation function to be a weighted average of the individual evaluation functions. The weighted average ex post method does the same for the group credences and group utilities. And of course our favoured method does the same: the utility it assigns to a given state $s$ is a weighted average of the utilities assigned to the world of that state, $w^s$, by the selves that have, do, and will exist in it.

What is much less obvious, but true all the same, is that the expected utility norm (EU1) from Chapter 2 above is itself a judgment aggregation norm, and indeed a weighted averaging or linear pooling norm at that. Recall: (EU1) says that your evaluation for an act should be your expectation of the utility of that act. That is,

(EU1) $V(a)$ is your subjective expectation of the utility of $a$.

That is,

$$V(a) = \sum_{s \in \mathcal{S}} P(s||a)U(a \,\&\, s) = \sum_{s \in \mathcal{S}} P^a(s)U^a(s)$$

Let's see what I mean by saying that (EU1) is a weighted averaging or linear pooling norm. Consider a standard decision problem, equipped with a set of states, $\mathcal{S}$, and a set of acts, $\mathcal{A}$. You are considering how to evaluate an act $a$ from $\mathcal{A}$. You know your utilities for each of the situations $a \,\&\, s$, where $s$ is a state in $\mathcal{S}$. But you don't know which state is the actual state, so you need to assign credences to each of the states under the assumption of $a$, and you need to set your evaluation for $a$. However, you do know what your credences *would* be if you *were* to know which state is actual—you'd give that state (and any proposition true at that state) credence 1, and you'd give any other state (and any proposition false at the state you know is actual) credence 0. What's more, you do know what your evaluation for an act *would* be if you *were* to know which state is actual—it would be your utility for that act at that state. Thus, you might view this as a judgment aggregation problem in which the individuals you wish to aggregate are these alternative know-it-all selves, your maximally well-informed counterparts at the different possible states who know what state it is that they are in. Thus, for each possible state of the world, you have what we might call a *know-it-all counterpart* at that state. This know-it-all counterpart assigns maximal credence—the full 100%, or credence 1—to all the propositions that are true in that state of the world; and they assign the minimal credence—0%, or credence 0—to all those that are false. That is, they get everything right about that state. And their utilities are just the same as yours. Moreover, since they live not in a state of uncertainty, but in a state of certainty, their utility function and their evaluation function coincide. Thus, if we consider your know-it-all counterpart at state $s$, then their evaluation of a particular act $a$ is just your utility for the situation $a \,\&\, s$. A little more formally: given a state $s$, let your know-it-all counterpart at $s$ have credence function $P^a_s$, utility function $U^a_s$, evaluation function $V_s$, and preference ordering $\preceq_s$. Your attitudes, on the other hand, are $P^a$, $U^a$, $V$, and $\preceq$. Then:

- $P^a_s(X) = 1$, if proposition $X$ is true at state $s$, and $P^a_s(X) = 0$, if $X$ is false at $s$;
- $U^a_s(s) = U^a(s)$;
- $V_s(a) = U^a_s(s)$.

Now, suppose you are considering how to set your attitudes. You have set your utility function $U^a$. But you have not yet set your credence function $P^a$, your evaluation function $V$, nor your preference ordering $\preceq$. A natural thing to say is that your credences and your evaluations are the aggregated credences and evaluations of your know-it-all counterparts. Thus, your credences are obtained by aggregating the credences of your know-it-all counterparts, while your evaluations are obtained by aggregating their evaluations. Suppose, then, that we aggregate these all together using the method of weighted arithmetic averages; that is, by linear pooling. Then there are weights—a weight $0 \leq \alpha_s \leq 1$ for each state of the world $s$, and thus each know-it-all counterpart—such that $\sum_{s \in \mathcal{S}} \alpha_s = 1$, and each attitude you have is the weighted average of the corresponding attitude of these individuals. Thus:

(a) Your credence in a proposition $X$ is

$$P^a(X) = \sum_{s \in \mathcal{S}} \alpha_s P^a_s(X) = \sum_{s \in X} \alpha_s$$

In particular, $P^a(s) = \alpha_s$.

(b) Your evaluation of $a$ is

$$V(a) = \sum_{s \in \mathcal{S}} \alpha_s V_s(a) = \sum_{s \in \mathcal{S}} \alpha_s U^a(s) = \sum_{s \in \mathcal{S}} P^a(s) U^a(s)$$

This entails two important conclusions. First, from (a), we can infer that my credences should be probabilities—that is, they should satisfy the axioms of the probability calculus. These make two demands:[55]

(P1a) My credence in a tautology should be 1.

[55] To see this, first note that it follows from (a) that $P(X) = \sum_{s \in X} P(s)$. Now:

(P1a) $P(s) = \alpha_s$ and $\sum_{s \in \mathcal{S}} \alpha_s = 1$. So, if $\top$ is a tautology, then $P(\top) = \sum_{s \in \top} \alpha_s = \sum_{s \in \mathcal{S}} \alpha_s = 1$. Thus, (P1a).

(P1b) My credence in a contradiction should be 0.

(P2) My credence in a disjunction, $X$ or $Y$, should be the sum of my credence in $X$ and my credence in $Y$ with my credence in their conjunction, $X$ and $Y$, subtracted.

Second, from (b), we can infer that my value for a particular act is my subjective expectation of its utility, just as the expected utility norm (EU1) requires.

## 9.1 The Argument for Linear Pooling

So we have been assuming at various points in the preceding chapters that the correct way to aggregate the numerically represented judgments of a group of individuals—their credences, their utilities, their values, or their Borda scores—is to take a weighted average of those judgments. But this is not the only way to aggregate such judgments. We might instead take a weighted *geometric* mean as opposed to a weighted *arithmetic* mean, which is the weighted average we have been using so far—this is known as *geometric pooling*. Or we might take the harmonic mean, the mid-range, the median, or the geometric median. Or we might try some quite different technique. Thus, if we are to provide a foundation for the approach that we have been taking so far, we must justify our use of weighted arithmetic averages, in particular. We do this in the present chapter. Having done it, we will have justified our favoured approach to the problem of choosing for changing selves—granted that it is a judgment aggregation problem, any attitudes to be aggregated (in our case, only the utilities) should be aggregated using weighted arithmetic averaging. But, as we saw at the end of the previous section, we will have done two more things as well: (i) we will have justified *probabilism*, which says that a rational individual's credences should be probabilities; and (ii) we will have justified (EU1), which says that an individual's evaluation of an act should be their subjective expectation of its utility.

(P1b) If $\perp$ is a contradiction, then $P(\perp) = \sum_{s \in \perp} P(s) = 0$. Thus, (P1b).

(P2)

$$\begin{aligned} P(X \vee Y) &= \sum_{s \in X \vee Y} P(s) \\ &= \sum_{s \in X} P(s) + \sum_{s \in Y} P(s) - \sum_{s \in X \,\&\, Y} P(s) \\ &= P(X) + P(Y) - P(X \,\&\, Y) \end{aligned}$$

### 9.1.1 The Principle of Minimal Mutilation

The argument that we will present is based on what I will call *the principle of minimal mutilation*.[56] Here's the idea. When we aggregate the attitudes of a group of individuals who disagree—that is, when we take a collection of different sets of attitudes towards the same set of items, and try to aggregate this into a single set of attitudes concerning those same items—what we end up with must differ from the sets of attitudes held by at least some of the individuals in the group, and will typically differ from all of them. But, with most sorts of attitudes, and certainly with the numerically represented attitudes that we are concerned to aggregate here, a miss is not as good as a mile—some sets of attitudes lie further from an individual's attitudes than others. Thus, a person who thinks it's 90% likely to rain is closer to someone who thinks it's 80% likely to rain than to someone who thinks that rain is 5% likely. Similarly, someone who assigns a utility of 10 to being a parent is closer to someone who assigns 9 to that experience than to someone who assigns 1 (when the utilities of all three measure the extent to which they value being a parent on the same scale). The principle of minimal mutilation says, roughly, that the aggregate of the attitudes of a group of individuals should lie as close as possible to the different attitudes of the individuals in the group—it should not lie further than necessary from them. Thus, consider Omar, Pepijn, and Quentin—here are their utilities for going to the cinema and going to play darts (see Figure 9.1):

| | Omar | Pepejn | Quentin | $U_G$ | $U_G^*$ |
|---|---|---|---|---|---|
| Cinema | 3 | 8 | 6 | 6 | 6 |
| Darts | 6 | 9 | 3 | 1 | 4 |

Suppose I were to offer $U_G$ as their aggregate, which assigns a utility of 6 for going to the cinema and 1 for playing darts. This seems obviously wrong, and the reason is that there are alternative putative aggregates that are closer

[56] See (Pettigrew, 2015a, taa) for earlier versions of this argument in particular cases. The idea that something like the principle of minimal mutilation should be used when aggregating doxastic attitudes originates in the computer science literature (Konieczny & Pino-Pérez, 1998, 1999; Konieczny & Grégoire, 2006). There, they are interested not in aggregating numerically represented attitudes, but categorical attitudes, such as full beliefs or commitments (Miller & Osherson, 2009). This method was studied first in the judgment aggregation literature by Gabriella Pigozzi (2006). In the case of probabilistic aggregation, something related has been considered by Predd et al. (2008); Pettigrew (2017). The claim that minimizing average or total distance from individual's attitudes is the correct way to aggregate conative attitudes—or mixtures of conative and doxastic attitudes, such as preference orderings—is much older (Kemeny, 1959; Fishburn, 1977; Young & Levenglick, 1978; Saari & Merlin, 2000).

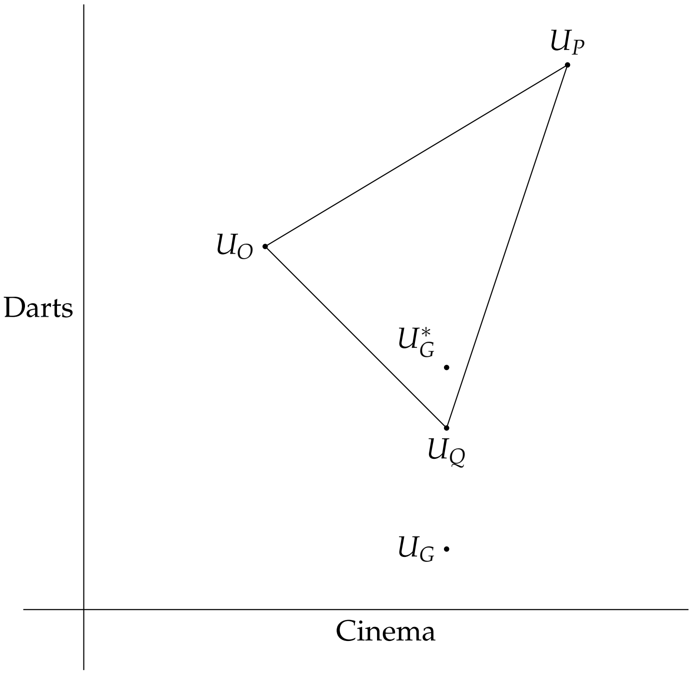


Figure 9.1: Here, we plot the utilities of Omar, Pepejn, Quentin, and the two putative aggregates on the Euclidean plane. The $x$-coordinate gives their utility for going to the cinema; the $y$-coordinate gives their utility for playing darts. The triangle formed by drawing straight lines between each of them contains all and only the weighted arithmetic averages of the three sets of attitudes.

to each of Omar, Pepejn, and Quentin. Take $U_G^*$, for instance, which assigns 6 to the cinema and 4 to darts. Then, intuitively, $U_G^*$ is closer to each of Omar, Pepejn, and Quentin than $U_G$ is: it assigns the same utility as Quentin to the cinema, but the difference between its darts utility and Quentin's is less than the difference between the darts utility of $U_G$ and Quentin's; moreover, the cinema utility of $U_G^*$ is exactly as far from Omar's as the cinema utility of $U_G$ is, and similarly for Pepejn; and the darts utility of $U_G^*$ is closer to Omar's than the darts utility of $U_G$, and similarly for Pepejn. Thus, the principle of minimal mutilation rules out $U_G$ as the aggregate—the attitudes represented by $U_G$ lie further from the attitudes to be aggregated than is necessary.

### 9.1.2 The Dominance Argument for Weighted Averages

We will argue that numerically represented attitudes should be aggregated by taking weighted arithmetic averages in this way. We will show that, for any putative aggregate that isn't a weighted arithmetic average of the individual attitudes, there is an alternative that is closer to all of the individuals than that putative aggregate; and we will show that the same issue does not arise for putative aggregates that are weighted arithmetic averages. In the presence of the principle of minimal mutilation, this shows that a putative aggregate that is not a weighted arithmetic average is no aggregate at all.

To do this, we must specify a measure of the distance between sets of numerically represented attitudes. It will, in fact, be what is usually called *Euclidean distance*, and it is the standard distance between two points in space.[57] The idea is this: While Euclidean distance might be the natural measure of distance between two points in space, there is no immediately obvious reason why it should be the measure of distance between sets of attitudes. To show that it is, we proceed as follows: First, we lay down a set of properties that we would like our measure of distance to have. Next, we show that only the members of a rather select group of distance measures have all of these properties together—they are the Euclidean distance

[57] Suppose $\langle a_1, a_2, a_3\rangle$ and $\langle b_1, b_2, b_3\rangle$ are coordinates for points in 3-dimensional space. Then the Euclidean distance from one to the other is

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+(a_3-b_3)^2}$$

In general, if $\langle a_1, \ldots, a_n\rangle$ and $\langle b_1, \ldots, b_n\rangle$ are points in $n$-dimensional space, the Euclidean distance between them is

$$\sum_{i=1}^{n}\sqrt{(a_i-b_i)^2}$$

measure together with any positive transformation of it.[58] We conclude that only these measures are legitimate. Having done this, we will show that, if a putative aggregate is not a weighted arithmetic average of the individuals' attitudes, there is an alternative that is closer to each of those individuals when that distance is measured using Euclidean distance, or any positive transformation of it. And thus, by the principle of minimal mutilation, we conclude that only weighted arithmetic averages can serve as aggregates.

Let's meet the properties we would like to see in a measure of distance.[59] Throughout, we will assume that all of the individuals between whom we might wish to measure the distance have attitudes towards the same set of items—these items may be states or propositions to which the individuals assigns credences, or they may be situations to which they assign utilities, or they may be acts to which they assign evaluations. Let's write the set of items towards which each of our individuals has attitudes $X_1, \ldots, X_m$. Thus, to specify an individual, we write the sequence $\mathbf{a} = \langle a_1, \ldots, a_m \rangle$ of numerical representations of their attitudes towards $X_1, \ldots, X_m$, respectively. Thus, when we ask for a measure of distance from one individual to another, we are asking for a function $d_m : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty]$. $d_m$ takes pairs of sequences of numerical representations, $\mathbf{a}$ and $\mathbf{b}$, and returns a measure of distance from one to the other, $d_m(\mathbf{a}, \mathbf{b})$, which is either a non-negative real number or infinity.

Let's meet the first feature that we would like our measure of distance to have. It is called *extensionality*, and it tells us something about what the distance between two individuals can depend upon. Given a pair of individuals, $\mathbf{a}$ and $\mathbf{b}$, the distance between them depends only on the following multiset:[60]

$$\{\{(a_1, b_1), \ldots, (a_m, b_m)\}\}$$

Formally:

[58] One measure of distance, $d'$, is a positive transformation of another, $d$, if there is a strictly increasing function that, when applied to the distance as measured by $d$ gives the distance as measured by $d'$. That is, there is a function $H$ from real numbers to real numbers such that (i) if $x < y$, then $H(x) < H(y)$, and (ii) $d'(\mathbf{a}, \mathbf{b}) = H(d(\mathbf{a}, \mathbf{b}))$.

[59] These are adapted from (D'Agostino & Sinigaglia, 2010), which are in turn adapted from (D'Agostino & Dardanoni, 2009).

[60] A multiset is a collection that, like a set and unlike a sequence, ignores order—so that while the sequences $\langle 1, 1, 2 \rangle$ and $\langle 2, 1, 1 \rangle$ are different, the multisets $\{\{1, 1, 2\}\}$ and $\{\{2, 1, 1\}\}$ are identical—but, unlike a set and like a sequence, can contain the same element more than once—so that while the sets $\{1, 1, 2\}$ and $\{1, 2, 2\}$ are the same, and the same as $\{1, 2\}$, the multisets $\{\{1, 1, 2\}\}$, $\{\{1, 2, 2\}\}$, and $\{\{1, 2\}\}$ are different.

**Extensionality** If

$$\{\{(a_1, b_1), \ldots, (a_m, b_m)\}\} = \{\{(c_1, d_1), \ldots, (c_m, d_m)\}\}$$

then

$$\begin{aligned} & d_m(\langle a_1, \ldots, a_m \rangle, \langle b_1, \ldots, b_m \rangle) \\ = \; & d_m(\langle c_1, \ldots, c_m \rangle, \langle d_1, \ldots, d_m \rangle) \end{aligned}$$

Thus, suppose the utilities of Raquel, Siobhan, Tilo, and Ursula are as follows:

| | Raquel | Siobhan | Tilo | Ursula |
|---|---|---|---|---|
| Cinema | 10 | 8 | 7 | 5 |
| Darts | 7 | 5 | 10 | 8 |

Then, according to Extensionality, Raquel lies exactly as far from Siobhan as Tilo lies from Ursula. After all:

$$\begin{aligned} & \{\{(U_R(C), U_S(C)), (U_R(D), U_S(D))\}\} \\ = \; & \{\{(10, 8), (7, 5)\}\} \\ = \; & \{\{(U_T(C), U_U(C)), (U_T(D), U_U(D))\}\} \end{aligned}$$

The second condition is *agreement invariance*. Suppose two individuals, Vivek and Winnie, start off with utilities in two situations, one in which they are outside without an umbrella in the rain, one in which they are outside without an umbrella in the dry. They then realise that there is a third possible situation, one in which they are outside without an umbrella in the snow. They both assign exactly the same utility to this third situation—that is, they agree perfectly upon it. Then Agreement Invariance says that the distance between them has not changed as a result of adopting this new attitude, since both of them adopted the same attitude. Formally:

**Agreement Invariance**

$$d_{m+1}(\langle a_1, \ldots, a_m, c \rangle, \langle b_1, \ldots, b_m, c \rangle) = d_m(\langle a_1, \ldots, a_m \rangle, \langle b_1, \ldots, b_m \rangle)$$

Another way of putting this: the distance between two individuals depends only on their attitudes towards items about which they disagree—adding new attitudes towards items about which they agree changes nothing.

The third condition, *difference supervenience*, says that, when we consider two individuals with attitudes only towards a single item—a utility in a single situation, for instance, or a credence in a single proposition—then the distance between those attitudes should be some increasing and continuous function of the difference between them. Formally:

**Difference Supervenience** There is a strictly increasing and continuous function $g : \mathbb{R} \to \mathbb{R}$ such that

$$d_1(\langle a \rangle, \langle b \rangle) = g(|a - b|)$$

Thus, looking back to the case of Raquel, et al. from above, we see that Raquel's and Siobhan's attitudes towards the cinema lie exactly as far apart as their attitudes towards darts—although they each have different values in the cinema and in darts, the difference between those attitudes is the same. Thus, according to difference supervenience, the distance between their attitudes towards the cinema is the same as the distance between their attitudes to darts.

The fourth condition is well known from discussions in social welfare theory. It is called *separability*.[61] It says that, if Xavier and Yasmin are equally far from Zola on items $X_1, \ldots, X_m$, and if Xavier is closer to Zola than Yasmin is on items $X_{m+1}, \ldots, X_{m+m'}$, then Xavier is closer to Zola than Yasmin is on $X_1, \ldots, X_m, X_{m+1}, \ldots, X_{m+m'}$. Formally:

**Separability** If

(i)

$$\begin{aligned} & d_m(\langle a_1, \ldots, a_m \rangle, \langle c_1, \ldots, c_m \rangle) \\ = \quad & d_m(\langle b_1, \ldots, b_m \rangle, \langle c_1, \ldots, c_m \rangle) \end{aligned}$$

(ii)

$$\begin{aligned} & d_{m'}(\langle a_{m+1}, \ldots, a_{m+m'} \rangle, \langle c_{m+1}, \ldots, c_{m+m'} \rangle) \\ < \quad & d_{m'}(\langle b_{m+1}, \ldots, b_{m+m'} \rangle, \langle c_{m+1}, \ldots, c_{m+m'} \rangle) \end{aligned}$$

then

$$\begin{aligned} & d_{m+m'}(\langle a_1, \ldots, a_{m+m'} \rangle, \langle c_1, \ldots, c_{m+m'} \rangle) \\ < \quad & d_{m+m'}(\langle b_1, \ldots, b_{m+m'} \rangle, \langle c_1, \ldots, c_{m+m'} \rangle) \end{aligned}$$

Taken together, these first four conditions—Extensionality, Agreement Invariance, Difference Supervenience, and Separability—already restrict the

[61] In the social welfare context, we find it first in (Fleming, 1952), but also in (Young, 1974; Arrow, 1977).

range of legitimate distance measures significantly. Any distance measure that satisfies all four conditions has the following form: for any $\mathbf{a}, \mathbf{b}$,

$$d_n(\langle a_1, \ldots, a_n \rangle, \langle b_1, \ldots, b_n \rangle) = H\left(\sum_{i=1}^{n} g(|a_i - b_i|)\right)$$

where $H$ and $g$ are strictly increasing and continuous functions from the real numbers to the real numbers.[62]

Our final condition ensures that $g(x) = x^2$. To motivate it, consider the following situation: three cousins, Anya, Anke, Aneri, and their friend Ben have utilities over four different options, *archery*, *badminton*, *curling*, and *darts*. Their utilities are as follows:

| | Archery | Badminton | Curling | Darts |
|---|---|---|---|---|
| Anya | 10 | 7 | 4 | 1 |
| Anke | 7 | 10 | 4 | 1 |
| Aneri | 10 | 7 | 1 | 4 |
| Ben | 8 | 4 | 6 | 2 |

Now, notice that Anya and Ben order Archery and Badminton the same way; and they order Curling and Darts the same way. Moreover, Anke's utilities are obtained from Anya's by swapping the utilities in Archery and Badminton, and keeping the utilities in Curling and Darts fixed, while Aneri's utilities are obtained from Anya's by swapping the utilities in Curling and Darts, and keeping the utilities in Archery and Badminton fixed. And notice that the difference between Anya's utilities in Archery and Badminton is the same as the difference between her utilities for Curling and Darts, and similarly for Anke, Aneri, and Ben.

Our final condition, The Badness of Order-Reversing Swaps, makes two claims. The first says that Anke and Aneri must lie further from Ben than Anya does. The idea is that, when we determine how far one set of numerically represented attitudes lies from another, we should look not only at how far the individual numerical values assigned to the various items lie from one another, as Difference Supervenience requires us to do, but also at the extent to which the first set of attitudes orders those items in the same way as the second. Thus, for instance, it says that Anke lies further from Ben than Anya does because Anya agrees with Ben on the ordering of Archery and Badminton, and Anke is obtained from Anya by swapping her utilities in those two options so that she disagrees with Ben on their ordering. And similarly for Aneri.

[62] See the proof of Theorem 1(1) in (D'Agostino & Dardanoni, 2009).

You might see this condition as complementing Difference Supervenience. In the presence of Separability, Extensionality, and Agreement Invariance, Difference Supervenience militates in favour of a rather local approach to assessing the distance between two sets of attitudes—it seems to suggest that we look at the pairs of attitudes individually, assess the distance between those, and then aggregate those individual distances. This might lead you to worry that there are global features of sets of attitudes that are thereby excluded from the assessment of distance. The Badness of Order-Reversing Swaps is intended to ensure that at least one global feature, or at least non-local feature, of a set of attitudes—namely, the way in which those attitudes order the items towards which they are directed—is included in the assessment of distance.

The second part of The Badness of Order-Reversing Swaps says that Anke and Aneri lie equally far from Ben, because they are both obtained from Anya by order-reversing swaps, and the differences between the two utilities that they swap are equal, as are the differences between Ben's utilities in those two options.

Putting these two together, we have the following formal version:

**The Badness of Order-Reversing Swaps** Suppose

(i) $a_i, a_j$ and $b_i, b_j$ are ordered in the same way;

(ii) $a_p, a_q$ and $b_p, b_q$ are ordered in the same way;

(iii) $|a_i - a_j| = |a_p - a_q|$ and $|b_i - b_j| = |b_p - b_q|$;

Then

$$\begin{array}{rl} & d_m(\langle a_1, \dots, a_i, a_j, a_p, a_q, \dots a_m \rangle, \langle b_1, \dots, b_i, b_j, b_p, b_q, \dots, b_m \rangle) \\ < & d_m(\langle a_1, \dots, a_j, a_i, a_p, a_q, \dots a_m \rangle, \langle b_1, \dots, b_i, b_j, b_p, b_q, \dots, b_m \rangle) \\ = & d_m(\langle a_1, \dots, a_i, a_j, a_q, a_p, \dots a_m \rangle, \langle b_1, \dots, b_i, b_j, b_p, b_q, \dots, b_m \rangle) \end{array}$$

Now, as D'Agostino & Dardanoni (2009, Theorem 1(1)) prove, when taken together, these five conditions—Extensionality, Agreement Invariance, Difference Supervenience, Separability, and The Badness of Ordering-Reversing Swaps—entail that our measure of distance between two sets of numerically represented attitudes has the following form:

$$d_m(\langle a_1, \dots, a_m \rangle, \langle b_1, \dots, b_m \rangle) = H\left(\sqrt{\sum_{i=1}^{m}(a_i - b_i)^2}\right)$$

where $H$ is a continuous and strictly increasing function. That is, $d_m$ is a positive transformation of Euclidean distance.

How does this help? Because of the following fact, which is illustrated in Figure 9.2:[63]

**Theorem 9.1.1 (Dominance Theorem)** *Suppose we have a set of $n$ individuals, $\mathbf{a}^1, \ldots, \mathbf{a}^n$, with numerically-represented attitudes towards items $X_1, \ldots, X_m$. And now suppose that $\mathbf{b}$ is a putative aggregate of the attitudes of these $n$ individuals. Then, if our distance measure $d$ satisfies the five conditions above—that is, if $d$ is a positive transformation of Euclidean distance—then:*

(I) *If $\mathbf{b}$ is* not *a weighted arithmetic average of the $\mathbf{a}^i$s, then there is an alternative $\mathbf{b}^*$ such that, for all $1 \leq i \leq n$,*

$$d_m(\mathbf{a}^i, \mathbf{b}^*) < d_m(\mathbf{a}^i, \mathbf{b})$$

*That is, each individual $\mathbf{a}^i$ lies closer to $\mathbf{b}^*$ than to $\mathbf{b}$.*

(II) *If $\mathbf{b}$ is a weighted average of the $\mathbf{a}^i$s, then, for any alternative $\mathbf{b}^* \neq \mathbf{b}$, there is $1 \leq i \leq n$ such that*

$$d_m(\mathbf{a}^i, \mathbf{b}^*) > d_m(\mathbf{a}^i, \mathbf{b})$$

*That is, there is some individual $\mathbf{a}^i$ that lies closer to $\mathbf{b}$ than to $\mathbf{b}^*$.*

By the principle of minimal mutilation, therefore, any aggregate $\mathbf{b}$ of the $\mathbf{a}^i$s should be a weighted arithmetic average of them. That is, there should be weights $0 \leq \alpha_1, \ldots, \alpha_n \leq 1$ such that $\sum_{i=1}^n \alpha_i = 1$ and, for each item $1 \leq k \leq m$,

$$b_k = \sum_{i=1}^n \alpha_i a^i_k$$

where $b_k$ is the attitude of individual $\mathbf{b}$ to item $X_k$, while $a^i_k$ is the attitude of individual $\mathbf{a}^i$ to item $X_k$. If there are not, there is some alternative aggregate $\mathbf{b}^*$ that is closer to the attitudes of every single one of the individuals, and thus a more minimal mutilation of those attitudes than $\mathbf{b}$.

[63] This theorem is discussed initially by de Finetti (1974), who proved it in the case that interests us here. It has been generalised significantly by Predd et al. (2009) so that it applies to a much broader range of distance measures.

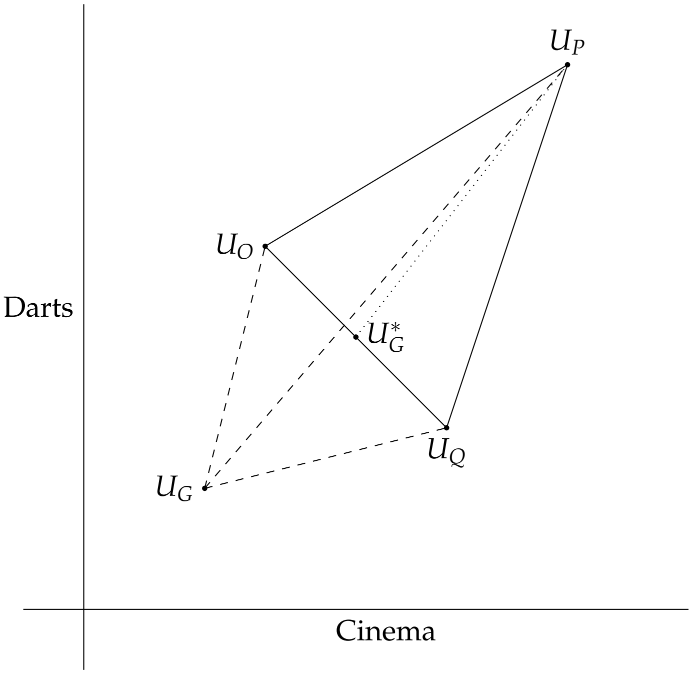


Figure 9.2: Again, we plot the utilities of Omar, Pepejn, and Quentin on the Euclidean plane (as $U_O$, $U_P$, $U_Q$, respectively). Consider the candidate aggregate utility function $U_G$, which assigns utility 2 to both Cinema and Darts. It is not a weighted average of the individual utility functions; all such weighted averages lie inside the triangle. Therefore, by Theorem 9.1.1 there is an alternative candidate aggregate, $U_G^*$, that is closer to each individual utility function than $U_G$ is. A particular $U_G^*$ is illustrated here. The dashed lines show the distance from $U_G$ to the various different individuals; the dotted lines show the distance from $U_G^*$ to those individuals.

## 9.2 The features of linear pooling

This completes our argument in favour of aggregating attitudes represented numerically by taking weighted arithmetic averages. However, we should not rest easy quite yet. All arguments in the judgment aggregation literature proceed in the same way. First, they present a set of features that an aggregation method might or might not have; second, they argue that these are desirable features of such a method, features that you would ideally want your method to have; third, they show that their favoured aggregation method and only that method boasts all of those features; and finally they conclude that their favoured method is the correct one. And indeed our argument has exactly this form: we showed that weighted arithmetic averaging and only weighted arithmetic averaging ensures that the aggregate of the attitudes of a group of individuals does not lie unnecessarily far from those attitudes. The problem is that now, seventy years after the modern version of the discipline was born, there is a vast array of these apparently desirable features, and it is well known that no method boasts all of them—that is, not only is there no existing, already-formulated method that boasts them all; we know that, *as a matter of mathematical fact*, there cannot possibly be any such method—this is the lesson of the myriad impossibility results. As a result, in order to persuade your audience that your particular favoured method of aggregation is the correct one, the sort of argument described above does not suffice. It is not enough merely to show that your method and yours alone has some desirable features. Given any particular rival method, there are likely to be some desirable features that it and it alone boasts. So, as well as showing that your method uniquely boasts certain desirable features, you must also show that the allegedly desirable features that it lacks—those that your rivals boast—either are not desirable after all, or are less desirable than the features that your method has, so that lacking the former is a price worth paying in order to secure the latter. That is what we will do now.

I'll begin by looking at the versions of Arrow's conditions that apply in the context of numerically-represented attitudes. I'll show that versions of linear pooling in fact satisfy all three. This is an illustration of the oft-noted fact that features that cannot be jointly satisfied when we aggregate attitudes represented ordinally often pose no problem when we aggregate attitudes represented cardinally. After that, I'll consider two apparently desirable features that aggregate credences lack when they result from taking weighted averages. In this case, I'll argue that one feature is not in fact desirable, and that our method satisfies the other when it is formulated correctly.

### 9.2.1 Arrow's conditions

Let's start with Arrow's Weak Pareto condition. Recall: in the preference ordering case, this says that if every individual prefers $b$ to $a$, then so does the group. Formally: if $a \prec_i b$ for each $1 \leq i \leq n$, then $a \prec_G b$. Of course, this is just one of a number of related principles, each one a unanimity preservation principle:

- If $a \preceq_i b$ for each $1 \leq i \leq n$, then $a \preceq_G b$.
- If $a \prec_i b$ for each $1 \leq i \leq n$, then $a \preceq_G b$.
- If $a \sim_i b$ for each $1 \leq i \leq n$, then $a \sim_G b$.

In the context of numerically represented attitudes, the following are natural unanimity preservation principles:

- If $a^i_k < a^i_l$, for each individual $i$ and items $X_k$, $X_l$, then $a^G_k < a^G_l$.
- If $a^i_k \leq a^i_k$, for each individual $i$ and items $X_k$, $X_l$, then $a^G_k \leq a^G_l$.
- If $a^i_k = a^i_l$, for each individual $i$ and items $X_k$, $X_l$, then $a^G_k = a^G_l$.
- If $a^i_k = r$, for each individual $i$ and item $X_k$, then $a^G_k = r$.

And it is easy to see that linear pooling satisfies all of these conditions.

Next, consider No Dictator. In the preference ordering case, this says that there is no individual such that, whatever her preferences and whatever the preferences of the other individuals, the aggregate agrees with her about everything. Formally: there is no $i^*$ such that for all $a, b$ in $\mathcal{A}$, $a \preceq_G b$ iff $a \preceq_{i^*} b$. In our context, where the attitudes aren't represented ordinally but cardinally, this becomes: there is no $i^*$ such that, for all $1 \leq k \leq m$, $a^{i^*}_k = a^G_k$. Now, there are certainly linear pooling methods that violate this: if there is an individual $i$ who receives all of the weight, regardless of what credences she assigns, so that $\alpha_j = 1$ if $i = j$ and $\alpha_j = 0$ if $i \neq j$, then the resulting method violates No Dictator. But if that isn't the case—for instance, if we insist that $0 < \alpha_1, \ldots, \alpha_n < 1$—then No Dictator is satisfied.

Next, consider the Independence of Irrelevant Alternatives. Again, we find ourselves in a situation in which linear pooling is compatible with this feature and also compatible with its lack. Whether a version of linear pooling satisfies the Independence of Irrelevant Alternatives depends on whether you set the weights you're going to use independently of the attitudes that you're going to aggregate, or whether you wait to see the attitudes you're going to aggregate and set the weights in the light of that.

Either way is permitted by linear pooling, which says only that any aggregate should be a weighted average of the attitudes to be aggregated; it does not specify that the same weights must be used regardless of the attitudes to be aggregated. Thus, for instance, we might weight an individual by their average distance to the other individuals in the group, so that we assign lower weight to outliers and higher weight to those who belong to clusters of individuals who agree or nearly agree on a lot. Or, in the credal case at least, we might determine an individual's weight by how opinionated they are, for instance, using Shannon's measure of entropy to measure the level of uncertainty present in their credence function. Since the aggregated attitude towards a given item, on a weighted average method, depends only on the individual attitudes towards that item and the weights used, if the weights do not depend on the individual attitudes towards other items, nor does the aggregate attitude, thus respecting the Independence of Irrelevant Alternatives. But if the weights do change depending on the individual attitudes to other items, then the aggregate attitude will typically change depending on those attitudes as well, thus violating the Independence of Irrelevant Alternatives.

### 9.2.2 Aggregating credences I: Independence Preservation

So we have seen that all weighted average methods satisfy Weak Pareto and its cousins; all but those determined by the most extremal weightings satisfy No Dictator; all but those for which the weighting of an individual is determined by that individual's attitudes and their relationships to the attitudes of other individuals satisfy the Independence of Irrelevant Alternatives. We turn now to two features that are often thought to be desirable for credal aggregation, but which are not features of any linear pooling method that assigns at least some weight to more than one individual. I will call such methods—where $\alpha_i, \alpha_j > 0$ for some $1 \leq i \neq j \leq n$—*non-extremal linear pooling methods*.[64]

You might wonder why we consider issues with linear pooling for credences here. After all, we concluded in Chapter 5 and reaffirmed in Chapter 7 that you should not aggregate the credences of your different selves and should instead make decisions based on your current credences. But note that the argument from minimal mutilation that we have given in this chapter for weighted average aggregation does not distinguish between credences and utilities—that is, it tells in favour such aggregations for one

[64] Much of this section is adapted from (Pettigrew, taa).

just in case it does for the other. Thus, any objection to linear pooling for credences is an indirect objection to our argument for weighted average aggregation for utilities.

While linear pooling satisfies a whole range of unanimity preservation principles, such as the Weak Pareto conditions and its cousins, it is often observed that, when it is applied to credences, there is a plausible unanimity preservation principle that it does not satisfy. To formulate this principle, we need to remind ourselves what it means for a probabilistic credence function $P$ to render two propositions $X$ and $Y$ independent. Recall from above:

$$\textit{X and Y are probabilistically independent relative to P} \Leftrightarrow P(X|Y) = P(X)$$

That is, two propositions are probabilistically independent relative to a probability function iff the probability of one doesn't change when you condition on the other. Equivalently:

$$\textit{X and Y are probabilistically independent relative to P} \Leftrightarrow P(XY) = P(X)P(Y)$$

That is, two propositions are probabilistically independent relative to a probability function iff the probability of their conjunction is the probability of one weighted by the probability of the other. Now, as is often observed, if two propositions are probabilistically independent relative to two credence functions $P_1$ and $P_2$, it is most likely that they will not be probabilistically independent relative to a weighted average of $P_1$ and $P_2$. The following theorem, which is in the background in (Laddaga, 1977; Lehrer & Wagner, 1983), establishes this:

**Theorem 9.2.1** *Suppose $P_1$, $P_2$ are credence functions, and $P = \alpha P_1 + (1 - \alpha)P_2$ is a weighted average of them (that is, $0 \leq \alpha \leq 1$). Suppose that $X$ and $Y$ are propositions and further that they are probabilistically independent relative to $P_1$ and $P_2$. If $X$ and $Y$ are also probabilistically independent relative to $P$, then at least one of the following is true:*

(i) $\alpha = 0$ *or* $\alpha = 1$. *That is, $P$ simply is one of $P_1$ or $P_2$.*

(ii) $P_1(X) = P_2(X)$. *That is, $P_1$ and $P_2$ agree on $X$.*

(iii) $P_1(Y) = P_2(Y)$. *That is, $P_1$ and $P_2$ agree on $Y$.*

We provide a proof in the Appendix to this chapter. On the basis of this well-known result, it is often said that there is a sort of judgment such that linear pooling does not preserve unanimity on that sort of judgment (Laddaga, 1977; Lehrer & Wagner, 1983; Wagner, 1984; Genest & Wagner, 1987; Dietrich & List, 2015; Russell et al., 2015). The kind of judgment in question is judgment of independence. According to this objection to linear pooling, an individual judges that two propositions are independent exactly when those propositions are probabilistically independent relative to her credence function. So, since two propositions can be independent relative to each of two different credence functions, but dependent relative to each of the non-extremal weighted averages of those credence functions, linear pooling does not preserve unanimous judgments of independence—two individuals may be unanimous in their judgment that $Y$ is independent of $X$, while at the same time all non-extremal linear pools of their credences judge otherwise. That is, linear pooling violates the unanimity preservation principle that we called Independence Preservation in Chapter 6:

> **Independence Preservation** If all individuals have credences functions on which propositions $A$ and $B$ are probabilistically independent of one another, then, on the group credence function, $A$ and $B$ should be independent of one another.

It seems to me that the mistake in this version of the objection to linear pooling lies in the account that it assumes of judgments of independence. I will argue that it is simply not the case that I judge $X$ and $Y$ to be independent just in case my credence in $X$ remains unchanged when I condition on $Y$: it is possible to judge that $X$ and $Y$ are independent without satisfying this condition; and it is possible to satisfy this condition without judging them independent. Let's see how.

First, suppose I am about to toss to coin. I know that it is either biased heavily in favour of heads or heavily in favour of tails. Indeed, I know that the objective chance of heads on any given toss is either 10% or 90%. And I know that every toss is stochastically independent of every other toss: that is, I know that, for each toss of the coin, the objective chance of heads is unchanged when we condition on any information about other tosses. Suppose further that I think each of the two possible biases as equally likely. I assign each bias a credence of 0.5. Then my credence that the coin will land heads on its second toss should also be 0.5. However, if I consider my credence in that same proposition *under the supposition that the coin landed heads on its first toss*, it is different—it is not 0.5. If the coin lands heads on the first toss, that provides strong evidence that the coin is biased towards heads and not tails—if it is biased towards heads, the evidence that it landed heads on the first toss becomes much more likely than it would if the coin is biased towards tails. And, as my credence that the coin has bias 90% increases, so does my credence that the coin will land heads on the second toss. So, while I know that the tosses of the coin are stochastically independent, the outcome of the first and the second toss are not probabilistically independent relative to my credence function.[65]

Next, we can easily find examples in which two propositions are independent relative to my credence function, but I do not judge them independent. Indeed, there are examples in which I know for certain that they are not independent. Suppose, for instance, that there are just two probability functions that I consider possible chance functions. They agree on the chance they assign to $XY$ and $Y$, and thus they agree on the conditional chance of $X$ given $Y$. Both make $X$ stochastically dependent on $Y$. By the lights of the first, $X$ depends positively on $Y$ — the conditional probability of $X$ given $Y$ exceeds the unconditional probability of $X$. By the lights of the second, $X$ depends negatively on $Y$ — the unconditional probability of $X$ exceeds the conditional probability of $X$ given $Y$; and indeed it does so by the same amount that the conditional probability of $X$ given $Y$ exceeds the probability of $X$ relative to the first possible chance function. Suppose I have equal credence in each of these possible chance hypotheses. Then my credence in $X$ lies halfway between the chances of $X$ assigned by the two possible chance functions. But, by hypothesis, that halfway point is just the

[65] More precisely: There are two possible objective chance functions $ch_1$ and $ch_2$. If we let $H_i$ be the proposition that the coin will land heads on its $i^{\text{th}}$ toss, then the following hold:

- $ch_1(H_i) = 0.1$ and $ch_2(H_i) = 0.9$, for all $i$;
- $ch_1(H_iH_j) = ch_1(H_i)ch_1(H_j)$ and $ch_2(H_iH_j) = ch_2(H_i)ch_2(H_j)$.

And if we let $C_{ch_i}$ be the proposition that $ch_i$ is the objective chance function, then given that I know that either $ch_1$ or $ch_2$ is the objective chance function, I should assign credence 1 to the disjunction of $C_{ch_1}$ and $C_{ch_2}$. That is, $P(C_{ch_1} \vee C_{ch_2}) = 1$. Now, given that $H_1$ and $H_2$ are independent relative to $ch_1$ and $ch_2$, it seems natural to say that I judge $H_1$ and $H_2$ to be independent: I know that they are; and I assign maximal credence to a proposition, $C_{ch_1} \vee C_{ch_2}$, that entails that they are. Now suppose I think it equally likely that the coin has the 0.1 bias or that it has the 0.9 bias. So $P(C_{ch_1}) = 0.5 = P(C_{ch_2})$. Then, by the Principal Principle, my credence in heads on the second toss should be 0.5, for it should be $P(H_2) = P(C_{ch_1})ch_1(H_2) + P(C_{ch_2})ch_2(H_2) = (0.5 \times 0.1) + (0.5 \times 0.9) = 0.5$. But suppose now that I condition on $H_1$, the proposition that the coin lands heads on the first toss. If I were to learn $H_1$, that would give me strong evidence that the coin is biased towards heads and not tails. After all, the second chance hypothesis, $C_{ch_2}$, makes heads much more likely than does the first chance hypothesis, $C_{ch_1}$. And, indeed, again by the Principal Principle, $P(H_2|H_1) = \frac{P(H_2H_1)}{P(H_1)} = \frac{P(C_{ch_1})ch_1(H_2H_1)+c(C_{ch_2})ch_2(H_2H_1)}{P(C_{ch_1})ch_1(H_2)+P(C_{ch_2})ch_2(H_2)} = \frac{(0.5\times0.1^2)+(0.5\times0.9^2)}{(0.5\times0.1)+(0.5\times0.9)} = 0.82 > 0.5 = P(H_2)$. So, while I know that $H_1$ and $H_2$ are independent, and judge them so, it does not follow that they are independent relative to my credence function. The upshot: an individual might judge two propositions independent without those two events being probabilistically independent relative to her credence function.

conditional chance of $X$ given $Y$, on which they both agree. So my conditional credence in $X$ given $Y$ is just my unconditional credence in $X$. So $X$ and $Y$ are probabilistically independent relative to my credence function. Yet clearly I do not judge them stochastically independent. Indeed, I know them to be stochastically dependent—what I don't know is whether the dependence is positive or negative.[66]

So it seems that, whatever is encoded by the facts that make $X$ and $Y$ probabilistically independent relative to my credence function, it is not my judgment that those two propositions are stochastically independent: I can know that $X$ and $Y$ are stochastically independent without my credence function rendering them probabilistically independent; and I can know that $X$ and $Y$ are stochastically dependent while my credence function renders them probabilistically independent. Perhaps, then, there is some other sort of independence that we judge to hold of $X$ and $Y$ whenever our credence function renders those two propositions probabilistically independent? Perhaps, for instance, such a fact about our credence function encodes our judgment that $X$ and $Y$ are *evidentially independent* or *evidentially irrelevant*? I think not. If you think that there are facts of the matter about evidential relevance, then these are presumably facts about which an individual may be uncertain. But then we are in the same position as we are with stochastic independence. We might have an individual who is uncertain which of two probability functions encodes the facts about evidential relevance. Each of them might make $Y$ epistemically relevant to $X$; but it might be that, because of that individual's credences in the two possibilities, her credence

[66] More precisely, suppose:

(i) $ch_1(XY) = ch_2(XY)$ and $ch_1(Y) = ch_2(Y)$

(ii) $ch_1(X|Y) - ch_1(X) = ch_2(X) - ch_2(X|Y) > 0$

(iii) $P(C_{ch_1}) = \frac{1}{2} = P(C_{ch_2})$

First, note that:

$$P(X|Y) = \frac{P(XY)}{P(Y)} = \frac{\frac{1}{2}ch_1(XY) + \frac{1}{2}ch_2(XY)}{\frac{1}{2}ch_1(Y) + \frac{1}{2}ch_2(Y)} = \frac{ch_i(XY)}{ch_i(Y)} = ch_i(X|Y)$$

Next, if we let $\beta = ch_1(X|Y) - ch_1(X) = ch_2(X) - ch_2(X|Y)$, then

$$\begin{aligned} P(X) &= \frac{1}{2}ch_1(X) + \frac{1}{2}ch_2(X) \\ &= \frac{1}{2}(ch_1(H|E) - \beta) + \frac{1}{2}(ch_2(H|E) + \beta) \\ &= ch_i(X|Y) = P(X|Y) \end{aligned}$$

function renders $X$ and $Y$ independent. If, on the other hand, you do not think there are facts of the matter about evidential relevance, it isn't clear how facts about my credence function could encode judgments about evidential relevance; nor, if they could, why we should care to preserve those judgments, even when they are made unanimously. Remember: we noted in section 6.2 that there will always be some features shared by all members of a group that cannot be shared with the group credence function.

So I think we shouldn't *require* our aggregation methods for credences to preserve probabilistic independence from individuals to aggregate. That is, I don't think we should accept Independence Preservation. As we will see below, there is a further objection in this vicinity, due to Elkin & Wheeler (2016), that it should at least be *permissible* to preserve independence—something that linear pooling cannot do, as Theorem 9.2.1 shows. But our response to that requires some ideas that arise most naturally from considering a different objection. So we turn to that next; and we return to Elkin and Wheeler's objection in Section 9.2.4.

### 9.2.3 Aggregating credences II: commuting with conditionalization

As is often pointed out, linear pooling does not commute with updating by Bayesian conditionalization (Madansky, 1964; Genest, 1984; Dietrich & List, 2015; Berntson & Isaacs, 2013; Russell et al., 2015). The idea is this: Suppose that two individuals, Adila and Benicio, have credences in a range of propositions; and we take their group credence to be the linear pool of those credences determined by the weighting $\alpha$ for Adila and $1 - \alpha$ for Benicio. At this point, some new evidence arrives that is available to both members of the group. It comes in the form of a proposition that they both learn with certainty. Bayesian conditionalization says that each individual, upon learning this evidence, should update their credences so that their new unconditional credence in a given proposition is just their old conditional credence in that proposition given the piece of evidence. How are we to update group credences in response to such evidence? There are two ways we might proceed: we might look to the individuals first, update their prior credence functions in accordance with the dictates of Bayesian conditionalization, and then take a linear pool of the resulting updated credence functions; or we might look to the group first, and update the group credence function in accordance with Bayesian conditionalization. Now suppose that, in the first approach, the weights used to pool the individuals' posterior updated credence functions to give the group's posterior updated credence function are the same as the weights used to pool the individual's prior credence functions to give the group's prior credence function—that is, Adila's updated credence function is given weight $\alpha$ and Benicio's is given $1 - \alpha$. Then, in that situation, the two methods will rarely give the same result: updating and then pooling will most likely give a different result from pooling and then updating; or, as it is often put, pooling and updating do not commute.[67] The following theorem makes this precise:[68]

**Theorem 9.2.2 ((Madansky, 1964))** *Suppose $P_1$, $P_2$ are credence functions, and $P = \alpha P_1 + (1-\alpha)P_2$ is a weighted average of them (that is, $0 \leq \alpha \leq 1$). And suppose that*

$$\alpha P_1(X|Y) + (1-\alpha)P_2(X|Y) = P(X|Y) \left( = \frac{\alpha P_1(XY) + (1-\alpha)P_2(XY)}{\alpha P_1(Y) + (1-\alpha)P_2(Y)} \right)$$

*Then at least one of the following is true:*

(i) $\alpha = 0$ *or* $\alpha = 1$. *That is, $P$ simply is one of $P_1$ or $P_2$.*

(ii) $P_1(X|Y) = P_2(X|Y)$. *That is, $P_1$ and $P_2$ agree on $X$ given $Y$.*

(iii) $P_1(Y) = P_2(Y)$. *That is, $P_1$ and $P_2$ agree on $Y$.*

This raises a problem for linear pooling, for it shows that the following are usually incompatible:

(1) The rational update rule for individual credences is Bayesian conditionalization.

(2) The rational update rule for group credences is Bayesian conditionalization.

(3) Group credences are always obtained from individual credences in accordance with Linear Pooling.

(4) The weights assigned to individuals do not change when those individuals receive a new piece of evidence.

[67] Mathematicians say that two operations commute if you arrive at the same result regardless of the order in which you apply them. For instance, the operation of rotating an object clockwise by 30° commutes with rotating it counterclockwise 40° since you arrive at the same point regardless of whether you do the former first, then the latter, or the latter first, then the former.

[68] Theorem 9.2.2 is a straightforward corollary of Theorem 9.2.3 below. We simply look at that consequences of letting $\alpha = \alpha'$. And we can see that, in this situation, either (i), (ii), or (iii) must hold.

The argument based on the principle of minimal mutilation from above seeks to establish (3), so we will not question that. What's more, there are strong arguments in favour of Bayesian conditionalization as well (Lewis, 1999; Greaves & Wallace, 2006; Briggs & Pettigrew, 2018). So we have (1) and (2).[69]

That leaves (4). In fact, denying (4) seems exactly right to me. To see why, let's begin by noting exactly how the weights must change to accommodate Bayesian conditionalization as the update plan for group credences in the presence of Linear Pooling. First, let's state the theorem, which is a particular case of the general result due to Howard Raiffa (1968, Chapter 8, Section 11):

**Theorem 9.2.3 ((Raiffa, 1968))** *Suppose $P_1$, $P_2$ are credence functions, and $0 \leq \alpha, \alpha' \leq 1$. And suppose that*

$$\alpha' P_1(X|Y) + (1-\alpha')P_2(X|Y) = \frac{\alpha P_1(XY) + (1-\alpha)P_2(XY)}{\alpha P_1(Y) + (1-\alpha)P_2(Y)}$$

*Then at least one of the following is true:*

(i)
$$\alpha' = \alpha \times P_1(Y) \times \frac{1}{\alpha P_1(Y) + (1-\alpha)P_2(Y)}$$

*and*
$$1-\alpha' = (1-\alpha) \times P_2(Y) \times \frac{1}{\alpha P_1(Y) + (1-\alpha)P_2(Y)}$$

(ii) $P_1(X|Y) = P_2(X|Y)$. *In this case, there are no restrictions on $\alpha'$.*

That is, to obtain the new weight, $\alpha'$, for the first individual (whose initial credence function is $P_1$), we take the old weight, $\alpha$, we weight that by the credence that the first individual initially assigned to $Y$, and we multiply by a normalizing factor. To obtain the new weight, $1-\alpha'$, for the second individual (whose initial credence function is $P_2$), we take the old weight, $1-\alpha$, we weight that by the credence that the second individual initially

[69] Hannes Leitgeb (2016) accepts (3) and (4), but rejects (1) and (2). Leitgeb notes that there is an alternative updating rule that does commute with linear pooling—indeed, it is the only one that does. This alternative updating rule is the extremal case of what Hannes Leitgeb and I called *Alternative Jeffrey Conditionalization*, and which has since become known as *Leitgeb-Pettigrew* or *LP Conditionalization* (Leitgeb & Pettigrew, 2010). Ben Levinstein (2012) raises worries about this updating rule, and I have raised objections to the argument that Leitgeb and I gave in its favour (Pettigrew, 2016a, Section 15.1).

assigned to $Y$, and we multiply by the same normalizing factor. That is, the new weight that is assigned to an individual is proportional to her old weight and the accuracy of her initial credence in the proposition that she has now learned to be true. And indeed that seems exactly right. For we might think of these weights as encoding some facts about the expertise or reliability of the individuals in the group. Thus, when we learn a proposition, we increase the relative weighting of an individual in proportion to how confident they were in that proposition initially—that is, we reward their reliability with respect to this proposition by assigning them greater weight in the future.

Julia Staffel (2015, Section 6) objects to linear pooling on the grounds that it can only accommodate Bayesian conditionalization as the updating rule for individuals and groups by changing the weights assigned to the individuals in this way.[70] Her worry is that, in certain cases, the required shifts in the weights are simply far more extreme than is warranted by the situation. Consider two polling experts, Nate and Ann. Over the course of their careers, they've been equally accurate in their predictions. As a result, when I ask for their group credence—the credence of Nate-Ann—I assign them equal weight: they both get a weight of 0.5. But then Ann has credence 0.8 in $X$ and Nate has credence 0.2 in $X$, and $X$ turns out to be true. When they both learn $X$, we have to shift the weights assigned to them in the group credence in order to preserve conditionalization—we have to shift Nate's from 0.5 to 0.2; and we have to shift Ann's from 0.5 to 0.8. That is, despite his long career of matching Ann's accuracy, one inaccurate prediction results in a drastic shift in the weight that Nate receives. Surely such an extreme shift is not justified by the situation. For instance, if Nate is now sceptical about a second proposition, $Y$, assigning it 0.1, while Ann is bullish, assigning it 0.9, then the group credence will be 0.74—Nate's scepticism will do little to temper Ann's confidence.

I agree that such shifts are counterintuitive. However, I don't agree that this is a reason to reject Linear Pooling. After all, such shifts also occur in credences about chance hypotheses for any individual who satisfies the Principal Principle, a central tenet of Bayesian reasoning. Suppose I am in possession of a trick coin. You know that the bias of the coin towards heads is either 20% or 80%. You've watched 1,000 coin tosses: 500 came up heads; 500 tails. You began with credence 0.5 in each of the bias hypotheses. And you satisfy the Principal Principle at all times. This entails that, at

[70] Thanks to Liam Kofi Bright, Julia Staffel, and Brian Weatherson for urging me to address this objection.

each moment, your credence function is a linear pool of the possible chance functions, where the weight that you assign to a particular possible chance function is just your credence that it is the true chance function. As a result, having witnessed an equal number of heads and tails, your current credence in each of the bias hypotheses has returned to 0.5. But now you toss the coin again, and it lands heads. Then the Principal Principle and Bayesian conditionalization demand that your credence that the bias is 80% must shift to 0.8; and your credence in the bias is 20% must shift to 0.2. So, after a long run of equally good predictions, a single coin toss can shift your credences in the bias hypotheses dramatically. In fact, that single coin toss can shift your credences in the bias hypotheses exactly as dramatically as the weights assigned to individuals might shift if you adhere to Linear Pooling. And this is just a consequence of satisfying the innocuous and widely-accepted Principal Principle.[71] This is my response to Staffel's objection.

### 9.2.4 Aggregating credences III: independence preservation revisited

In Section 9.2.2, we considered an objection to linear pooling from the fact that no non-extremal version of it satisfies Independence Preservation, the

[71] More precisely: There are two possible objective chance functions $ch_1$ and $ch_2$. If we let $H_i$ be the proposition that the coin will land heads on its $i^{\text{th}}$ toss, then the following hold:

- $ch_1(H_i) = 0.2$ and $ch_2(H_i) = 0.8$, for all $i$;
- $ch_1(H_iH_j) = ch_1(H_i)ch_1(H_j)$ and $ch_2(H_iH_j) = ch_2(H_i)ch_1(H_j)$

Let $C_{ch_k}$ be the proposition that $ch_l$ is the objective chance function. And let $P_i$ be my credence function after the $i^{\text{th}}$ toss. Thus, by hypothesis, $P_0(C_{ch_1}) = cr_0(C_{ch_2}) = 0.5$. Also, I assume that $P_i$ satisfies the Principal Principle at all times: that is,

$$cr_i(-|C_{ch_k}) = ch_k(-)$$

One consequence of this is:

$$cr_i(-) = cr_i(C_{ch_1})ch_1(-) + cr_iC_{ch_2})ch_2(-)$$

Thus, my credence function at any point is a linear pool of the possible objective chance functions $ch_1$ and $ch_2$, where the weights are determined by my credences in the chance hypotheses $C_{ch_1}$ and $C_{ch_2}$. Now, after witnessing 500 heads and 500 tails, my credences are thus: $P_{1,000}(C_{ch_1}) = cr_{1,000}(C_{ch_2}) = 0.5$. Now suppose I learn that the $1,001^{\text{st}}$ toss landed heads—that is, I learn $H_{1,001}$. Then

$$cr_{1,001}(C_{ch_1}) = cr_{1,000}(C_{ch_1}|H_{1,001}) = cr_{1,000}(H_{1,001}|C_{ch_1})\frac{cr_{1,000}(C_{ch_1})}{cr_{1,000}(H_{1,001})} = ch_1(H_{1,001})\frac{0.5}{0.5} = ch_1(H_{1,001}) = 0.2$$

And similarly, $P_{1,001}(C_{ch_2}) = 0.8$.

unanimity preservation principle that demands that, whenever two propositions are probabilistically independent relative to all individuals' credences, they should be independent relative to the aggregate credences as well. In this section, we consider a further objection, due to Elkin & Wheeler (2016). They accept that Independence Preservation is false: it is not *always mandatory* to preserve unanimous judgments of independence. However, they do wish to say that is it *sometimes permissible* to do so. That is, they claim that there are groups of individuals, each of whose members judge two propositions independent, for which we wish that it be at least permissible for the aggregate also to judge them independent. As Theorem 9.2.1 shows, if it is always mandatory to aggregate by linear pooling, there will be very few situations in which it is permissible to preserve judgments of independence without using extremal weightings—indeed, it will only be permissible when the individuals all assign the same credence to one or other of the propositions they deem independent of one another.

Let's explore this using an example. Let $R$ be the proposition *It will rain in Bristol tomorrow*, and let $L$ be the proposition *The Pope is left-handed*. Now suppose Adila and Benicio have credences in both of these propositions, and suppose they both consider them independent:

| | $R$ | $L$ | $R\|L$ | $L\|R$ |
|---|---|---|---|---|
| $P_A$ | 0.8 | 0.2 | 0.8 | 0.2 |
| $P_B$ | 0.4 | 0.1 | 0.4 | 0.1 |

Now, Theorem 9.2.1 shows that, if our aggregate is required to be a non-extremal linear pool of Adila's and Benicio's credences—if the aggregate credence function is $P(-) = \alpha_A P_A(-) + \alpha_B P_B(-)$, where $0 \leq \alpha_A, \alpha_B \leq 1$—then it cannot make $R$ and $L$ independent. That is, non-extremal linear pooling requires of the aggregate that, if the group learns the handedness of the Pope, the aggregate will change its opinion about the weather (and vice versa). And, on the face of it, this seems an unreasonable demand. Elkin and Wheeler consider this grounds to abandon linear pooling.

I agree that, on the face of it, this seems an unreasonable demand. But when we consider why linear pooling has this effect, I think our initial reaction should soften. Indeed, I think linear pooling aggregates the opinions of individuals who share independence judgments in exactly the right way. Suppose, for instance, that we aggregate Adila and Benicio using non-extremal linear pooling. For definiteness, let's suppose they both receive the same weight, so that $\alpha_A = \alpha_B = \frac{1}{2}$. So their aggregate credences are:

| | $R$ | $L$ | $R\|L$ | $L\|R$ |
|---|---|---|---|---|
| $P$ | 0.6 | 0.15 | 0.666 | 0.1666 |

So, in line with Theorem 9.2.1, the aggregate credences do not treat $R$ and $L$ as independent. If the group learns that the Pope is left-handed ($L$), its credence in rain ($R$) rises; if it learns that it will rain tomorrow ($R$), its credence that the Pope is left-handed ($L$) increases. This is the source of Elkin and Wheeler's objection.

Now, suppose that the group does learn $L$. Then, updating on this evidence, neither Adila nor Benicio change their credence in $R$. But the aggregate does. However, as Theorem 9.2.3 shows, the updated aggregate credence function is still a linear pool of the updated individual credence functions. What has happened is that the weights have changed. And, as Theorem 9.2.3 also shows, the weights have changed in a principled way. Prior to learning $L$, Adila had a higher credence in that proposition than Benicio did. And thus, once it is revealed as true, her greater accuracy is rewarded with greater weight in the future. Thus, while the prior weights were $\alpha_A = \alpha_B = \frac{1}{2}$, the posterior weights are

$$\alpha'_A = \frac{\alpha_A P_A(L)}{\alpha_A P_A(L) + \alpha_B P_B(L)} = \frac{0.5 \times 0.2}{0.5 \times 0.2 + 0.5 \times 0.1} = \frac{2}{3}$$

and

$$\alpha'_B = \frac{\alpha_B P_B(L)}{\alpha_A P_A(L) + \alpha_B P_B(L)} = \frac{0.5 \times 0.1}{0.5 \times 0.2 + 0.5 \times 0.1} = \frac{1}{3}$$

So, when a non-extremal linear pooling method fails to preserve independence, it does so because, when the group learns a new proposition, and when some of the group have different credences in that proposition, updating the aggregate, linearly pooled credences is equivalent to updating the weights applied to the individuals in such a way that you reward the individuals whose prior credences in the learned proposition were more accurate, and then pooling the individual, unchanged credences using those new weights. And this seems to me to provide a principled reason to reject even the permissibility of preserving independence in these situations. When our aggregate preserves judgments of independence, the credence that our group assigns to $R$ does not change when it learns $L$. But, while all in the group agree that $L$ does not provide information directly about $R$, it does provide information about the accuracy of the individuals in the group, that is, the individuals whose credences we are aggregating. And this should lead us to update the weights these individuals receive, and in turn the aggregate credence in $R$. So, I submit, linear pooling is right to outlaw independence preservation when the individuals have different credences in the propositions they consider independent.

In sum: when we aggregate the numerically-represented judgments of a group of individuals, we should do so by linear pooling. We have argued for this by appealing first to a principle of minimal mutilation—only if we aggregate by linear pooling can we ensure that our aggregates are not needlessly different from the individuals whose credences we are aggregating. But, as we noted, this does not suffice. We must also consider how linear pooling fares when we look also to other desirable features. As we saw, it performs well when we consider the relevant versions of the Arrow conditions, but it does poorly when we consider the preservation of independence judgments, and it seem to perform poorly when we consider its interaction with conditionalization. However, as I argued, it is in fact right that our aggregation method does not preserve judgments of independence, and linear pooling in fact interacts well with conditionalization when the interaction is properly understood.

## 9.3 Appendix: Proofs of Theorems 9.2.1 and 9.2.3

### 9.3.1 Proof of Theorem 9.2.1

**Theorem 9.2.1** *Suppose $P_1$, $P_2$ are credence functions, and $P = \alpha P_1 + (1-\alpha)P_2$ is a weighted average of them (that is, $0 \leq \alpha \leq 1$). Suppose that $X$ and $Y$ are propositions and further that they are probabilistically independent relative to $P_1$ and $P_2$. If $X$ and $Y$ are also probabilistically independent relative to $P$, then at least one of the following is true:*

(i) $\alpha = 0$ *or* $\alpha = 1$. *That is, $P$ simply is one of $P_1$ or $P_2$.*

(ii) $P_1(X) = P_2(X)$. *That is, $P_1$ and $P_2$ agree on $X$.*

(iii) $P_1(Y) = P_2(Y)$. *That is, $P_1$ and $P_2$ agree on $Y$.*

*Proof.* Suppose $P_1(X|Y) = P_1(X)$, $P_2(X|Y) = P_2(X)$, and $P(-) = \alpha P_1(-) + (1-\alpha)P_2(-)$. And suppose

$$\begin{aligned} P(X|Y) &= \frac{P(XY)}{P(Y)} \\ &= \frac{\alpha P_1(XY) + (1-\alpha)P_2(XY)}{\alpha P_1(Y) + (1-\alpha)P_2(Y)} \\ &= \alpha P_1(X) + (1-\alpha)P_2(X) \\ &= P(X) \end{aligned}$$

Then

$$\begin{aligned}&\alpha P_1(X)P_1(Y) + (1-\alpha)P_2(X)P_2(Y) = \\ &\qquad \alpha^2 P_1(X)P_1(Y) + \alpha(1-\alpha)P_2(X)P_1(Y) + \\ &\qquad\qquad \alpha(1-\alpha)P_1(X)P_2(Y) + (1-\alpha)^2 P_2(X)P_2(Y)\end{aligned}$$

So

$$\begin{aligned}&\alpha(1-\alpha)P_1(X)P_1(Y) + \alpha(1-\alpha)P_2(X)P_2(Y) = \\ &\qquad \alpha(1-\alpha)P_2(X)P_1(Y) + \alpha(1-\alpha)P_1(X)P_2(Y)\end{aligned}$$

Suppose (i) is false – that is, $0 < \alpha < 1$. Then

$$P_1(X)P_1(Y) + P_2(X)P_2(Y) = P_2(X)P_1(Y) + P_1(X)P_2(Y)$$

And so

$$(P_1(X) - P_2(X))(P_1(Y) - P_2(Y)) = 0$$

From which it follows that either (ii) $P_1(X) = P_2(X)$ or (iii) $P_1(Y) = P_2(Y)$. □

### 9.3.2 Proof of Theorem 9.2.3

**Theorem 9.2.3** *Suppose $P_1$, $P_2$ are credence functions, and $0 \leq \alpha, \alpha' \leq 1$. And suppose that*

$$\alpha' P_1(X|Y) + (1-\alpha')P_2(X|Y) = \frac{\alpha P_1(XY) + (1-\alpha)P_2(XY)}{\alpha P_1(Y) + (1-\alpha)P_2(Y)}$$

*Then at least one of the following is true:*

(i)
$$\alpha' = \alpha \times P_1(Y) \times \frac{1}{\alpha P_1(Y) + (1-\alpha)P_2(Y)}$$

*and*
$$1 - \alpha' = (1-\alpha) \times P_2(Y) \times \frac{1}{\alpha P_1(Y) + (1-\alpha)P_2(Y)}$$

(ii) $P_1(X|Y) = P_2(X|Y)$. *In this case, there are no restrictions on $\alpha'$.*

*Proof.* Suppose

$$\alpha' P_1(X|Y) + (1-\alpha')P_2(X|Y) = \frac{\alpha P_1(XY) + (1-\alpha)P_2(XY)}{\alpha P_1(Y) + (1-\alpha)P_2(Y)}$$

Then, multiplying both sides by the denominator of the right-hand side gives:

$$\begin{aligned}&\alpha P_1(X|Y)P_1(Y) + (1-\alpha)P_2(X|Y)P_2(Y) = \\ &\qquad \alpha\alpha' P_1(X|Y)P_1(Y) + \alpha(1-\alpha')P_2(X|Y)P_1(Y) + \\ &\qquad\qquad \alpha'(1-\alpha)P_1(X|Y)P_2(Y) + (1-\alpha)(1-\alpha')P_2(X|Y)P_2(Y)\end{aligned}$$

So

$$\alpha'(1-\alpha)P_1(Y)(P_1(X|Y) - P_2(X|Y)) = \alpha(1-\alpha')P_2(Y)(P_1(X|Y) - P_2(X|Y))$$

Now, suppose (ii) is false – that is, $P_1(X|Y) \neq P_2(X|Y)$. Then

$$\alpha'(1-\alpha)P_1(Y) = \alpha(1-\alpha')P_2(Y)$$

And thus,

$$\alpha' = \frac{\alpha P_1(Y)}{\alpha P_1(Y) + (1-\alpha)P_2(Y)}$$

That is, (i), as required. □

# Chapter 10

# Do we know enough to make decisions this way?

I was inspired to write this book partly by reading Edna Ullmann-Margalit's paper 'Big Decisions: Opting, Converting, and Drifting' (Ullmann-Margalit, 2006), and partly by reading L. A. Paul's enormously influential book, *Transformative Experience* (Paul, 2014a).[72] Ullmann-Margalit and Paul both discuss a particular version of the problem of choosing for changing selves, namely, that which arises when one of the choices that is open to us might actively cause my values to change—choosing to adopt a child, for instance, or to embark on a new career, or to move to another country; these are all examples of such a choice, and thus Aneri's choice to become a police officer, Cheragh's decision to write her novel, and Deborah's dilemma whether to have a baby now or later all fall in this category, while Blandine's decision to study particle physics, and Erik's and Fernando's pension scheme choices do not. But Paul also raises what she takes to be a further problem for orthodox decision theory. As we will see, the natural solution to that problem shares features with our favoured solution to the problem of choosing for changing selves, that is, the Aggregate Utility Solution. What's more, Paul thinks there is a fundamental problem with those features, and if she is right, her objection causes problems for the Aggregate Utility Solution as well.

[72] Material in this chapter is adapted from (Pettigrew, 2015b, 2016b, tab).

## 10.1 The deliberative conception of decision theory

Before we describe the problem that Paul raises, it will be useful to specify the version of decision theory to which she takes it to apply. Firstly, she takes it to apply primarily to the realist conception that we introduced in Chapter 2.[73] But secondly, she takes it to apply to what I will call the *deliberative* understanding of decision theory, as opposed to the *evaluative* (though she sometimes hints that the objection is intended to apply to the evaluative interpretation as well). The deliberative and evaluative understandings of decision theory differ on which elements of a decision are relevant to its rationality. For those who favour a deliberative understanding, decision theory governs not only the choice that an individual makes in a given situation, but also the deliberation by which she comes to make that choice. In contrast, those who favour an evaluative understanding say that decision theory evaluates the choice only. Thus, for instance, suppose I must decide whether or not to take an umbrella when I leave my house. As it happens, I would maximise my expected utility by taking the umbrella—I think it's pretty likely to rain, I hate getting wet, and it doesn't much bother me to carry the umbrella. Now suppose that I do indeed end up taking the umbrella. But my reason for doing so was not that it would maximise my expected utility—it was not by calculating which action would maximise expected utility and then picking it that I reasoned to my conclusion. Rather, I chose the action I did simply using the rule *Always pick the action that involves approximating most closely the sartorial choices of Mary Poppins*. Then, according to the evaluative understanding of decision theory, I am fully rational, because I chose the option that maximises expected utility, while according to the deliberative understanding, I am not, because I did not deliberate correctly concerning my choice—my decision was not sensitive to the expected utility of the actions between which I had to choose. As we will see below, Paul's challenge applies primarily to the deliberative understanding of decision theory, though I will also ask whether Paul's insight supports a stronger argument, which tells against the evaluative understanding as well.

[73] Recall: for both realist and constructivist, credences, utilities, and preferences are all psychologically real; but for the realist, credences and utilities are fundamental and determine preferences, while for the constructivist, preferences are fundamental and determine credences and utilities.

## 10.2 Paul's Utility Ignorance Objection

Paul's first objection to decision theory is that it cannot accommodate choosing for changing selves—or, in her terminology, how to make a decision when one of the options might lead to what she calls a *personally transformative experience*, that is, an experience that will lead you to change your values. It is the purpose of this book to explore how we might answer that objection. But her second objection to decision theory is based on the possibility of a different sort of transformative experience, which she calls an *epistemically transformative experience* (or *ETE*). This is an experience that teaches you something that you couldn't come to know without having that experience. Thus, for Frank Jackson's scientist, Mary, who has lived her whole life in a monochrome black-and-white room, the experience of stepping outside and seeing the colour red for the first time is an ETE. However much Mary learned about the physical properties of red objects during her time in that room, she could not know what it is like to see red (Jackson, 1986). Similarly, for some people, becoming a parent for the first time is an ETE. However much they attend to the testimony of people who already have children, however much they read novels about parenting, however much they care for their friends' children or their nephews and nieces, they cannot know what it is going to be like to be a parent until they become one themselves (Paul, 2014a).

In Mary's case, what she learns from her ETE is a phenomenological fact—she learns what it is like to see red. In the case of the new parent, there is likely a phenomenological component to what they learn from the experience as well—they learn what it is like to feel a particular sort of bond with another person; and they might learn for the first time what it is like to have sustained responsibility, either solely or in partnership with others, for another life. But there may well be other components—the experience might teach you some moral facts, for instance. For Paul's objection, she needs only this: ETEs teach you something that you cannot learn any other way and that you need to know in order to know the utility that you assign to the outcomes of certain actions that are available to you.

For instance, suppose I must decide whether or not to apply to adopt a child and become a parent. If I choose to apply and my application is successful, I will become a parent. In order to calculate the expected value of choosing to apply, I must therefore know the utility I assign to the outcome on which I apply and my application is successful. But in order to know that, I need to know what it will be like to be a parent—the phenomenal experience of the parental bond and the phenomenal experience of bearing sustained responsibility for this particular life are components that will at least partly determine my utility for being a parent. And that, for some people, is something that they can know only once they become a parent. For such people, then, it seems that the ingredients that they require in order to calculate their expected utility for applying to adopt a child are not epistemically available to them. And thus they are barred from deliberating in the way that the realist-deliberative understanding of decision theory requires of them. They are unable to make the decision rationally.

Using the ingredients of orthodox expected utility theory—as introduced above in Chapter 2—we can state the problem as follows: there are two actions between which I must choose—apply to adopt a child (*Apply*); don't apply (*Don't Apply*). And let's say that there are two states of the world—one in which I become a parent (*Parent*) and one in which I don't (*Child-free*). To choose whether or not to apply, I must determine whether I prefer applying to not applying—that is, whether $\mathit{Apply} \prec \mathit{Don't\ Apply}$ or $\mathit{Apply} \sim \mathit{Don't\ Apply}$ or $\mathit{Apply} \succ \mathit{Don't\ Apply}$. And, for the realist, in order to determine that, I must calculate the expected utility of those two actions relative to my credence function and my utility function. And to calculate that, I must know what my credence is in each of the two possible states of the world given each of the two possible actions—that is, I must know $P(\mathit{Parent}||\mathit{Apply})$, $P(\mathit{Child\text{-}free}||\mathit{Apply})$, and so on. And I must know my utilities for the different possible outcomes—that is, I must know $U(\mathit{Apply}\ \&\ \mathit{Parent})$ and $U(\mathit{Apply}\ \&\ \mathit{Child\text{-}free})$ and $U(\mathit{Don't\ Apply}\ \&\ \mathit{Parent})$ and $U(\mathit{Don't\ Apply}\ \&\ \mathit{Child\text{-}free})$ (we ignore for the moment the possibility that my utilities might change if I adopt). The problem that Paul identifies is that it is impossible to know $U(\mathit{Apply}\ \&\ \mathit{Parent})$ prior to making the decision and becoming a parent; and thus it is impossible to deliberate about the decision in the way that the realist-deliberative understanding of decision theory requires.[74] Paul concludes that there is no rational way to make such decisions. This is the Utility Ignorance Objection to the realist-deliberative understanding of decision theory.

Before we move on to consider how we might respond to this objection, let us pause a moment to consider its scope. First, note that the challenge targets only the realist understanding of decision theory, not the constructivist. For the constructivist, my credence and utility functions are determined by my preference ordering. Thus, to know them I need only know my pref-

[74] In fact, you might go further and say that applying and failing in your application is also an epistemically transformative experience, given your emotional investment in the application. If that's the case, it is impossible to know $U(\mathit{Apply}\ \&\ \mathit{Child\text{-}free})$, just as it's impossible to know $U(\mathit{Apply}\ \&\ \mathit{Parent})$. This would compound the problem Paul identifies.

erence ordering. And for many constructivists I can know that simply by observing how I choose between given sets of actions—I prefer *a* to *b* just in case I would choose *a* over *b* in a binary choice between them; this is so-called *revealed preference theory* (Samuelson, 1948). Paul's challenge applies only when we take the preference ordering, and thus the set of rationally permissible actions, to be determined at least in part by the utility function, as the realist does. Second, note that the challenge targets only the deliberative understanding of decision theory, not the evaluative. On the realist-evaluative understanding, I do not need to know my credences or my utility function in order to be rational. On this understanding, in order to be rational, I need only choose the action that *in fact* maximises expected utility; I need not choose it *because* it maximises expected utility. Thus, Paul's argument has no bite for the evaluative understanding.

Now, we might try to extend Paul's argument so that it does apply to the realist-evaluative understanding. To do that, we need to argue not only that I do not *know* the utility $U(\mathit{Apply}, \mathit{Succeed})$ prior to my choice between *Apply* and *Don't Apply*, but indeed that $U(\mathit{Apply}, \mathit{Succeed})$ is not even *determined* prior to that choice—and indeed Paul herself hints at that interpretation in some places (Paul, 2015a, 494). If that were the case, then there would be no way to make the choice rationally, even according to the realist-evaluative understanding. I am sceptical that the argument as it stands can support this conclusion. The examples that Paul gives motivate the claim that we cannot know our utilities for certain outcomes. To move from that to the claim that those utilities are not determined requires a particular sort of account of what they are. For instance, if we are hedonists, and the utility we assign to an outcome measures the intensity of the pleasure that we would experience in that outcome, then it is quite possible that those utilities might be determined but unknowable. In order to extend Paul's argument, then, we must at least rule out such a conception of utilities, and that requires further argument.

## 10.3 Paul's Revelation Response

Let's begin with Paul's own response to her own objection—I will call it *the Revelation Response*. Paul does not wish to abandon the machinery of decision theory; she does not wish to reject expected utility theory. Rather, in cases where one of the outcomes involves an ETE, she wishes to reconfigure the decision problem and apply standard decision theory to that reconfigured problem. Here she is talking of the decision whether or not to try the much-loved but also widely-hated durian fruit for the first time:

> The relevant outcomes, then, of the [reconfigured] decision to have a durian are discovering the taste of durian versus avoiding the discovery of the taste of durian, and the values attached reflect the subjective value of making (or avoiding) this discovery, not whether the experience is enjoyable or unpleasant. (Paul, 2014a, 113)

Thus, Paul's Revelation Response has two parts: first, we reconfigure the decision so that it is a decision between discovering what it's like to eat durian and avoiding discovering what it's like to eat durian, rather than a decision between eating durian and not eating durian; second, when calculating our subjective expected utility for each of these new possible actions, the only factor that is taken to feed into the utilities that we assign to the outcomes is the value we attach to having a novel experience.

The first part of the solution in fact does not change anything, since both of these decisions have the same outcomes. I can only discover what it's like to eat durian by eating it; and by eating it I discover what it's like to eat durian—indeed, this is what makes eating durian an epistemically transformative experience. So the outcomes of choosing to make that discovery are exactly the same as the outcomes of choosing to eat durian, and the decision problem has not been altered: the outcomes are the same and the problems with accessing our utilities for the outcomes remain in place. The second part of Paul's solution is intended to deal with the access problem. My worry about this part of the solution is that it seems to ignore the very problem that Paul has raised. Underpinning Paul's central objection to expected utility theory is the observation that my overall subjective utilities for the possible outcomes of my actions are determined by a number of factors, including, for instance, the moral value of the outcome, its aesthetic value, the extent to which it realizes my long-term goals, and so on. One of these factors will be the extent to which I value the novelty of any new experiences it affords me; and another will be the extent to which I value what it's like to experience that outcome. Now one of Paul's main objections to expected utility theory is that, for many important decisions, there are possible outcomes of our actions such that we cannot know what it's like to experience them. If this is true, and if the nature of this first-personal experience is indeed one of the factors that determines the overall utility of that outcome, we cannot know what our overall utilities are for those outcomes. Paul's solution is that we should ignore this factor, along with all other factors except the extent to which the outcome affords the individual a new experience. The problem with this, however, is that it does not seem to solve the problem that Paul has posed any better than a solution that says, for instance, that we should make such major transformative life decisions by simply ignoring all factors that determine the overall utility of the outcomes except their moral value; or by ignoring all factors except their aesthetic value; or by ignoring only what it's like to have the transformative experience, which is, in any case, the only factor to which we do not have access. So I think the Revelation Response favoured by Paul fails.

## 10.4 The Fine-Graining Response

Next, we turn to the Fine-Graining Response—this is a response to Paul's objection for which I have argued along with a number of other philosophers (Dougherty et al., 2015; Harman, 2015; Pettigrew, 2015b, 2016b). It is the similarity between this response and certain features of our favoured response to the problem of choosing for changing selves that makes Paul's challenge relevant to us in this context.

Expected utility theory is designed to deal with decisions made in the face of uncertainty. Usually that uncertainty concerns the way the world is beyond or outside of the individual. For instance, suppose I'm uncertain whether my adoption application would be successful if I were to apply. Then, when I'm making my decision, I ensure that the set of possible states of the world includes one in which my application succeeds and one in which it fails. I then quantify my uncertainty concerning these two possibilities in my credence function, and I use that to calculate my expected utility—perhaps I know that only 12% of adoption applications succeed, and I set my credence that mine will succeed to 0.12 in line with that, so that $P(Parent||Apply) = 0.12$. However, there is no reason why the uncertainty quantified by my credence function should concern only the way the world is beyond me. What Paul's argument shows is that I am uncertain not only about such worldly matters as whether I would be successful if I were to apply, but also about the utility that I assign to becoming a parent; I am uncertain not only about whether *Parent* or *Child-free* will be true if I apply, but also about the value $U(Apply \,\&\, Parent)$. Thus, just as I ensured that my decision problem includes possible states of the world at which I succeed in my application and possible states where I fail, similarly I should respond to Paul's challenge by ensuring that my decision problem includes possible states of the world at which I become a parent and value it greatly, possible states at which I become a parent and value it a moderate amount, states at which I become a parent and value it very little, and so on. Having done this, I should quantify my uncertainty concerning the utility I assign to being a parent in my credence function, and use that to calculate my expected utility as before.

More precisely, and simplifying greatly, suppose the possible utility values that I might assign to being a parent are $-12$, 3, and 10 (when measured on the same scale as the rest of my utilities). Then, while my original set of possible states of the world is $\mathcal{S} = \{Parent, \textit{Child-free}\}$, my new expanded set of possible states of the world is

$$\begin{aligned}\mathcal{S}^* = \{&\textit{Parent \& utility of being a parent is } -12,\\ &\qquad\textit{Parent \& utility of being a parent is } 3,\\ &\qquad\qquad\textit{Parent \& utility of being a parent is } 10,\\ &\qquad\qquad\qquad\textit{Child-free}\}.\end{aligned}$$

Now, recall the problem that Paul identified. Given the original way of setting up the decision problem, in order to deliberate rationally between *Apply* and *Don't Apply*, I need to know the utilities I assign to each possible outcome of each of the possible actions. In particular, I need to know $U(\textit{Apply} \ \& \ \textit{Parent})$. But I can't know that until I make the decision and become a parent. However, on the new formulation of the decision problem, with the expanded set of states $\mathcal{S}^*$, I do know the utilities I assign to each possible outcome of each of the possible actions. For I know that:

- $U(\textit{Apply} \ \& \ \textit{Parent \& utility of being a parent is } -12) = -12$,
- $U(\textit{Apply} \ \& \ \textit{Parent \& utility of being a parent is } 3) = 3$,
- $U(\textit{Apply} \ \& \ \textit{Parent \& utility of being a parent is } 10) = 10$,

Next, I quantify my uncertainty in these new possible states to give:

$$\begin{aligned}&P(\textit{Parent \& utility of being a parent is } -12 || a),\\ &\qquad P(\textit{Parent \& utility of being a parent is } 3 || a),\\ &\qquad\qquad P(\textit{Parent \& utility of being a parent is } 10 || a),\\ &\qquad\qquad\qquad P(\textit{Fail} || a),\end{aligned}$$

where $a$ is either *Apply* or *Don't Apply*. And, given this, I can calculate my expected utility and discharge the obligations of rationality imposed by the realist-deliberative understanding of decision theory. Paul's Utility Ignorance Objection, it seems, is answered. Call this the Fine-Graining Response, since it involves expanding, or fine-graining, the set of possible states of the world.

Now, notice how the states of the world to which the fine-graining response appeals resemble the states to which I appeal in my favoured response to the problem of choosing for changing selves. In both cases, they specify not only how the world is beyond the individual, but also how things are inside the individual; in particular, their utilities. Thus, if there is a problem for the Fine-Graining Response to Paul's Utility Ignorance Objection, it likely carries over to my favoured solution to the problem of choosing for changing selves. I'll consider two such objections: the first due to Paul herself (section 10.5), the second to Sarah Moss (section 10.6).

## 10.5 Paul's Authenticity Reply

Paul is not satisfied with the Fine-Graining Response. She allows that I can expand the set of possible states of the world in the way described. And she allows that I can form credences in those different states of the world. But she worries about the sort of evidence on which I might base those credences.

Let's start with an ordinary decision that does not involve an ETE. Suppose I am deciding whether to have chocolate ice cream or strawberry ice cream. I have tasted both in the past, so I know what both experiences will be like—neither experience would be transformative. As a result, when I come to make my decision, I know the utility I assign to the outcome in which I eat chocolate ice cream. I know it by imaginatively projecting myself forward into the situation in which I am eating chocolate ice cream. And I can do this because I have tasted chocolate ice cream in the past. And similarly for the utility I assign to the outcome in which I eat strawberry ice cream. I know what it is, and I know it because I've tasted strawberry ice cream in the past and so I can imaginatively project myself forward into the situation in which I'm eating it.

When I consider the utility I assign to becoming a parent, I can't imaginatively project in this way, since I'm not a parent and becoming a parent is an ETE. As described above, I respond to this epistemic barrier by expanding the set of possible states of the world I consider in my decision problem. I expand them so that they are fine-grained enough that each specifies my utility for becoming a parent at that world; and my credences in these different possible states quantify my uncertainty over them. But how do I set those credences? I cannot do anything akin to imaginatively projecting myself into the situation of being a parent, as I did with the chocolate ice cream, because becoming a parent is an ETE. What can I do instead?

Well, the natural thing to do is to seek out the testimony of people who have already undergone that transformative experience.[75] Perhaps I cannot discover from them exactly what it is like to be a parent—since it's an ETE, the only way to learn what it's like is to undergo the experience. But perhaps I can learn from them how much they value the experience. And after all, that's all that I need to know in order to make my decision rationally, according to the realist-deliberative understanding of decision theory—expected utility theory doesn't require that you know what an outcome will be like; it requires only that you know how much you value it and thus how much it contributes to the expected utility calculation. However, as we all know, different people value being a parent differently. For some, it is an experience of greater value than all other experiences they have in their life. For others, it is a positive experience, but doesn't surpass the value of reciprocated romantic love, or extremely close friendships, or succeeding in a career, or helping others. And for yet others, it is a negative experience, one that they would rather not have had. Simplifying greatly once again, let's assume that all parents fall into these three groups: members of the first assign 10 utiles to the outcome in which they become a parent; members of the second assign 3; and members of the third assign -12. And let's assume that 10% fall into the first group; 60% into the second; and 30% into the third. Now, suppose that I learn this statistical fact by attending to the testimony of parents. Then I might set my credences as follows (where we assume for convenience that I am certain that my adoption application will be successful, so $P(\mathit{Parent}||\mathit{Apply}) = 1$):

- $P(\mathit{Parent\ \&\ utility\ of\ being\ a\ parent\ is\ } {-12} \,||\, \mathit{Apply}) = 0.3$,
- $P(\mathit{Parent\ \&\ utility\ of\ being\ a\ parent\ is\ } 3 \,||\, \mathit{Apply}) = 0.6$,
- $P(\mathit{Parent\ \&\ utility\ of\ being\ a\ parent\ is\ } 10 \,||\, \mathit{Apply}) = 0.1$,

With these in hand, I can then calculate the expected utility of *Apply* and *Don't Apply*, I can compare them, and I can make the choice between them in the way that the realist-deliberative decision theorist requires.

However, Paul claims that if I choose in this way then my decision is badly flawed. She holds that an individual who made the decision to become a parent in this way would be "alienated" from that decision; the choice thus made would be "inauthentic":

[75] See (Dougherty et al., 2015) for two further ways in which I might set these credences. I focus on testimonial evidence here since it is the sort of evidence that Paul and Moss both consider.

> A [...] problem with leaving your subjective perspective out of your decisions connects to the Sartrean point that making choices authentically and responsibly requires you to make them from your first personal perspective. A way to put this is that if we eliminate the first personal perspective from our choice, we give up on authentically owning the decision, because we give up on making the decisions for ourselves. We give up our authenticity if we don't take our own reasons, values, and motives into account when we choose. To be forced to give up the first person perspective in order to be rational would mean that we were forced to engage in a form of self-denial in order to be rational individuals. We would face a future determined by Big Data or Big Morality rather than by personal deliberation and authentic choice. (Paul, 2014a, 130)

For Paul, then, the problem lies in the way that I set my credences in the fine-grained states of the world. I set my credences concerning my own utilities by deferring to statistical facts about how others assign different utilities. My evidence does not sufficiently concern *my* utilities; and thus I am alienated from any decision based on the credences that I form in response to that evidence. I am like the individual who makes a moral decision by deferring to societal norms or the value judgements of the majority group, rather than making those decisions herself. Paul contrasts this statistical method of forming opinions about my own utilities with the method described above in the case of the chocolate and strawberry ice cream, where I imaginatively project myself into the situation in which I have the experience based on my own memory of previous similar experiences. In those cases, the opinions formed do not give rise to the same sort of alienation and inauthenticity, since they are connected in the right way to my own utilities. They are more akin to the individual who makes the moral decision for themselves.

I don't think Paul's Authenticity Reply will work. After all, when I set my credences concerning my own utilities by appealing to the statistical evidence concerning the utilities of others, I do so because I think that this statistical evidence tells me something about *my own utility*; it is good evidence concerning *my own utilities*. In contrast, when I defer to societal norms to make a moral decision, I do so not because I think that those norms tell me anything about my own values; I do not think they provide good evidence concerning what I think is the correct moral action. I do so because I can't decide what I think is the correct moral action, or I do not have the courage to follow my own moral compass.

## 10.6 Moss' No Knowledge Reply

Like Paul's Authenticity Reply to the Fine-Graining Response, Moss' No Knowledge Reply also argues that the problem with such decisions lies in the nature of the evidence on the basis of which I form my credences about my utilities, and on the sort of connection between my utilities and the resulting credences. Let's turn to Moss' reply now.

Suppose I set my credences in *Parent & utility of being a parent is* $-12$, etc., as above. That is, I set them on the basis of statistical evidence concerning the utilities that existing parents assign to being a parent. For Paul, the problem is that such evidence does not sufficiently concern my utilities in particular; it is too much concerned with the utilities of other people. For Moss, the problem with those credences is not that they are not sufficiently concerned with me, or at least that is not the primary problem. Rather, the problem is that those credences do not constitute knowledge, and rational decisions must be based on credences that constitute knowledge (Moss, 2018, Section 9.5).

To those unfamiliar with Moss' work, it might sound as if she is making a category mistake. Credences, you might think, are simply not the sort of thing that can constitute knowledge. Full beliefs can—if I believe that it's raining, then that belief might count as knowledge. But credences, or partial beliefs, cannot—if I have credence 0.6 that it's raining, then it makes no more sense to say that my credence counts as knowledge than it does to say that a colourless idea sleeps furiously. Or so you might think. But Moss denies this (Moss, 2013, 2018). Let's see why.

First, it is worth saying what Moss takes credences to be. Suppose I say that I'm 50% confident that Kenny is in Hamburg. On the standard interpretation, this means that I have a precise graded attitude—a credence—towards the standard, non-probabilistic content *Kenny is in Hamburg*, where the latter might be represented by a set of possible worlds. In particular, I have a 0.5 credence in that non-probabilistic content. For Moss, in contrast, a credence is not a graded attitude towards a standard propositional content; rather, it is a categorical attitude towards what she calls a *probabilistic content*. For instance, to say that I'm 50% confident that Kenny is in Hamburg is to say that I have a categorical attitude—in fact, a belief—towards the probabilistic content *Kenny is 50% likely to be in Hamburg*.

What are these probabilistic contents? Well, just as a standard propositional content, such as *Kenny is in Hamburg*, can be represented by a set of possible worlds, so a Mossian probabilistic content, such as *Kenny is 50% likely to be in Hamburg*, is represented by a set of probability spaces, where a probability space is a set of possible worlds together with a probability distribution defined over those worlds. Thus, the probabilistic content *Kenny is 50% likely to be in Hamburg* is represented by the set of those probability spaces in which the probability distribution assigns 50% to the proposition *Kenny is in Hamburg*—that is, the set $\mathbf{P} = \{P : P(\textit{Kenny is in Hamburg}) = 0.5\}$.

Another example: Suppose I say that I'm more confident than not that Kenny is in Hamburg. On the standard interpretation, this means that I have an imprecise graded attitude towards the propositional content *Kenny is in Hamburg* (Joyce, 2010). Imprecise graded attitudes are also represented by sets of probability spaces—these are usually called *representors*. In this case, my imprecise graded attitude is represented by the set of those probability spaces in which the probability distribution assigns more than 50% to the proposition *Kenny is in Hamburg*—that is, the set $\mathbf{P} = \{P : P(\textit{Kenny is in Hamburg}) > 0.5\}$. That set is my representor. For Moss, in contrast, I do not have a graded attitude towards the propositional content *Kenny is in Hamburg*, but rather a categorical attitude towards the probabilistic content *Kenny is more likely than not to be in Hamburg*. The probabilistic content towards which I have that categorical attitude is in fact represented by the same set of probability spaces that is used to represent the imprecise graded attitude that is usually attributed to me—that is, my representor, $\mathbf{P} = \{P : P(\textit{Kenny is in Hamburg}) > 0.5\}$.

Now, citing a large body of examples, Moss argues that we often say that, just as beliefs in standard, non-probabilistic contents—viz., propositions—can count as knowledge, so can beliefs in probabilistic contents—viz., the contents represented by sets of probability functions. For instance, I might say that Patricia knows that Kenny is 50% likely to be in Hamburg, or that Jason knows that Kenny is more likely than not to be in Hamburg.

As well as citing intuitive examples in which we ascribe probabilistic knowledge, Moss also gives examples that show that there are distinctions between categorical beliefs in probabilistic contents that are analogous to the distinctions that we mark between different categorical beliefs in propositions by categorising one as merely justified and the other as knowledge. For instance, suppose that I know that the objective chance of this coin landing hands is 60%. And my credence that it will land heads is 0.6—that is, in Moss' framework, I believe that the coin is 60% likely to land heads. Next, suppose that you also set your credence in heads to 0.6—that is, you also believe the coin is 60% likely to land heads. But you set your credence in this way not because you know the objective chance, but because you know that Sarah's credence in heads is 0.6 and you have good reason to take Sarah to be an expert on the bias of coins. However, while you are right that Sarah is generally expert on such matters, in this case she hasn't actually inspected the coin and instead just plucked a number from thin air. In such a case, it seems that, while both of us have justified credences that are correct in a certain sense, yours is merely justified, while mine counts as knowledge.

Moss furnishes us with a splendidly detailed account of probabilistic knowledge, which includes a Bayesian expressivist semantics for probabilistic knowledge ascriptions as well as an account of the factivity, safety, and sensitivity conditions on probabilistic knowledge. But her No Knowledge Reply to the Fine-Graining Response does not depend on the more sophisticated or radical elements of her account. Rather, it depends on just three claims about probabilistic knowledge.

The first, we have met already: it is the claim that credences—and, more generally, beliefs in probabilistic contents—can count as knowledge, just as beliefs in non-probabilistic contents can.

The second claim concerns a certain sort of case in which the credences you form don't count as knowledge. Suppose we meet. Noting that I am a living human being, and knowing that about 0.7% of living human beings will die in the next year, you form a credence of 0.007 that I will die in the next year. Then, for Moss, your credence does not count as knowledge. The problem is that you cannot rule out relevant alternative reference classes to which I belong and among which the frequency of death within the next year is quite different. For instance, you know that I am 35 years old. And you can't rule out that the likelihood of death among living 35 year olds is quite different from the likelihood among all human beings. You know that I am male. And you can't rule out that the likelihood of death among living males is different from the likelihood among all human beings. And so on. You believe that it's 0.7% likely that I will die in the coming year, but you can't rule out that my death is $X$% likely, for a range of alternative values of $X$. Moss likens the case to Goldman's fake barn scenario (Goldman, 1976). I am travelling through Fake Barn County, and I stop in front of a wooden structure that looks like a barn. I form the belief that the structure in front of me is a barn because that's what it looks like. But my visual experience cannot distinguish a barn from a barn facade. So I cannot rule out the alternative possibility that the structure is a barn facade. And this alternative is relevant because Fake Barn County lives up to its name: it's full of fake barns. Therefore, my belief cannot count as knowledge. Similarly, since you cannot rule certain alternative reference classes among which my likelihood of death within the next year is quite different from 0.7%, your credence of 0.007 that I will die in the next year cannot count as knowledge.

Or so Moss says.

Now, recall our response outlined above to Paul's Utility Ignorance Objection to decision theory. Since I cannot know the utility I assign to being a parent, I expanded the set of possible states of the world so that, in each, my utility is specified; and then I quantified my uncertainty concerning these different utilities in my credences. Since I could not set those credences by imaginatively projecting myself into the position of being a parent, I had to set them by appealing to the statistical evidence concerning the utilities that existing parents assigned to being parents. Since the evidence for my credences is statistical, if it is to count as knowledge, I must be able to rule out relevant alternative reference classes to which I belong on which the statistics are quite different. For instance, suppose I set my credences in the different possible utilities by appealing to the statistics among *all* existing parents. Then there are certainly relevant alternative references classes that I should consider: the class of all parents who are men; the class of all gay parents who are cis men; the class of adoptive parents; the class of all parents with family and social support network similar to mine; and so on. Given the evidence on which I based my credences, I cannot rule out the possibility that the distribution of the three candidate utilities for being a parent is different in these reference classes from the distribution in the reference class on which I based my credences. Thus, according to Moss, my credences cannot count as knowledge.

Finally, the third claim upon which Moss bases her No Knowledge Reply to the Fine-Graining Response is a conjunction of a probabilistic knowledge norm for reasons and a probabilistic knowledge norm for decision—together, we refer to these as the *Probabilistic Knowledge Norms for Action*, following Moss.

> **Probabilistic Knowledge Norm for Reasons** Your credal state can only provide a reason for a particular choice if it counts as knowledge.
>
> **Probabilistic Knowledge Norm for Decisions** Suppose the strongest probabilistic content you know is represented by a set **P** of probability functions; and suppose you are faced with a choice between a range of options. It is permissible for you to chose a particular option iff that option is permissible, according to the correct decision theory for imprecise credences, for an individual whose imprecise credal state is represented by **P**.

For instance, suppose you must choose whether to take an umbrella with you when you leave the house. The strongest proposition you know is represented by the set of probability spaces, $\mathbf{P} = \{c : 0.4 < P(Rain) < 0.9\}$. If rain is 90% likely, then taking the umbrella maximises expected utility; if it is only 40% likely, then leaving the umbrella maximises expected utility. Now imagine an individual whose credal state is represented by **P**—in the language introduced above, **P** is her representor.[76] Which actions are permissible for this individual? According to some decision theories for imprecise credences, an action is permissible iff it maximises expected utility relative to *at least one member of the representor*. We might call these *liberal* decision theories, since they make many actions permissible. On this decision theory, it is permissible to take the umbrella and permissible to leave it. Thus, according to the Probabilistic Knowledge Norm for Decisions, both actions are also permissible. According to other decision theories, an action is permissible iff it maximises expected utility relative to *all members of the representor*. We might call these *conservative* decision theories, since they make few actions permissible. On this decision theory, neither taking nor leaving the umbrella is permissible for the individual with representor **P**, and thus, according to the Probabilistic Knowledge Norm for Decisions, neither is permissible for me.

Thus, putting together the various components of Moss' No Knowledge Reply, we have:

(i) the only precise credences I could form concerning the utility I assign to being a parent do not count as knowledge, because my statistical evidence doesn't allow me to rule out alternative reference classes that are made salient, or relevant, by the high stakes decision I wish to make based on those credences;

(ii) by the Probabilistic Knowledge Norm for Reasons, these credences can therefore not provide a reason for me to act in any particular way, so that if I choose to do whatever maximises expected utility relative to those credences, my reason for choosing in that way cannot be that the choice maximised expected utility for me, since that invokes my credences as a reason;

(iii) by the Probabilistic Knowledge Norm for Decisions, I am not necessarily required to choose the action that maximises expected utility relative to those credences—they do not correspond to the strongest

[76] For more on the correct decision theory for imprecise credences, see (Seidenfeld, 2004; Seidenfeld et al., 2010; Elga, 2010; Joyce, 2010; Rinard, 2015).

> probabilistic content I know, and thus what is permissible for me is not determined by maximising expected utility with respect to them.

What, then, am I required to do? That depends on what my statistical evidence allows me to know, and what the correct decision theory is for imprecise credences. As I mentioned already, there are many candidate theories, including the liberal and conservative versions described above. And on the question of what my statistical evidence allows me to know, we will have more to say below.

## 10.7 Assessing Moss' No Knowledge Reply: the Paulian view

We have now seen Paul's Utility Ignorance Objection to decision theory, the Fine-Graining Response, Paul's Authenticity Reply, and Moss' No Knowledge Reply. Given this, we can ask two questions: Does Moss' reply work from Paul's point of view? Does Moss' reply work independently of Paul's point of view? Paul emphasises four important features of her objection. As we will see, Moss' reply to the Fine-Graining Response preserves two of those to some extent and two not at all. We begin with those it doesn't preserve.

First, Paul claims that the challenge to decision theory raised by ETEs is unique to those experiences. Whatever problem they raise, it is not raised by any other sort of phenomenon. And yet that isn't true on Moss' interpretation. Consider the doctor who must choose a treatment for her patient. She has the following statistical evidence: in 98% of trial cases, the treatment cures the illness; in 2% of trial cases, the patient deteriorates severely. She sets her credences in line with that. The illness is serious, so this is a high stakes decision. Thus, other reference classes are relevant, and the doctor's evidence cannot rule out that the frequency of successful treatment is very different in those. So, by Moss' lights, the doctor's credence of 0.98 that the treatment will succeed and 0.2 that it will fail do not count as knowledge and so cannot provide a reason for action. Now, you might consider that the wrong conclusion or the right one—you might think, for instance, that the doctor's credences can provide reason for action, even if the doctor would prefer to have better evidence. But that is not the issue here. The issue is only that this other decision faces exactly the same problems that, for Moss, any decision faces that involves ETEs. That is, ETEs do not pose any new or distinctive problem for decision theory. And thus, on Moss' account, we lose this crucial feature of Paul's account.

The second distinctive feature of Paul's account is that, in decisions that involves ETEs, the problem is first-personal. When I am choosing whether or not to become a parent, the problem arises, according to Paul, because I am trying to make a decision for myself about my own future and yet I cannot access a part of my self that is crucial to the decision, namely, my utilities. This is why Paul turns to concepts like *alienation* and *authenticity* to account for the phenomenon: they apply to first-personal choices in a way that they don't to third-personal ones. However, as the example of the doctor from above shows, there is nothing distinctively first-personal in Moss' diagnosis of the problem with decisions that involve ETEs—the problem arises just as acutely for a doctor making a major decision for a patient as it does for me when I try to choose whether or not to adopt.

The first feature of Paul's account that Moss' No Knowledge Reply does preserve and explain, though for quite different reasons, is the importance of what is at stake in the decision that we wish to use our credences to make. As Paul and Moss both acknowledge, there are trivial ETEs and important ones. When I choose whether to spread Vegemite or Marmite on my toast—having tried neither—I am choosing which ETE to have. But neither thinks that this poses a problem for decision making in the way that choosing to become a parent does. Both think it is quite acceptable to use statistical evidence about the utilities that others assign to eating those two condiments as reasons I might cite when making my decision. Paul's explanation: only in significant life decisions do alienation and inauthenticity threaten. Moss' explanation: in low stakes cases, there are no alternative reference classes that are relevant, and so my credences will constitute knowledge even if my evidence cannot rule out any alternative reference classes. Different explanations, but both agree that stakes matter.

The second feature of Paul's account that Moss' reply preserves, though again for quite different reasons, is the attitude to decision theory. It is important to note that neither Paul nor Moss wish to abandon the machinery of decision theory in the face of the Utility Ignorance Objection; neither wishes to reject expected utility theory. Rather, in the case of significant life decisions that might give rise to ETEs, they advocate changing the decision problem that we feed into that decision theory. For instance, on the Fine-Graining Response, when I am deciding whether or not to adopt a child, I formulate the following decision problem:

- the set of possible acts is

$$\mathcal{A} = \{Apply, Don't\ Apply\};$$

- the set of possible states is

$$\begin{aligned}\mathcal{S} = \{&\textit{Parent \& utility of being a parent is } -12,\\ &\textit{Parent \& utility of being a parent is } 3,\\ &\textit{Parent \& utility of being a parent is } 10,\\ &\textit{Fail}\};\end{aligned}$$

- the doxastic states are my credences over those states, on the supposition of those acts;
- the conative states are my utilities over the conjunctions of acts and states, which encode the overall value I attach to these conjunctions, incorporating the quality of the phenomenal experience they give me, the moral and aesthetic values they boast, and so on.

I then feed this decision problem into the machinery of decision theory, which then tells me which of the possible acts are permitted by rationality and which are not.

For Paul, as we saw in Section 10.3, the new decision problem that we feed into the machinery of decision theory is this:

- the set of possible acts is

  $$\mathcal{A} = \{\textit{Choose to discover}, \textit{Choose not to discover}\};$$

  though note that *Choose to discover* is equivalent to *Apply* and *Choose not to discover* is equivalent to *Don't Apply*;
- the set of possible states is

  $$\mathcal{S} = \{\textit{Discover}, \textit{Don't discover}\}$$

  though note that *Discover* is equivalent to *Parent* and *Don't discover* is equivalent to *Child-free*;
- the doxastic states are my imprecise credences over the states, on the supposition of the acts;
- the conative states are my utilities over the conjunctions of acts and states, but instead of encoding the overall value I attach to these conjunctions, which Paul has shown we cannot access prior to making the decision, they encode *only the value I assign to the revelatory experiences involved in those conjunctions*.

Thus, the conative state specified in Paul's version of the decision problem is different from that in the orthodox version, while the doxastic state remains the same.

In contrast, as we saw in Section 10.6, for Moss, the new decision problem is this:

- the set of possible acts is

$$\mathcal{A} = \{\textit{Apply}, \textit{Don't Apply}\};$$

- the set of possible states is

$$\begin{aligned}\mathcal{S} = \{&\textit{Parent \& utility of being a parent is } -12,\\ &\textit{Parent \& utility of being a parent is } 3,\\ &\textit{Parent \& utility of being a parent is } 10,\\ &\textit{Fail}\};\end{aligned}$$

- the doxastic states are not my precise credences over the states, but rather *the strongest imprecise states that count as knowledge for me*;
- the conative states are my utilities over the conjunctions of acts and states, which encode the overall value I attach to these conjunctions, as in the orthodox approach.

Thus, the doxastic state specified in Moss' version of the decision problem is different from that in the orthodox version, while the conative state remains the same.

So, again, Paul and Moss agree—the orthodox decision problem should be replaced. But they agree for different reasons—Paul thinks that the conative state should be specified differently, while Moss thinks the doxastic state should be specified differently.

## 10.8 Assessing Moss' No Knowledge Reply: the independent view

In this section, we continue to consider Moss' No Knowledge Reply to the Fine-Graining Response to Paul's Utility Ignorance Objection to orthodox decision theory. But this time we consider it independently of its relationship to Paul's own reply to that response to her objection. We can read Moss' No Knowledge Reply in one of two ways. On the one hand, granted the possibility of probabilistic knowledge and the accompanying probabilistic versions of the knowledge norms for action—Moss' Probabilistic Knowledge Norm for Reasons and Probabilistic Knowledge Norm for Decisions—we can read it as trying to establish that the Fine-Graining Response is wrong. On the other hand, if we start at the other end and assume that the Fine-Graining Response is wrong, then the need to appeal to probabilistic knowledge to explain why it is wrong is supposed to furnish us with an argument in favour of probabilistic knowledge, its possibility and its use as a concept in epistemology.

The first worry I describe concerns the second reading. I will argue that a notion of probabilistic *knowledge* is not, in fact, required in order to explain the problem with decisions involving ETEs in the way Moss wishes to. The explanation can be given better, in fact, using only the familiar notion of probabilistic *justification*. The central point is this: the feature of first-personal utility credences based on statistical evidence that prevents them from counting as knowledge on Moss' account also prevents them from counting as justified.

In the Fine-Graining Response outlined in Section 10.4 above, I have credence 0.3 in *Parent & utility of being a parent is* $-12$, 0.6 in *Parent & utility of being a parent is* 3, and 0.1 in *Parent & utility of being a parent is* 10. I base these credences on my statistical evidence that 30% of parents assign utility $-12$ to being a parent, 60% assign utility 3, and 10% assign utility 10. Moss claims that these credences do not count as knowledge. I claim that, if they don't, they also don't count as justified.

Moss claims that these credences don't count as knowledge because my evidence doesn't allow me to rule out alternative reference classes that are rendered relevant by the high stakes of the decision I am making. I claim that they don't count as justified for the same reason. After all, the ability to rule out relevant alternatives is important for justification too. Suppose Charlie and Craig are identical twins. I know this; I've known them for years. I also know that I can't tell them apart reliably. I see Craig in the supermarket and I form the belief that Craig is in front of me. Now, while true, my belief does not count as knowledge because I can't rule out the relevant alternative possibility that it is Charlie in front of me, not Craig. But equally my inability to rule out this possibility of which I'm fully aware also renders my belief unjustified. In general, if I believe $p$ and there is an alternative possibility to $p$ such that (i) I'm aware of it, (ii) I'm aware that it's relevant, and (iii) I can't rule it out, then my belief in $p$ is not justified. The cases in which my inability to rule out an alternative precludes knowledge but not justification are those where either I am not aware of the possibility or not aware that it is relevant. For instance, in Goldman's Fake Barn County example, either I am not aware of the possibility of barn facades—perhaps I've never heard of such a thing—or, if I am aware of that possibility, I am not aware that it is relevant—because I don't know that I am in Fake Barn County. Thus, while I might be justified in believing that the structure in front of me is a barn, my belief doesn't count as knowledge. However, as soon as I learn about the possibility of barn facades and learn that I'm currently in Fake Barn County, my belief is neither justified nor knowledge. And the same goes for my credences about my utilities in the case of ETEs. Almost whatever statistical evidence I have about my utilities for becoming a parent, there is some relevant alternative reference class in which there are different frequencies for the various possible utility assignments such that (i) I'm aware of that reference class, (ii) I'm aware it's relevant, and (iii) I can't rule it out. Thus, any precise credence that I assign on the basis of that statistical evidence is not justified.

Thus, it seems to me that Moss' diagnosis of the problem with the Fine-Graining Response is wrong. The problem is not that the credences based on statistical evidence are not *knowledge*, it's that they're not *justified*. If that's right, then the argument in favour of the possibility of probabilistic knowledge that Moss bases on that diagnosis fails.[77]

But this seems a Pyrrhic victory. If I am right, surely this only makes the problem worse for the Fine-Graining Response itself. After all, the possibility of probabilistic knowledge and the putative norms that link it with reasons and decisions are controversial, whereas the possibility of probabilistic justification and the norms that link it with reasons and decisions are not. I think most decision theorists would agree that, while there is sense in which an individual with unjustified credences should maximise expected utility with respect to those credences, such an individual will nonetheless not be fully rational. Thus, we seem to be left with a stronger reply to the Fine-Graining Response than we had before: we might call it the *No Justification Reply*.

[77] Of course, a knowledge-firster will claim that the credence based on the statistical evidence fails to be justified *because* it fails to be knowledge, not the other way around. While I am not a knowledge-firster myself, I think I can remain neutral on that claim here. I wish to say nothing about whether there is such a thing as probabilistic knowledge, nor if there is whether it plays the fundamental role in credal epistemology that the standard knowledge-firster claims non-probabilistic knowledge plays in the epistemology of non-probabilistic belief. I only claim that Moss cannot mount a certain sort of argument in favour of probabilistic knowledge, namely, that it is an essential ingredient in a plausible explanation of the difficulty of decision-making in the presence of epistemically transformative experience. That role can be played just as well by the notion of justification.

But this is too quick. All that the considerations so far have shown is that, if I take a single statistical fact based on the distribution of utilities among people in a single reference class, and set my credences about my own utilities exactly in line with that, without considering anything else, then those credences will typically neither be knowledge nor justified. But there are other, better ways to respond to statistical evidence, and these can give justified credal states that can then be used to make our ETE decisions.

For instance, suppose I have the statistical evidence from above: 10% of all parents assign 10 utiles to being a parent, 60% assign 3 utiles, and 30% assign -12. But I also realise that I have properties that I share with some but not all parents: I enjoy spending time with my nieces and nephew; and I am a moderately anxious person. Let's suppose I think that the latter is the only property I have that affects the utilities I assign to being a parent. That is, I think that the distribution of utilities in the reference class of people who enjoy being around children is much the same as the distribution of utilities in the reference class of all parents, but the distribution among the reference class of moderately anxious people is quite different from the distribution in the class of all parents. And let's suppose that this belief is justified by my background evidence. Now, I don't know exactly what the latter distribution is, since that isn't included in my body of statistical evidence, but I have credences in the various possible distributions that are based on my background evidence. Let's assume again that those credences are also justified by my background evidence. I then use these credences, together with my statistical evidence concerning the distribution of utilities in the reference class of all parents, to set my credences concerning my own utilities for being a parent. The resulting credences will be justified.

Now notice: these credences will be justified not because I've *ruled out* the alternative distributions of utilities among the alternative reference classes, but rather because I've *incorporated my uncertainty* about those different distributions into my new credences concerning my utilities for parenting. And indeed that is the natural thing to do in the probabilistic setting. For many Bayesian epistemologists, nothing that is possible is ever completely ruled out; we just assign to it very low credence. This is the so-called *Regularity Principle*, and there are various versions determined by the various different notions of possibility (Shimony, 1955; Stalnaker, 1970; Lewis, 1980; Jeffrey, 1992). If the Regularity Principle is true, it is too demanding to require of an individual with probabilistic attitudes that they rule out alternative possibilities before they can know anything. Rather, we might say: in order for a probabilistic attitude to be justified, the individual must have *considered* all relevant alternative possibilities and must have determined their attitude by *incorporating their attitudes towards those possibilities*. And we can do that in the case of credences concerning ETEs, even when those credences are based on statistical evidence, as we can see from the example of my adoption decision described above.

Now, I imagine that Moss might reply: while such credences might be justified, they will rarely count as knowledge. In order to count as knowledge, she might say, I must not only consider the properties I have *that I think might* affect the utility I assign to being a parent, and incorporate into my credences concerning that utility my uncertainty about the distribution of utilities for being parent among the reference classes defined by those properties; I must also consider the properties I have *that will in fact* affect that utility, and incorporate my uncertainty about the distribution of utilities for being a parent among the corresponding reference classes. Failing to consider those other properties might not preclude justification—I might be perfectly justified in not having considered those properties, and indeed justified in not even being aware of them. But it does preclude knowledge. Thus, just as I am perfectly justified in ignoring the possibility that the structure in front of me is a fake barn, but will be unable to know various propositions if that possibility is relevant in my situation, similarly, I might be justified in not considering various reference classes and the distribution of utilities within them, but nonetheless will be unable to know various probabilistic content if those reference classes are relevant in my situation. And thus, Moss might claim, by the Probabilistic Knowledge Norms for Action, the justified credences that I formed by incorporating my uncertainty about distributions among alternative reference classes cannot be used in rational decision making in the usual way.

The problem with this claim is that it asks too much of us. If, in order to know a probabilistic content concerning an event in a high stakes situation, you must have considered all of the causal factors that contribute to it being likely to a certain degree, there will be almost no probabilistic contents concerning complex physical phenomena that we'll know. In a high stakes situation, I'll never know that it's at least 50% likely to rain in the next ten minutes, even if it is at least 50% likely to rain in the next ten minutes, since I simply don't know all of the causal factors that contribute to that—and indeed knowing those factors is beyond the capabilities of nearly everyone. There are many situations where, through no fault of our own, we just do not have the evidence that would be required to have credal states that count as knowledge. And this is not peculiar to credences concerning utilities for ETEs, nor even to credences based on statistical evidence.

Now, Moss might reply again: yes, it's difficult to obtain probabilistic knowledge; and perhaps we rarely do; and it's true that people shouldn't be held culpable if they violate the Probabilistic Norms of Actions; but that doesn't mean that we shouldn't strive to satisfy them, and it doesn't mean that the norms are not true. On this reply, Moss considers the Probabilistic Norms of Action as analogous to the so-called Truth Norm in epistemology, which says that we should believe only truths. Certainly, no-one thinks that those who believe falsehoods are always culpable. But nonetheless the Truth Norm specifies an ideal for which we should strive; it specifies the goal at which belief aims; and it gives us a way of assigning epistemic value to beliefs by measuring how far they fall short of achieving that ideal. Perhaps that is also the way to understand the Probabilistic Knowledge Norms for Action. They tell us the ideal towards which our actions should strive; and they give a way of measuring how well an action has been performed by measuring how far it falls short of the ideal.

But that can't be right. To see why, start by considering the following Non-Probabilistic Knowledge Norm for Reasons: a proposition *p* can count as your reason for performing an action just in case you know *p*. Now, that can legitimately be said to set an ideal, because there really is no extra feature of a categorical attitude towards *p* that we would want to add once we know *p*; it just doesn't get any better than that. The problem is that the same cannot be said in the case for probabilistic knowledge. Why? Well, suppose I know that it is at least 50% likely to rain. And suppose I am deciding whether or not to take my umbrella when I go outside. The higher the likelihood of rain, the higher the expected utility I assign to taking my umbrella. If it's over 40% likely to rain, I maximise my utility by taking it when I leave. Thus, since I know it's at least 50% likely to rain, I should take it. But this probabilistic belief concerning rain is not as good as it could be. If it's going to rain, it would be better if I were to believe that it is 100% likely to rain; if it's not going to rain, it would be better if I were to believe that it is 0% likely to rain. What's more, suppose I believe that it's at least 50% likely to rain. And suppose further that my belief is justified but not yet knowledge. Now suppose that I am going to gain one of two possible pieces of evidence. Either (i) I will gain evidence that turns my justified belief that it's at least 50% likely to rain into knowledge; or (ii) I will gain evidence that justifies a belief that it's at least 90% likely to rain, but will not turn that belief into knowledge. Which should I prefer, (i) or (ii)? If probabilistic knowledge is the aim of our probabilistic beliefs, and probabilistic knowledge is the ideal at which we should strive when we form beliefs that ground our decisions, we should prefer (i)—we should prefer to obtain knowledge that it's at least 50% likely, rather than justified belief that it's at least 90% likely. But, I submit, (ii) seems just as good, if not better.

Before we wrap up, I'd like to draw attention to one final point, which is apt to be neglected. On the orthodox version of decision theory, an individual is bound to choose in line with her credences and her utilities—in the precise version of decision theory, for instance, she must pick an act that maximises expected utility by the lights of her current precise credences. Both Moss and Paul argue that this is too demanding in the case of an individual who has adopted the Fine-Graining Response and who sets her credences in the fine-grained states in line with the statistical evidence. Requiring that she chooses in line with her credences, Paul argues, is tantamount to requiring that she makes her decision by deferring to the utilities of others—and that way inauthenticity and alienation lie. For Moss, on the other hand, it is not reasonable to demand that an individual choose in line with beliefs in certain probabilistic contents—which is, after all, what her credences are—when she cannot rule out other probabilistic contents.

However, it is worth noting that the demand that orthodox decision theory makes is in fact rather weak. Suppose **P** is the set of credence functions that represents the strongest probabilistic content that you know. Then, in many cases, and certainly the cases under consideration here, **P** is also the set of all and only the credence functions that you are justified in adopting in light of your evidence. Then, while it is true that, once you have picked your credence function $P$ from **P**, you are bound to maximise expected utility with respect to $P$, you are not bound to pick any particular credence function from **P**—you might pick $P$, but equally you might pick any other $P' \neq P$ from **P**, and you would be equally justified whichever you picked. Thus, the set of permissible choices for you is in fact exactly the same according to the orthodox view and according to Moss' Probabilistic Knowledge Norm for Decisions, when that is coupled with a liberal decision theory for imprecise credences. In each case, an act is permissible if there is a credence function $P$ in **P** such that the act maximises expected utility from the point of view of $P$.

I conclude, then, that Moss' No Knowledge Reply to the Fine-Graining Response does not work. I agree with Moss that credences that are based directly on sparse statistical evidence do not constitute probabilistic knowledge. But I argue that they are not justified either. And it is their lack of justification that precludes their use in decision-making, not their failure to count as knowledge. What's more, there are ways to set credences in the light of purely statistical evidence that gives rise to justified credences. Moss may say that these do not count as knowledge, and I'd be happy to accept that. But if she then also demands that credences used in decision making should be knowledge, I think the standard is set too high. Or, if she thinks that probabilistic knowledge simply serves as an ideal towards which we ought to strive, then there are times when I ought to abandon that ideal—there are times when I ought to pass up getting closer to knowledge in one probabilistic content in order to get justification in a more precise and useful probabilistic content.

I conclude this chapter, then, optimistic that there is no substantial problem with the Fine-Graining Response to Paul's Utility Ignorance Objection, and thus no analogous problem for my favoured solution to the problem of choosing for changing selves.

# Part II

# Setting the weights

# Chapter 11

# The problem of weighting changing selves

In Part I, we made some progress on the problem of choosing for changing selves. We began by considering three putative solutions—the Unchanging Utility, Utility of Utility, and One True Utility Solutions—and we concluded that they don't work. The best version of the Unchanging Utility Solution said that, when you are choosing between options that will affect you in the future, you must choose the one that will obtain for you in the future what you most value in the future, and that gives too little weight to what you value in the present. The Utility of Utility Solution simply ascribes to our higher-order utilities too much authority, giving them veto power over our first-order utilities. And the One True Utility Solution requires an implausibly strong version of objectivism about values on which, as the name suggests, there is one true objectively correct set of utilities.

We then proposed to treat the problem as a judgment aggregation problem. That is, we proposed to consider my various past, present, and future selves as individual members of a corporate entity, and we proposed to consider my present self as making its decision on behalf of that corporate entity. We surveyed the ways we might do that—we might aggregate their preferences over the possible actions, or their evaluations of those actions, or their credences and utilities separately—and we concluded in favour of the latter. Indeed, we concluded that there was no call to aggregate their credences at all—we would, instead, simply use my present self's credences. But we would aggregate their utilities. What's more, we would do this by taking weighted averages. So, at this point in our exploration, we have the shape of a solution. We have the form of the quantity that you should maximise when you make your decision. It is this:

$$V_G(a) = \sum_{s \in \mathcal{S}} P^a_G(s) U^a_G(s) = \sum_{s \in \mathcal{S}} P^a_p(s) \sum_{i=1}^{n} \alpha_{s,i} U^a_{s,i}(w^s)$$

where $P_p$ is your credence function at the present time $t_p$, and $U_{s,i}$ is your utility function at time $t_i$ in state $s$. Notice, however, that this formulation includes parameters that have yet to be fixed, namely, the weights $\alpha_{s,i}$ that apply to each $U_{s,i}$ to produce your aggregate utility function for state $s$. It is the purpose of this part of the book to say something about how those weights might be set.

Often, when decision theorists discuss how to set certain parameters that feed into a decision problem—your credences, perhaps, or your utilities, or your attitudes towards risk—they proceed by specifying rational constraints on these parameters. Probabilism says that you are rationally required to have credences that satisfy the axioms of the probability calculus (recall these axioms from Footnote 55). Orthodox expected utility theory says that you are required to evaluate an act at its expected utility relative to your credences and your utilities (recall (EU1) from Chapter 2). Decision theorists often demand on behalf of rationality that, if you discount the future at all, you must do so exponentially (see Chapter 13 below). And so on. But it is often assumed that, if you satisfy those constraints, there is nothing further to say. You have fulfilled the requirements of rationality, and we can leave you to make your decision. Whatever you choose within these rational requirements is then rationally permissible, and no option better or worse than any other from the point of view of rationality. In what follows, I will also canvas some putative rational constraints. But in nearly all cases, I will find that they are false. Using weights, you take the local utilities that measure the values of your individual selves and aggregate them to give the global utilities that your present self should use in decision making—in Section 2.5, we called these your present self's *decision-making utilities* to distinguish them from their *local utilities*, which record your present values. But, just as many hold with Hume that the local utilities of our individual selves cannot be rationally criticised, so it seems that, beyond the constraint that our global utilities at a time should be a weighted average of our local utilities at all times, our ways of aggregating those local utilities to give the decision-making utilities of my present self can rarely be criticised from the point of view of rationality. In the few cases below where we do wish to impose constraints on the weights we assign to different selves, we will see these are closer to moral requirements, rather than rational ones.

However, while we might conclude that there are few constraints on the way we set our weights, nonetheless, by considering the putative constraints that we will below, we encounter important considerations that we might wish to take into account when we set our weights. For instance, I will argue that there is no rational requirement that you assign equal weight to all of your past, present, and future selves within a state because to do so would be to require certain you to be alienated from the decisions your present self makes. Now, I do not go further and argue that being alienated in this way is irrational, and thus that equal weightings are rationally forbidden. But, nonetheless, by considering why you are not rationally required to set equal weightings, we encounter a consideration, namely, alienation, that you might want to take into account when you set your weights.

This part of my investigation, then, is rather atypical amongst inquiries within decision theory. The upshot will be that there are few absolute constraints on the weights we use to aggregate our past, present, and future selves, and those we do encounter have more of a moral flavour than a rational one. But I will enumerate a number of different considerations that might lead you to set your weights in one way rather than another, even though doing so is not required of you—considerations that it is reasonable for you to use to set your weights; considerations you might find it useful to reflect upon when you set your weights. It seems to me that this is an important role that philosophy can play in the theory of decision making—a role that, too often, it does not play. In that theory, we often assume that your utility function is exactly as you would like it to be—we assume that it is as it should be by your lights. We might say that it is the decision theorist's job to elicit your utility function, perhaps using the techniques described in Chapter 8. And we will certainly say that it's her job to describe how you should use that utility function in your decision making. But it is much less often said that it is the decision theorist's job to give you the tools to change your utility function, to correct or amend it, or indeed to help you set it in the first place. So we rarely hear decision theorists talk about considerations upon which you might reflect in order to determine your utility function or to improve your utilities by your own lights. In this part of our investigation, however, this is exactly what we will do. More precisely, we will explore considerations upon which you might reflect in order to determine the weights you assign to the utility functions of your past, present, and future selves when you are calculating the weighted average that gives the decision-making utilities at the present time.

A common thread runs through the considerations that we identify in this part of the book: all appeal to the Parfitian claim that we should have greater care for the selves to whom we are more closely or strongly connected—in our setting, this means that we should give greater weight to those selves. One of the central themes of Parfit's early work is that the relation of personal identity—which holds between two selves when they are both selves of the same enduring person—is often less important in ethics than the relation of psychological connectedness—which holds between two selves when they have mental lives that overlap significantly, so that they share beliefs, memories, experiences, and so on. An example: he seeks to show that, when we think of some future calamity that will occur to us and wonder whether we will survive it, we care not so much whether the enduring person of which we are currently a part will continue to exist beyond that event, but rather whether a self will exist beyond that event with whom we are strongly psychologically connected (Parfit, 1971, 1984). Another example: he seeks to show that, in order to cancel the moral obligations created by making a promise, you must not only bear the relation of personal identity to the self to whom the promise was made, but you must also bear to them a strong relation of psychological connectedness (Parfit, 1984, 327). Our theme here will be that connectedness is often more important than personal identity not only in ethics, but also in decision theory. In particular, we should assign weight to past, present, and future selves in line with the degree of connectedness between our present self and those selves. In what follows, we will consider three ways in which one self might be connected to another: the first self might have made a sacrifice from which the second self benefits (Chapter 12); the first and second selves might be cognitively connected, so that they share experiences and cognitive states, such as memories, opinions, and beliefs (Chapter 13); and the first and second might be conatively connected, so that they share many values in common (Chapter 14). In Chapter 15, we consider what we might call self-reinforcing and self-frustrating choices. And in Chapter 16, we conclude by pointing to future work that is required to complete, extend, and complement the account of choosing for changing selves that we have been pursuing.

# Chapter 12

# The weight of the past

To introduce the question of this chapter, here are Derek Parfit and Richard Hare talking about times when they had to choose for changing selves:

> When I was young what I most wanted was to be a poet. This desire was not conditional on its own persistence. I did not want to be a poet only if this was what I later wanted. Now that I am older, I have lost this desire. [...] Does my past desire give me a reason to try to write poems now, though I now have no desire to do so? (Parfit, 1984, 157)

> I wanted, when a small boy, to be an engine-driver when I grew up; when I have graduated as a classical scholar at the age of 18, and am going on to take the Ph.D. in Greek literature, somebody unexpectedly offers me a job as an engine-driver. In deciding whether to accept it, ought I to give any weight to my long-abandoned boyhood ambition? (Hare, 1989, 156)

And here is the example of Blandine from the introduction:

> Blandine is pondering her career. For years, she wanted to be a musician in a band. She always placed enormous value on the emotional side of life, and she wished to devote her career to exploring and expressing those emotions through music. Recently, however, she has abandoned this desire completely. She no longer wishes to be a musician, and no longer values the emotional side of life. Indeed, she is now committed to pursuing studies in particle physics. Some friends ask her to join a new band that they are putting together; and on the same day

> she receives an offer to study physics at university. Which path should Blandine choose?

In each of these cases the question is this: should our past values receive any weight at all when we make decisions? Many philosophers claim that they shouldn't. For instance, here is Richard Brandt:

> If a person is deciding what to do for himself, we should think it strange for him to decide on the basis partly of what he wanted or did not want ten years ago. (Brandt, 1992, 171)

Indeed, Brandt appeals to the alleged oddity of giving weight to one's past selves in order to argue against a desire satisfaction version of utilitarianism, since the latter seems unable to diagnose this oddity. In our framework, we might sum up Brandt's claim as follows:

> **The Irrelevance of Past Values (IPV)** According to the judgment aggregation solution to the problem of choosing for changing selves, you should choose in order to maximise the following quantity:
>
> $$V_G(a) = \sum_{s \in \mathcal{S}} P^a(s) \sum_{i=1}^{n} \alpha_{s,i} U^a_{s,i}(s)$$
>
> Then, if $p$ is the present time, and $i < p$, then it ought to be the case that $\alpha_{s,i} = 0$.

In this chapter, I wish to argue against this claim.

Let's begin by noting just how strong IPV is. It says that, for *any* decision problem, faced at *any* time in an agent's life, *every single one* of her past selves *must* receive *no* weighting at all. Thus, when Blandine is choosing whether to study particle physics or to join a band, she is obliged to ignore completely the high utility she assigned in the past to being in a band and exploring the emotional side of life through music. Even though she endorsed that value for many years, and even though her current value is of a very recent vintage, Blandine should decide exactly as she would decide had she always valued studying particle physics in the past, or if she had chopped and changed between wanting to be in a band and wanting to study particle physics, or if she had held entirely different values altogether in the past. According to Brandt and IPV, whatever she held in the past is irrelevant to her decision now (unless it somehow counts as evidence in favour of what she will value in the future). And the same goes for Parfit's choice to become a poet or philosopher, and Hare's decision whether or not to become a train driver.

Since the claim against which we are arguing is so strong, our counterclaim can be correspondingly weak. We need only find *one* decision problem such that it is *at least permissible* that the agent gives *some* positive weight to the values of *at least one* of her past selves. We will begin by considering a more detailed version of an apparent counterexample to IPV that Brandt considers—he attributes it to Derek Parfit and James Griffin. Brandt rejects it. I do too. But I think Brandt's objection to it fails, so I offer an alternative. After that, I will turn to a different sort of case.

## 12.1 Deathbed decisions

Let's begin, then, with the sort of example that Richard Brandt attributes to Derek Parfit and James Griffin (Brandt, 1992, 171):

> **Deathbed** All of his adult life, Bill has been an atheist. But this has always been more than a belief for him. It is also a set of values. Throughout his life, he has always valued facing the fact of his mortality with dignity and without the crutch of believing that there is an afterlife; and he has valued greatly his ability to find meaning in his life without positing some divine creator with a grand design for the universe into which he fits. On his deathbed, however, his mortality terrifies him and his imminent non-existence (which he believes awaits him) makes his life seem meaningless to him. He comes to want a visit from a priest, so that he might receive some of the comfort of a blessing, the church's forgiveness, and the hope of eternal life in heaven. Should he call for a priest?

According to the Irrelevance of Past Values, he should. Though he used to value a life without the comfort of the church, he no longer does. He is minutes from death and so we can assume that these new values are not just those of his current self, but also the values of all of his future selves—he will not change his values again. Thus, if we assign no weight to the past selves, the new values are all there is—he should choose in line with them and call the priest.

For many people, this is an unpalatable conclusion, and Parfit and Griffin suggest that we should reject IPV in the light of this consequence. Brandt, on the other hand, thinks that we can explain away such a putative counterexample as follows:

> Some persons will not feel comfortable about [the fact that IPV exhorts Bill to call the priest], although one can avoid this consequence [...] by claiming that [Bill's] present desire is not "ideal" and hence should be ignored. (Brandt, 1992, 172)

It's not entirely clear what Brandt has in mind here when he says that Bill's new preferences are not "ideal". Is it that they are impermissible, like the desire to kick puppies or pull a cat's tail? Surely that is too strong. I am an atheist myself, but I don't take theists to have impermissible values. Is it that they are formed in an impermissible way, like a desire you form following a nasty knock on the head or because you take a certain drug? Bill's new preferences are the result of fear and self-preservation; they are formed in the cognitively chaotic environment of the deathbed. Do these count as impermissible ways to form preferences? Perhaps. But we needn't pursue the question further, for whatever Brandt means, his solution will not work. He says that, given that they are not "ideal", Bill's new values can be ignored. In our framework, this amounts to assigning them a weight of zero. But, since he advocates IPV, he also claims that Bill's past values should be ignored—so they too receive a weight of zero. If that's the case, then, *all* of Bill's values receive weight zero, and Bill's decision whether to call the priest to his deathbed is not constrained by his past, present, or future values nor by anything else—in other words, anything goes. But that's not the outcome Brandt wants—he wants it to be required of Bill that he should not call for the priest. So Brandt's objection to this putative counterexample to IPV cannot be right.

Here's a more plausible alternative, which also seeks to show that, in Parfit's and Griffin's example, IPV does not entail that Bill should call the priest. When we hear a case like Bill's, we suspect that it has been misdescribed—we suspect that Bill's values have not changed at all. Rather, even on his deathbed, Bill is still an atheist at heart, and he still values the strength and effort it takes to confront eternal non-existence without religious belief. What happens, though, is that, lying on his deathbed, overwhelmed by existential terror, Bill is visited by temptation; he is tempted to act against his current, enduring, atheist values.

Now, economists often understand temptation as a temporary shift of preferences or utilities (Thoma, ta). Thus, when you are tempted by the extra chocolate when the box comes your way, you switch from valuing your long-term health more than your short-term pleasure to valuing pleasure over health—and then you switch back so that you can regret it moments later. If that is the case, then Bill really does shift his values on his deathbed, and it furnishes us with a counterexample to IPV, as Parfit and Griffin claim. On this account, if Bill is tempted, his preferences shift from atheist to theist, and once again we can only save the intuition that he should not call for the priest by saying that he should give weight to his past values.

But I think this is the wrong understanding of temptation. Of course, if you are a behaviourist and take utilities to be mere shadows cast by preferences, which are in turn determined by choice behaviour, then this account of temptation is exactly what you must say—when I am tempted, my choice behaviour differs from what my prior preferences suggest it should be, and so the behaviourist must say that my preferences have changed. But for us realists, this orthodox account of temptation seems to miss something important. Acting in accordance with temptation involves a tension in one's decision-making process—that's why you feel indecisive in cases of temptation; it's why we talk of temptation pulling you in one direction while your conscience pulls you in the other. If you simply switch utilities and then act in accordance with your new utilities, there is no such tension—your actions and your preferences line up perfectly at all times.

What's the alternative? Here's a proposal. In cases of temptation, we retain our underlying values, but do not act on the basis of them. Rather, we act on the basis of some other values—or, at least, a different balance of our values—or we act as a result of other influences on our actions that lie beyond our preferences. Thus, I retain my balance in favour of long-term health over short-term pleasure, even when I am tempted by the chocolate box, but I act instead on the basis of values that balance those two goods differently; or I act because the visceral urge for the serotonin hit that the chocolate will deliver interferes with the usual pathway between my values and my actions, influencing my actions independently of my values. And similarly for Bill. He values the integrity that comes from living out the practical consequences of his values. But he also values comfort. In his true, underlying values, he gives greater weight to integrity than to comfort. And these values endure right to the end of his life. But, in his final moments, he acts on utilities that arise from giving greater weight to comfort and less to integrity. They aren't his true values, but he acts on them all the same. They might arise from a surge of existential terror that visits him as he contemplates his infinite non-existence and skews his thinking. And that creates the familiar tension we associate with temptation.

On this account, temptation arises when the usual connection between mental attitudes and actions breaks down. Bill retains his atheist values, but they don't guide his actions in the way they should; he acts in line with other values. I retain my value for health over pleasure, but that doesn't guide my action when confronted with the chocolate box; it is my urge for a serotonin hit that does that. In this sense, then, this account renders temptation similar to cases in which, for instance, I have an extremely low credence that something will happen, but I act as if I have a much higher credence in it. This is sometimes what happens in cases of anxiety. If I am anxious about the possibility of being in an airplane crash, I might take a lengthy land route that would only be rational if my credence in the possibility of an airplane crash were much higher than it in fact is. In both cases—the case of temptation and the case of the airplane crash—something interferes with the usual way in which my beliefs and desires determine my actions: in the anxiety case, it might be the rush of cortisol that begins when I imagine my plane crashing; in the temptation case, it might be a visceral urge for pleasure or comfort.

Let's now apply this account to Bill's decision. I suggested above that, on his deathbed, Bill is drawn by temptation towards calling the priest. And, if the account of temptation sketched above is correct, this means that Bill retains his atheist values, even when he is feeling the pull of temptation, but he is pulled by temptation because something interferes with the usual pathway between his beliefs and desires, on the one hand, and his actions, on the other. If this is correct, it explains why Bill should not call for the priest—doing so would conflict with his current values. There is no need to appeal to his past values in order to achieve this verdict.

Further evidence that this is the correct interpretation of the case comes when we think of another sort of conversion late in life. Unlike Bill, Brenda's conversion occurs before she is on her deathbed. Indeed, it occurs while she believes herself to be many years away from death. On a weekend trip to the Vatican as an interested, atheist art lover, she finds himself face-to-face with Michelangelo's *Pietà*. It moves her profoundly. She returns each day during her visit and stands before it silently. It changes something in her during that trip—she finds God and converts. On her journey home, she is in a car accident. And as she lies on her deathbed in the hospital, she calls for a priest. In this case, I think, we are much more inclined to say that it is at least permissible for her to do this. The reason, I submit, is that we are inclined to say here that Brenda is choosing based on her true current, deathbed values. She is not drawn by temptation. Rather, she has genuinely changed.

In the end, then, I think Parfit's and Griffin's case is not, in fact, a counterexample to IPV. To find a case that is, we will have to look harder.

## 12.2 Past values and permissibility

Consider the following interpersonal example. We will use it to motivate an intrapersonal/interself example. And indeed this will be our methodology throughout the remainder of this chapter. We will appeal to interpersonal examples to motivate principles that we will then test and apply in the intrapersonal/interself case.

> **Roshni** has just lost her grandmother, who made a number of sacrifices for her while she was alive. When Roshni's mother was young, Roshni's grandmother worked long hours in hard, manual employment to provide money to put her daughter through school. Later, when Roshni was born, her grandmother would spend a lot of time teaching her outside school—introducing her to poetry, music, and mathematics. She had other interests—she was an avid painter; she liked to walk in the countryside around their home, observing the animals that lived nearby; she enjoyed reading intellectual biographies of the scientists of her era. But she put all of those on hold to give Roshni the best possible education. Roshni is now trying to decide whether or not to go to university. Though she never said so, Roshni knows that her grandmother wanted her to go. What part, if any, should this play in Roshni's reasoning?

It is pretty clear that it is permissible for Roshni to take her grandmother's preferences into account, even though these are in the past. Suppose she goes to university and, on the first day, her new friends ask her why she's there. Roshni explains that she was reasonably equally balanced between going and not, leaning slightly towards not going. But taking into account her grandmother's past desire and the sacrifices she made tipped her the other way. We would, I think, judge this a sound reason—we would not judge that she had made the wrong decision; we would not say she should have ignored her grandmother's preferences because they are past.

Perhaps, at base, Roshni's desire at least to take her grandmother's preferences into account might stem from two reasonable sources: first, a desire to do good for a person who has made sacrifices in order to benefit you; second, a belief that it is possible to do good for a person who has died by fulfilling the preferences they had when they were alive. The first is uncontroversially permissible. Soon we will come to the claim that you are sometimes *obliged* to do good for a person who has made sacrifices in order to benefit you. And that is certainly controversial. But the mere claim that it is reasonable or permissible to want to do this is not.

The second claim is more controversial. Can we benefit the dead? Can we harm them? My view is that we can do both. However, for our purposes here, we need only say that we can benefit the dead. And we do so, I claim, at least when we bring about an outcome that they valued consistently throughout their life, and which it was permissible for them to value during that period. Thus, for instance, Roshni would benefit her grandmother by going to university. On this account, of course, it is possible to benefit someone without changing any intrinsic feature of them—you need not bring about happiness or pleasure or any other intrinsic state in that person in order to benefit them. You can change relational facts; and indeed benefitting someone is simply bringing the world a little more into accord with their permissible values (Papineau, 2012). So I benefit you if I donate to your favourite charity on your behalf, even if I never get a chance to tell you this, and you remain ignorant of my generosity and its consequences.

So it is permissible for Roshni to take her grandmother's wishes into account in her decision-making. It is reasonable—her grandmother's preferences are legitimate reasons that she may weigh against her own. Giving them some weight is not mere sentimentality. Now, the same, I think, is true of past selves. It is permissible to take their values into account when you choose what to do now. They are past, of course, just as Roshni's grandmother's are. But, in just the same way that Roshni's grandmother made sacrifices in order to benefit Roshni and put her in the position she was in at 18 years old, choosing whether or not to go to university, so it is often the case that my past selves have made sacrifices in order to benefit my current and future selves. And, in such a situation, it does not seem any odder for me to give at least some weight to the values of those past selves when I make my decision now than it does for Roshni to give weight to her grandmother's preferences.

Recall Blandine, for instance: until recently, what she valued most was exploring the emotional side of human life. She wanted to find ways of conveying complex emotions through music. Above all else, she valued the possibility of breaking down barriers of communication between people by producing music that elicited emotional responses in them that she couldn't express using language. It was her life's project. To the end of pursuing it, she made sacrifices: for instance, instead of spending as much time with her school friends as she would have liked, she spent time practising her musical instruments and studying at school. She realised she was unlikely to make much of a living from her music and wanted to ensure she could earn money from other skills to support herself. Then, two weeks ago, she stopped valuing the pursuit so intensely. In fact, her values flipped. She had never held that a life spent investigating the physical, non-human world was wrong or impermissible or otherwise reprehensible. She just always valued the emotional side a great deal more. But now, above all else, she values exploring that physical, non-human world—and, in particular, the objective features of the subatomic world. Her past selves would have been distraught to learn this. The years they spent in preparation will have been largely in vain if her current self makes its decisions entirely without reference to their values. The sacrifices they made will have benefitted Blandine—she would not have secured the place to study physics without them—but they will not serve the project for which they were made. So, when Blandine faces the choice two weeks later between joining her friend's band and studying particle physics, it seems quite reasonable for her to give some weight to the values endorsed by these past selves whose sacrifices bestowed on her the benefits she now enjoys, just as it is reasonable for Roshni to give some weight to her grandmother's values.

Before we move on, let me consider a possible worry about this line of argument. In this section, I have drawn parallels between Blandine's case and Roshni's. In both cases, they have benefitted from the actions of a self other than their current self—Roshni benefitted from various of her grandmother's past selves; Blandine from various of her own past selves. But there seems to be an important disanalogy. The value that Roshni's grandmother attached to education is now past—but not because she abandoned it; it is past because she died. In Blandine's case, on the other hand, her value for conveying the emotional side of life is also now past—but not because she died; it is past because she abandoned it. If we are to use the interpersonal cases like Roshni's to inform our thinking about intrapersonal cases like Blandine's, we must ask whether this disanalogy is relevant. Personally, I think it isn't.

First, note that the disanalogy disappears if we think of Blandine as a group agent whose members are her different selves at different times. In this case: yes, there is an enduring entity who abandons their values at one time in favour of new values—namely, Blandine; but, on the more fundamental level, there is a series of past selves that made sacrifices—like Roshni's grandmother—and never changed their values—also like Roshni's grandmother—but rather went out of existence with their values intact—again, like Roshni's grandmother. If that's right, then it seems that the strength of the analogy between Roshni and her grandmother, on the one hand, and past-Blandine and present-Blandine, on the other, turns on your metaphysics of persons—is a person a group agent composed of different selves at different times; or is it a single undecomposable entity? And which is the correct metaphysics of persons for Blandine's case?

I think it is important to conceive of Blandine as composed of different selves. Indeed, this is already suggested by the judgment aggregation solution for which we argued in the previous section, which treats a person as a group agent whose members' various opinions and attitudes must be aggregated to provide the group agent's opinions and attitudes. It is also reflected in the way we talk about Blandine's case and cases like it. When someone's values change in this sort of way, we are inclined to say not that the person *changed their mind*, but rather that the person themselves *changed*—in extreme cases, we say that they became a different person; indeed, Blandine might look back on the past and say, 'I was a different person then'. And this fits better with the view of Blandine on which she is a group agent composed of different selves.

Let me describe one further consideration against the alleged disanalogy between Roshni's case and Blandine's. I think we are sometimes prepared to treat the two differently because we are tacitly (or explicitly) working with an objectivist account of subjective utility. For the objectivist, recall, your subjective values are just shadows cast by your credences concerning the objective values—in particular, they are your expectations of the objective values. Thus, on this account, the change in Blandine's subjective values must be the result of a change in her credences concerning the objective values. And the difference between Roshni and her grandmother must be that her grandmother has a very high credence that education has great objective value, while Roshni has a more middling credence in this. Now, recall from our discussion of the Reflection Principle in Chapter 5: we tend not to take past credences into account when we are making decisions, because we assume that whatever wisdom or information was contained in them is already incorporated into your current credences, which were formed on the basis of those past credences. But we don't believe anything similar about Roshni's grandmother. There is no reason to assume that the wisdom or information that was contained in Roshni's grandmother's credences concerning the objective value of education is already incorporated into Roshni's. And thus it seems reasonable for Roshni to give some weight to her grandmother's subjective values (which are determined by her credences), even though it is not reasonable for Blandine to give any weight to the values of her past selves (which are determined by their credences). Roshni's grandmother's subjective values are evidence of her credences about the objective values, and those credences might well be the sort of evidence that should influence Roshni's own credences about the objective values, and thus her own subjective values. In contrast, while Blandine's past subjective values are evidence of her past credences about the objective values, those credences are not the sort of evidence that should influence her current credences about the objective values, and thus her current subjective values, because we assume that they have already been taken into account. Thus, for the objectivist, there is a clear disanalogy between Roshni's case and Blandine's.

However, recall: we are taking a subjectivist line in this book. And so we should not take the above considerations as a reason to judge Roshni's case and Blandine's differently. So if the apparent disanalogy we perceive between Roshni's case and Blandine's is grounded in this objectivist thought, we should abandon it.

Indeed the sort of case we deal with here suggests a line of objection against the objectivist. They cannot honour the intuitive difference between the case we are considering, in which Roshni's grandmother made sacrifices that benefitted Roshni, and an alternative, in which she did not. I hope we'll agree that it is more reasonable for Roshni to take her grandmother's wishes into account in the first case than in the second. But, according to the objectivist, there should be no difference between these cases. In both cases, Roshni's grandmother's subjective values are simply indicators of her credences about the objective values and these are just as good as evidence about the objective values themselves whether or not she made sacrifices for Roshni. Thus, Roshni's reaction to learning about them should be the same either way.

In sum: we have described a certain sort of case in which it is at least permissible for an agent to give some weight to her past selves when she makes a decision. If those past selves made sacrifices that benefitted the agent, we would not find it unreasonable for her to give some weight to those past values when she makes her decision. Thus, I submit that the Irrelevance of Past Values is false.

## 12.3 Past values and obligations

I expect that this much will not seem very controversial. Those, such as Brandt, who argue for IPV, cite examples like Parfit's or Hare's autobiographical cases outlined above, and simply puzzle at the peculiarity of paying attention to a set of values you've since discarded. But if we pick our examples a little more carefully, and we fill in the details a little more clearly, and we ask what reasons people might have for paying such attention, we see that they exist and can be legitimate. What's more, when you think about your past values as the values had throughout their existence by your past selves, rather than as the abandoned values of the enduring person you are, the analogy with Roshni and her grandmother is more pertinent, and we let our judgment in that case inform our judgment in this case.

Our next question is whether you are ever *obliged* to give past selves some weight. Is there ever a situation in which it would be wrong for you to assign no weight whatsoever to the values held by some of your past selves?

As before, let's start with the interpersonal case and use that to inform our thinking about the intrapersonal/interself case, which is our main concern here. Roshni's case at least shows that sometimes you are *permitted* to take into account the past values of the dead—or, more generally, the values of past selves, whether or not the person to which they belong still exists. Are you ever *obliged* to do so? And, if you are, under what circumstances, and to what extent?

In the end, I wish to argue for a version of the following principle:

> **The Beneficiary Principle (intrapersonal, sacrifice version)** A current self that has justly benefitted from certain sorts of sacrifice made by some of its past selves has an obligation to give a certain amount of weight to the preferences of these past selves.

Different versions of this arise from specifying different exchange rates for translating sacrifices into obligations. We'll discuss how these might work below when we consider the following interpersonal version of this principle:

> **The Beneficiary Principle (interpersonal, sacrifice version)** An individual (or group) who has justly benefitted from certain sorts of sacrifice made by another individual (or group) has an obligation to give a certain amount of weight to the preferences of this individual (or group).

### 12.3.1 Three related principles

In this form, the Beneficiary Principle is close to a principle that H. L. A. Hart calls the *Principle of Mutual Restriction*, and which John Rawls' adapts to produce his *Principle of Fairness*. Here's Hart's formulation:

> when a number of persons conduct any joint enterprise according to rules and thus restrict their liberty, those who have submitted to these restrictions when required have a right to a similar submission from those who have benefited by their submission. (Hart, 1955, 185)

And here's Rawls' principle (under Robert Nozick's formulation):

> when a number of persons engage in a just, mutually advantageous, cooperative venture according to rules and thus restrain their liberty in ways necessary to yield advantages for all, those who have submitted to these restrictions have a right to similar acquiescence on the part of those who have benefited from their submission. (Rawls, 1975, 108-14), (Nozick, 1974, 90)

For Hart and Rawls, such principles are supposed to explain why it is permissible to charge taxes for such public goods as the security provided by a nation's army and police, the healthcare services provided by its ambulances and emergency rooms, the education supplied by its schools and colleges, the infrastructure built by its government, etc. They are supposed to explain, furthermore, why it is permissible to demand those taxes from any person who benefits from these goods, whether or not they requested or desired those services, and whether or not they consented to the scheme in the first place. If others are restricting themselves—paying their taxes—to provide these public services, then Hart's and Rawls' principles say that those people have a right to expect others who benefit from the services to restrict themselves in the same way—that is, again, to pay their taxes. Those who don't, as Arneson (1982) points out, are free riders.

How do the Principles of Mutual Restriction and Fairness relate to the Beneficiary Principle?

First: note that my principle speaks of *obligations* while Hart and Rawls speak of *rights*. I take it that rights always create obligations, but not all obligations are created in this way. So, in this sense, my principle is weaker.

Second: Hart and Rawls think that when others sacrifice in a certain way—by restraining their own liberty—those who benefit should sacrifice in the same way. But it seems to me that this cannot be the most basic principle, and indeed that it must only hold in certain circumstances. After all, if those who have sacrificed have no desire that others sacrifice in the same way, those others are surely not obliged to do so. For instance, there are often residents of a country who are exempt from paying taxes to fund the security, healthcare, education, infrastructure, etc. that their government provides. These are typically residents on low incomes. We might assume that those who do pay these taxes do not desire that those on low incomes should do so as well. And, in that situation, surely those on low incomes are not obliged to pay taxes. For this reason, I drop this requirement and say instead that the beneficiaries of the sacrifices should assign some weight to the preferences of those making the sacrifices when the beneficiaries make their decisions. Sometimes, and in many of the cases that Hart and Rawls have in mind, those making the sacrifice will desire that others make it as well. So, in these situations, the Beneficiary Principle will entail the obligations that the Principles of Mutual Restriction and Fairness entail.

Third: both Hart and Rawls place restrictions on the way the sacrifice—or restriction of liberty—must come about in order to generate the obligation—or right. For them, it must be a result of cooperative activity governed by rules. But this restriction seems arbitrary. It is hard to see why haphazard sacrifices that arise organically and not from cooperation should have less power to create obligations than orderly ones that grow from coordinated collective effort. Consider a social movement in which individuals begin to speak out against a certain form of oppression—sexual harassment, racism, homophobia, or transphobia, for instance. Perhaps a handful of courageous individuals do it first, without coordinating, simply spurred by an incident in which the oppression becomes very visible. Each of them makes a sacrifice, because their visibility makes them vulnerable, costs them opportunities within their career when they are viewed as troublemakers, etc. Then others start to join. It seems that those who could effectively and visibly speak out come under an obligation to do so as they begin to benefit from the movement that has been created. But no rules governed the movement. For this reason, I remove Hart's and Rawls' requirement that the benefits must result from a coordinated and rule-governed set of sacrifices.

Fourth: one difference between Hart's principle and mine, on the one hand, and Rawls', on the other, is that Rawls insists that the benefits achieved by the sacrificers are shared by all, those who sacrifice and those who don't alike. Yet surely this is not needed to generate the obligation—if those making the sacrifice do not receive any benefit from it, it can hardly diminish their right to demand similar sacrifice from those who do benefit.

Finally, fifth: a difference between Rawls' principle and mine, on the one hand, and Hart's, on the other, is that Rawls and I insist that the obligation or right is generated only if the benefit received is just; that is, if it is just that the beneficiary receives this benefit. For me, this is motivated by the following sort of case: suppose two parents make substantial sacrifices in order to send their daughter to a very expensive private high school—they both take second jobs, increase their stress, remortgage their house, etc. The society into which their daughter passes after her schooling is still riddled with class-based discrimination, and as a result, she benefits greatly from the social capital provided by her expensive schooling—her accent, her contacts, the line on her CV, a certain confidence that such schools often instil. But the benefit is unjust. And, for this reason, she has no obligation to take the preferences of her parents into account.

### 12.3.2 The libertarian objection

So much for the relationships between the three principles—the Beneficiary Principle and the Principles of Mutual Restriction and Fairness. Let's turn now to their soundness. The standard objection to all three comes from the libertarian. Here is the heart of Robert Nozick's version:

> You may not decide to give me something, for example a book, and then grab money from me to pay for it, even if I have nothing better to spend the money on [...] You have, if anything, even less reason to demand payment if your activity that gives me the book also benefits you; suppose that your best way of getting exercise is by throwing books into people's houses, or that some other activity of yours thrusts books into people's houses as an unavoidable side effect. . . . One cannot, whatever one's purposes, just act so as to give people benefits and then demand (or seize) payment. Nor can a group of persons do this. (Nozick, 1974, 95)

Now, notice here that Nozick's central objection concerns not just obligations or rights, but the legitimacy of enforcing them. Thus, his complaint against the Principle of Fairness is that, if it were true, someone else's action could lead to you being coerced into action. And he thinks such coercion is illegitimate.

There are two responses to Nozick available here. First, we might point out that nearly all political systems short of pure anarchism require some positive obligations that can be legitimately enforced by coercion. Indeed, Nozick's own favoured Lockean system does so. It posits the obligation to respect private property that has been appropriated legitimately. And it seems unlikely that Nozick can argue that it is *permissible* to coerce people not to use land that you appropriated without their consent, but *impermissible* to coerce them to contribute to a scheme from which they have benefitted but which was created without their consent (Arneson, 1982). The problem for Nozick is that the cases are too similar for there to be a substantial normative difference between them. In the first case, the appropriator creates an obligation for others to which those others do not consent. This is supposed to be made palatable for the libertarian because the appropriator thereby also grants a right to those others, namely, the right to appropriate land as she has done. So, in this case, an obligation is created without consent, but this is offset with a benefit, in the form of a right, also distributed without consent. In the second case, the sacrificer also creates an obligation for others to which those others do not consent. But, again, this seems to be made palatable because the sacrificer thereby also distributes a benefit to those others. Now, the benefit does not usually come in the form of a right. But it is hard to see why that difference would be sufficient to create a normative difference between them.

The second response, at least on behalf of the Beneficiary Principle, is this: we might agree with Nozick that some of the obligations you have because you have benefitted from the sacrifices of others should not be enforced coercively, but that nonetheless, such benefits do create these obligations. After all, there are plenty of obligations that may not be enforced coercively. And many of those arise for us even though we do not voluntarily agree to them. Uncontroversially, there are negative obligations—the obligation not to be unkind to my friends and my obligation not to turn your friends against you—which apply even to those who do not consent to be ruled by it. And only slightly more controversially, there are also positive obligations, such as the obligation to help a drowning child or an elderly person who has fallen in the street, providing you need put yourself at no great risk in order to do so.[78]

However, it is not clear whether such analogies serve our purpose. In the standard examples of positive obligations that govern you without your consent, the obligation is often to act as if you have a desire that morality demands you should have anyway. I am obliged to help a drowning child even though I never promised I would, nor entered into any contract binding me to do it. But the obligation here is to do something that I should want to do anyway. Anyone who doesn't value saving that child is something of a monster. However, the positive obligations entailed by the Beneficiary Principle will not always be like that. They exhort beneficiaries to take into account the preferences of those whose sacrifices have benefitted

[78] Indeed, some of these latter obligations are even enforceable in certain jurisdictions. These are known as *duty to rescue laws*, and they exist in Canada, Germany, and a number of other countries.

them. Now, those preferences *might* be ones that the beneficiary themselves ought to have, but there is no reason why they should be. So it would better support the Beneficiary Principle if there were situations in which we have positive obligations to take account of the preferences of others.

We might, for instance, imagine a group of people who find themselves in a situation in which, at any time, only one of them can make decisions; but that decision-making role rotates around the group. Perhaps an evil kidnapper has placed them in this situation; perhaps it is the result of a natural occurrence. In this case, I think, we would say that each individual in the group has an obligation, when their time comes around, to take into account the preferences of the others in the group. And that is true regardless of whether or not those preferences are ones that the appointed decision-maker at that time should have.

Notice here that the obligation arises not because of any benefit that the decision-maker receives as a result of the sacrifices of the others, but rather simply from the difference between their power to make decisions and the lack of decision-making power enjoyed by the others. It is worth remarking that this interpersonal case is structurally analogous to the intrapersonal case of choosing for changing selves—in the latter case, the power to make decisions moves from earlier selves to later selves as a result of time passing.

### 12.3.3 Exchange rates

As I mentioned above, as it stands, the Beneficiary Principle is really a schema. Without exchange rates, which specify which sacrifices resulting in which benefits generate which obligations, it does not say anything very specific. With that in mind, we now turn to those exchange rates.

When we specify the obligations generated, there are two variables we must set: first, the strength of the obligation—how bad would it be not to fulfil it?; second, the range from which a beneficiary is obliged to pick the weight they will assign to the preferences of the sacrificers. Let's see what might determine those.

First: as we noted above, all agree that no obligation is generated if the benefit received is not just—better: no obligation *to give weight to the sacrificers' preferences* is generated; as we will see in Section 12.3.6 below, when we consider the obligation to provide reparations, other obligations may well arise. The benefit received must be just, but the means by which it is acquired must also be just: no obligation arises to give weight to the sacrificers' preferences if they created the benefit by taking a significant risk and expending a significant amount of time on stealing from the elderly and infirm.

Second: the strength of the obligation increases with the extent to which the benefits were received voluntarily. We might distinguish three levels:[79] (I) the benefits received were actively sought; (II) the benefits were not actively sought, but they were not relinquished when they were received, even though they could have been; (III) they were not actively sought and they could not be relinquished. The obligations generated for beneficiaries should diminish as we move through these levels. But the weights assigned to the preferences of those whose sacrifices created the benefits should remain fixed.

Third: the strength of the obligation and the minimum weight that must be assigned to the sacrificers' preferences both increase as the extent of the sacrifice increases. The more someone sacrifices for your benefit, the greater your obligations is to give their preferences some weight, and the greater that weight should be. Suppose, for instance, that your college tuition is paid for by a generous benefactor. You benefit from this and so have some obligation to give some weight to that benefactor's preferences in your decision-making. But the strength of that obligation and the weight you're required to give is much greater if the benefactor has a modest income and remortgaged their house to bestow the benefit upon you than if they are a very wealthy person who barely felt the loss of money. Indeed, in the case of an extremely wealthy person, paying one person's college tuition might constitute no sacrifice at all, since at their level of wealth the cost might make no different to their own utility.

Fourth: how does the extent of the benefit received affect the strength of the obligation and the minimum weight? I am less confident that I know what to say here, since considerations seem to pull in two directions. On the one hand, we can surely think of plenty of examples in which it should make no difference. Two friends each spend a full week planning birthday surprises for you. The first happens to have the better idea—she gives you something that you enjoy enormously, and thereby benefits you greatly. The second gives you something you like well enough and which benefits you only moderately. It seems that you are nonetheless equally obliged to give weight to their preferences, and the minimum weight you are obliged to give is equal. Similarly when two strangers spend a whole afternoon helping you look for your lost son. You don't seem to have any greater obligation to the one who happens to find him. And yet, on the other hand, this suggests that you can be laden with great obligations even though you

[79] Though see (Pasternak, 2014) for a more fine-grained account.

have benefitted hardly at all, so long as the tiny benefits you have received have taken enormous effort. And that, too, seems wrong.

### 12.3.4 Moral blackmail

A final point about the applicability of the Beneficiary Principle. The most significant concern about that principle is that it opens up the possibility of what I will call *moral blackmail*. We all know of people who do indeed make sacrifices that benefit others, but who make those sacrifices with the express intention of creating a moral obligation in the beneficiary to pay some attention to their preferences. In some of its most sinister manifestations, we have people who bestow gifts, do favours, listen to woes, and so on, in an attempt to create in the recipient a moral obligation to have a romantic or sexual relationship with them—you have only to look to many toxic discussions of 'friend-zoning' to find ample evidence of this. Whatever our account of the exchange rates that generate obligations from sacrifices, it must somehow preclude this possibility.

In fact, I think the problem of moral blackmail does not arise. Suppose Nat performs an action that is detrimental to him at the time he performs it. And suppose further that he does this with the intention of creating a moral obligation in Neil to acquiesce of certain of his requests. In this case, I would say, Neil's actions do not constitute a sacrifice. Rather, he has made what he considers an investment. He takes a hit at this time in the hope that he will be compensated at a later time. Thus, the case is analogous to a company that invests time in training an employee at the beginning of their career in the hope that they will reap the benefits of this employee's expertise later in their career. In such a case, we would never say that the company makes a sacrifice; and we would never say that the employee is under any obligation to give weight to the company's preferences in their decision-making. And similarly in the case of Nat and Neil. Nat's action would only count as a sacrifice if it did not hold promise of compensation. And if his action does not count as a sacrifice, the Beneficiary Principle does not apply; in particular, it does not entail that Neil has any obligation to give weight to Nat's preferences.

Now, you might think that, given this defence, the Beneficiary Principle will no longer entail the sorts of obligation that interest Hart and Rawls. After all, in those cases, sacrifices made by many in the group result in a collective good that benefits the whole group. Surely, then, the actions of the many in the group do not count as sacrifices, since they come with the hope of compensation, namely, the collective good.

My response: it is true that they come with that hope; but nonetheless, in many of the cases, they count as sacrifices. A sacrifice is an action that leaves you worse off than if you had not performed it, or at least leaves you with the expectation of being worse off than if you had not performed it. Nat's action, therefore, does not count as a sacrifice because he expects to be better off in the long run as a result of performing—indeed, that is his motivation for performing it. But the many who sacrifice for the collective good are not usually in that position. The cases that interest Hart and Rawls are those in which any individual member can reasonably expect that if they don't perform their action, the collective good will arise all the same. That is, they are true free rider cases. For instance, if I fail to pay my taxes, I can reasonably expect that my country's security provision will continue to function and my security will be ensured. So, if an individual member does perform their action—e.g., if I do pay my taxes—then they do something that they can expect will leave them worse off than if they had not. The good will arise whether they perform it or not, and the action is detrimental to them.

You might also worry, however, about a different case. Again, Nat performs actions that are detrimental to him and beneficial to Neil. And again he does so with the intention of generating an obligation for Neil; he wishes to ensure that Neil is obliged to give weight to his preferences. But this time he does so with no expectation that Neil will provide compensation by giving weight to Nat's preferences—perhaps he knows that Neil is not moved by moral considerations; or perhaps he knows that Neil is rarely aware of his moral obligations. In that case, surely we would say again that Nat's actions fail to generate an obligation for Neil. I agree. And to avoid that conclusion, I say again that Nat's actions in this case do not count as sacrifices. Your action counts as a sacrifice, I say, only if (i) you don't expect compensation, and (ii) were your intentions in making the sacrifice satisfied, you would still not receive compensation. In the case we have considering, Nat's actions satisfy (i) but not (ii). He doesn't expect compensation, but were his intentions satisfied, he would get it. So Nat does not make a sacrifice, and thus the Beneficiary Principle entails no obligations.

### 12.3.5 Applying the Beneficiary Principle

In the light of these remarks about the exchange rates that convert sacrifices and benefits into obligations, and the circumstances under which actions count as sacrifices, we are now in a position to see how the Beneficiary Principle might apply in the case of Roshni and her grandmother.

Our first question: does her grandmother make sacrifices, or is she like Nat in our example from the previous section? I used the term when I described Roshni's case initially at the beginning of this chapter. But since then we have taken greater care to define it. Should we still describe Roshni's grandmother's actions thus? Roshni's grandmother performed actions that were detrimental to her—they took her away from the pursuits she loved. But that is not enough. In order for such an action to count as a sacrifice, her grandmother must not expect compensation, and she must perform the action with a certain sort of intention, namely, one that will not result in compensation if it is satisfied. Is this the case here? That depends. If Roshni's grandmother provided her with an education with the intention that she would then go on to university, then she has not made a sacrifice; if, on the other hand, she did so with the intention that Roshni should be in a position to go to university if she so wished, then she has made a sacrifice.

Let's suppose that she does make a sacrifice. Our next question: are the benefits that Roshni received just? This is not uncontroversial, since you might think that everyone deserves equal educational support and encouragement and that Roshni's grandmother bequeaths on her granddaughter an unfair and unjust advantage. But let us assume it here it is just. As a result, the Beneficiary Principle applies, and Roshni is not only permitted to give some weight to her grandmother's past preferences, but indeed obliged to do so.

Next, we consider the strength of that obligation, and the minimum weight she is obliged to give. Note that Roshni receives the benefits involuntarily: she isn't able to refuse the time her grandmother spends on her education when she is young; and she certainly isn't able to prevent her grandmother from spending the money on her mother's education that she did before Roshni was born. What's more, the benefits of an education cannot be relinquished. So, based on this alone, Roshni's obligation won't be very strong. However, the benefit received and the sacrifices that brought it about are both significant, and this increases not only the strength of Roshni's obligations, but also the minimum weight she can give her grandmother's preferences when she makes her decision. So Roshni has a moderately strong obligation to give at least some weight to her grandmother's preferences. Now, recall: before she takes account of those preferences, her decision was almost on a knife edge—she had a preference for not going to university, but it was very slight. It seems plausible, then, that the obligation to give weight to her grandmother's preferences is sufficient to tip her decision in the other direction and make her choose university.

### 12.3.6 Obligations and reparations

Shortly, we will move to the intrapersonal version of the Beneficiary Principle, and there we will see how it can create obligations to give weight to past preferences. But first I make a short digression to argue that a principle closely related to the Beneficiary Principle provides a good codification of certain moral intuitions we have. The intuitions in question are those in favour of obligations to provide reparations for harms inflicted by some members of past generations on other members of past generations. Think, for instance, of the wrongs inflicted on enslaved people by slavers during the transatlantic slave trade; or by the perpetrators of a genocide, such as The Holocaust, on its victims; or the harm inflicted on certain low-income, low-polluting countries—such as Bangladesh—by high-income, high-polluting countries—such as the UK, the US, and countries in Europe—as a result of the historic destruction of their shared natural resources, such as Earth's atmosphere and its oceans.

In these cases, there are usually four groups: two exist at an earlier time—the harmed group (let's call them G1, e.g. black Africans enslaved during the transatlantic slave trade) and the harming group (G2, e.g. the white transatlantic slavers)—and two exist at the present time—one of these groups bears an important relationship (R1) to the harmed group (G3, e.g. the current population of Black Americans and Black British) while the other bears an important relationship (R2) to the harming group (G4, e.g. the current population of white Americans and white British). The reparations claim says:

> **The Principle of Reparations** G4 owes reparation to G3 for the harms inflicted by G2 on G3.

And there are two puzzles about this principle:

- *Sins of the Fathers puzzle* First, given that the harm was inflicted by G2, why does the obligation to provide reparations fall on G1?
- *Inheritance puzzle* Second, given that the harm was inflicted on G3, why should the reparations be paid to G4?

I propose that both puzzles are solved if we accept the following principle:

> **The Beneficiary Principle (interpersonal, harm version)** An individual (or group) who has benefitted from certain harms inflicted upon another individual (or group) has an obligation to give a certain amount of weight to the preferences of the harmed group, at least to some extent.

Assuming this, we solve the Sins of the Father Puzzle as follows: G4 is not obliged to provide reparations because they inflicted any harm, nor because they inherit the sins of their ancestors, but rather because G4 has benefitted from the harms inflicted on G1 by G2. And we solve the Inheritance Puzzle as follows: G4 is obliged to take into account the preferences of G1, and this entails that they are obliged to take into account the preferences of G3, since G1 has reason to want outcomes after their death that are in line with whatever G3 wants.

Solved in this way, we can see what the relationship R1 between G1 and G3 must be, and what the relationship R2 between G2 and G4 must be, in order for reparations to be owed—R1 must be such that G4 benefits from G2's harm to G1; R2 must be such that we can reasonably assume that members of G1 would want, after their death, that the preferences of G3 be satisfied. Often, the relevant relationships are simply those of genealogical descendants: such descendants usually inherit benefits from their ancestors; and ancestors typically desire the satisfaction and happiness of their descendants.

Of course, this is a very indirect argument for the harm version of the Beneficiary Principle, not least because the obligation to provide reparation is not by any means universally accepted. And so it is an even more indirect argument for the sacrifice version. But I set it out here because it does provide a unified route to solving the Sins of the Fathers puzzle and the Inheritance Puzzle.

### 12.3.7 The intrapersonal version of the Beneficiary Principle

So much, then, for the interpersonal version of the Beneficiary Principle. Having presented it, honed it, and defended it, we are now ready to consider the intrapersonal version that is our true focus here.

> **The Beneficiary Principle (intrapersonal, sacrifice version)** A current self that has justly benefitted from certain sorts of sacrifice made by some of its past selves has an obligation to give a certain amount of weight to the preferences of these past selves.

As it stands, this is a schema. As with the interpersonal version, we must specify exchange rates before it will entail any specific obligation. Those exchange rates, I suggest, are the same as those in the interpersonal case. So no obligation is generated if the benefit is unjust; the obligation is stronger the more voluntary the receipt of benefits; the obligation is stronger and the minimum weight greater when the sacrifice is greater; and actions count as sacrifices only if the agent does not expect compensation, or if the undertake their action with an intention that, if fulfilled, would not necessarily result in their compensation.

How would this play out in the case of Blandine from above? Again, our first question is whether Blandine's past selves made genuine sacrifices. When they stayed home to study whilst their schoolfriends were out at the bowling alley, did they do so with the expectation of compensation by their future selves, or on the basis of intentions that, if satisfied, would lead to compensation? If not, you might wonder why they did these things? By analogy with Roshni's case, we might imagine that they did this to put Blandine in a strong position to make her choice when the time came. In that case, Blandine's past selves sacrificed for her current self. And, like Roshni, the benefits that Blandine's current self thereby received are just. So the intrapersonal Beneficiary Principle applies, and an obligation is generated. What's more, those benefits were received involuntarily, and they cannot be relinquished; the benefits were substantial, and while the sacrifices were not as significant as those Roshni's grandmother made, they were also not trivial. Thus, like Roshi, Blandine is obliged to give at least some weight to the preferences of her past selves.

We began, in this chapter, with Brandt's claim that it is never permissible to give any weight whatsoever to values held in the past—we summarised this as the Irrelevance of Past Values. Since then, we have seen a number of reasons to reject it. First, we saw cases in which it is permissible to give weight. And soon after we saw how, in some of those cases, it is also mandatory.

# Chapter 13

# Discounting the future

As most people will attest, we prefer getting good things—like money or chocolate or pure pleasure—sooner rather than later, while we prefer getting bad things—like dental examinations or bad news or pure pain—later rather than sooner. For instance, I will take two blocks of chocolate today instead of three bars in a month's time. I'll pay now to ride my favourite rollercoaster tomorrow rather than next year. And I'll opt to undergo the painful surgery I need next month rather than next week, even if I will become slightly less healthy as a result. In Parfit's terminology, we are *near biased* (Parfit, 1984, 313). When it comes to acquiring good things and avoiding bad ones, we favour those of our future selves that lie in the nearer future—we do more to obtain for them the good and shelter them from the bad than we do for the future selves that lie in the further future. In the terminology of the economists, we engage in *temporal discounting* (Strotz, 1955; Frederick et al., 2002; Doyle, 2013). When we make decisions, we discount the goodness of things more the further into the future we acquire them. And here's Socrates describing the phenomenon in Plato's *Protagoras* (356a5-d3), diagnosing its cause, and using that diagnosis to debunk the attitude in question:[80]

> For if someone were to say: 'But Socrates, the immediate pleasure is very much different from the pleasant and the painful at a later time,' I would reply, 'They are not different in any other way than by pleasure and pain, for there is no other way that they could differ. [...] Answer me this: Do things of the same size appear to you larger when seen near at hand and smaller when seen from a distance, or not? [...] If then our well-being

[80] The following translation is by Stanley Lombardo and Karen Bell in (Plato, 1997).

> depended upon this, doing and choosing large things, avoiding and not doing the small ones, what would we see as our salvation in life? Would it be the art of measurement or the power of appearance?'

Like Plato, many more recent philosophers and economists have held that most of the ways in which we are biased towards the near future and against the far future are irrational. The idea is this: if you are biased in this way, then you assign different values to two different possible states of the world that contain the same total goods and bads for you, but distributed differently over time; what's more, you assign those different values solely on the basis of where the goods and bads are positioned in the history. And that, they contend, is not a legitimate reason. So temporal discounting is irrational in the same way that it is irrational for me to prefer the £10 note in your right hand to the £10 note in your left hand, and thus to pay some money to receive the former rather than the latter; or the way it is irrational for me to prefer visiting towns that are a prime number of miles from my home town. In each of these cases, the difference between the outcomes is not a legitimate reason for the preference I have between them. That, I take it, is Socrates' point when he says: "They are not different in any other way than by pleasure and pain, for there is no other way that they could differ."

Below, we will describe Parfit's response to this argument, and we will endorse it. However, as Arif Ahmed (2018) shows, Parfit's account renders temporal discounting vulnerable to a different charge of irrationality. According to this charge, which is familiar from the economics literature since Strotz (1955), all but one particular type of temporal discounting will leave you vulnerable to exploitation. If you are near-biased but not near-biased in this particular way, there is a series of decision problems you might confront in which the choices you make will leave you at least as badly off at all times, and worse off at some times, as some other series of choices you might have made instead. And this, many argue, makes you irrational. Ahmed argues convincingly that, if you discount in the way Parfit recommends, you will be exploitable in this sense and thus irrational. I will agree that typically you will be exploitable, but I will deny that this makes you irrational.

At the end of the chapter, I will conclude that this holds two lessons for the Aggregate Utility Solution to the problem of choosing for changing selves. Firstly, if it works, Parfit's defence of giving less weight to a good thing when you will receive it further in the future is also a defence of giving less weight to the utility function of a future self when they lie further in your future. So if we can save it from Ahmed's criticisms, we furnish decision makers with another consideration on which they may reflect when they set the weights they apply to the utility functions of their other selves. Secondly, on one formulation, the problem with Parfit's solution is that, if agents discount the future in the way Parfit suggests, they will prefer one option to another at one time, and then reverse that preference between those options at a later time. Such preference reversals are exploitable and thus, the standard argument goes, irrational. But of course, without stringent constraints on the weights you might attach to different selves at different times, the Aggregate Utility Solution will endorse such preference swaps in many cases. So, if they are irrational, then the Aggregate Utility Solution is in trouble. I hope to show that they are not.

## 13.1 The epistemic defence of temporal discounting

Before we consider Parfit's defence of temporal discounting, let's briefly consider a popular alternative. When we first think about our bias towards the near future, we are often tempted to offer an epistemic defence. I am happy to pay to take the rollercoaster tomorrow rather than next year not because I bizarrely think a history in which it happens tomorrow is better than one in which it happens next year. I don't—I value those two histories equally. Rather, I am happy to pay because I am more confident I will be alive tomorrow than I am that I will be alive in a year, and so my expected value for the option in which I pay and ride tomorrow if I'm still alive is higher than for the option in which I don't pay and ride in a year if I'm still alive.

Now, surely this does account for *some* of our near bias. When I decline to sign up for a government saving scheme that will add £1 when I'm sixty to every £4 I deposit in a savings scheme before I'm forty, I prefer the £4 now to the £5 in twenty-five years because I am not sure I will be alive in twenty-five years, whereas I am quite sure I'm alive now. But it cannot account for all of it. I would have to have irrationally low confidence in my continued existence to account for much of my temporal discounting. To see this, consider an example. Suppose I value chocolate as what economists are wont to call a *non-dependent good*.[81] This means that an extra block of chocolate adds exactly the same amount to my utility regardless of how

[81] Money is perhaps the classic *dependent* good. How much extra utility I obtained from £10 depends greatly on how much I have already. If I am poor, it adds a great deal of utility. If I am a millionaire, it adds very little. Pure pleasure is perhaps the classic *non-dependent* good.

much chocolate I already have. As it is sometimes said, my utility is linear in chocolate, so that the utility I assign to a quantity of chocolate is a linear function of that quantity. So, for instance, as a risk-neutral agent, I'll be indifferent between a guaranteed block of chocolate and a coin toss than will give me two blocks if it lands heads and none if it lands tails. Now, suppose I'd prefer one block of chocolate today to two pieces in a week's time—not an unrealistic assumption. Then, if temporal discounting is simply an epistemic matter, I would have to think it 50% likely that I will die in the coming week—a very unrealistic assumption. So I don't think this can account for all cases of our near biases. Indeed, I think it can account for rather few.

## 13.2 Parfit's defence of temporal discounting

Parfit offers an alternative defence of the rationality of temporal discounting. While my past, present, and future selves are all members of the same corporate entity—namely, me—my current self is not connected equally to all of them. My recent future selves will share more memories with my current self than with far future selves; similarly for my beliefs, hopes, fears, anxieties, and conceptual resources. Not all of these differences will make a difference to how much I care about those other selves, but some of them might. I will likely feel more inclined to help some than others. I will favour some over others, so that some will receive greater weighting in my consideration than others because I am more strongly connected with them, and identify more strongly with them. And, most often, the degree of connection that is relevant will diminish as we consider further and further future selves.

This, Parfit contends, is a legitimate reason to discount the further future. I do not discount future goods merely because they are further away in time—that would be irrational. Rather, I discount them because they will be enjoyed by selves with whom I have a weaker connection and to whom I therefore give less consideration—this is rational. Thus, just as I prefer a situation in which my closest friend gets to ride her favourite rollercoaster over a situation in which a stranger gets to ride theirs, and just as I have those preferences because I am more closely connected to my closest friend than to the stranger, so I prefer those near-future selves with whom I'm closely connected receiving a block of chocolate rather than a far future self, thirty years from now, with whom I feel little connection.

Parfit's proposal is not only a successful rationalisation of the way we value the future. It seems to me that it correctly accounts for the phenomenology of discounting as well—I do indeed feel differently towards my far future selves and my near future selves. I identify less with the former than the latter. And, as a result, I care less about them. Nonetheless, it faces an apparent problem. According to Arif Ahmed, the sort of discounting it endorses often, perhaps always, seems to be open to rational criticism. To see how, we must introduce a very simple formal framework in which we can consider the different ways in which we might discount the future.

## 13.3 A formal framework for temporal discounting

In this framework, there is just one commodity that you value—let's say it's chocolate, though it might be episodes of pure pleasure. Whatever it is, though, it must be a commodity whose impact on the utility of your whole life is localised to the time at which you receive it—that is, having this commodity on Monday makes a difference to how well your life goes on Monday, but not to how well it goes on Tuesday. As a result, money is not a quantity we might use for this purpose. If you receive some money on Monday, you might spend some of it that day, but save some until Tuesday; or you might spend it all on Monday but on something from which you continue to obtain pleasure well beyond Monday. Chocolate that you must consume on the day you receive it will serve us well.

At any given time, your self at that time assigns utilities to receiving different quantities of chocolate at that time. We assume that these utilities are comparable between your different selves at different times, and that they are assigned in the same way by all selves. What's more, for the sake of simplicity, we'll assume that chocolate is a non-dependent good for you, and so you utility is linear in chocolate. Indeed, we can therefore measure your utilities on a scale such that each block of chocolate adds exactly one extra utile. Thus, given a quantity of chocolate $q$—two blocks, say, or fifteen—let $u(q)$ be the utility that each of your past, present, and future selves assigns to receiving that much chocolate at the time they exist. Then $u(q) = q$. Now, when we specify a state of the world, we only need to specify, for each time, how much chocolate you receive at that time in that state of the world. Thus, in state $s = \langle t_1, q_1; t_2, q_2; \ldots; t_n, q_n \rangle$, you receive $q_1$ blocks of chocolate at time $t_1$, $q_2$ blocks at $t_2$, ..., and $q_n$ blocks at $t_n$.

Next: at each time $t$, you have a *discount function* $D_t$ such that, if $t' > t$ is a later time, $D_t(t')$ is the unique real number such that you're indifferent between $N$ utilies at $t'$ and $D_t(t') \times N$ utiles at $t$. Thus, in our current framework, where $u(q) = q$ for any quantity of chocolate $q$, $D_t(t')$ is the unique real number such that you're indifferent between $N$ blocks of chocolate at $t'$ and $D_t(t') \times N$ blocks at $t$. So, for instance, if on Monday I am indifferent between two blocks of chocolate tomorrow and one block today, then $D_{\text{Monday}}(\text{Tuesday}) = 0.5$. The blocks of chocolate tomorrow count only half as much towards my total utility as the blocks of chocolate today. Another way to think of $D_t(t')$ is as the exchange rate between two currencies—block-of-chocolate-at-$t$ and block-of-chocolate-at-$t'$. Given that we tend to discount the future, rather than mark it up, $D_t(t')$ is typically at most 1—that is, the future currency is less valuable than the current currency. And of course, by definition, $D_t(t) = 1$—that is, you are indifferent between $N$ blocks of chocolate at $t$ and $N$ blocks of chocolate at $t$!

Now, given all of this, we can give the following account of my utility for a given state $s = \langle t_1, q_1; t_2, q_2; \ldots; t_n, q_n \rangle$ of the world at a time $t_k$:

$$\begin{aligned} u_{t_k}(s) &= u(q_1) + \ldots + u(q_k) + D_{t_k}(t_{k+1})u(q_{k+1}) \ldots D_{t_k}(t_n)u(q_n) \\ &= \sum_{i=1}^{k} u(q_i) + \sum_{i=k+1}^{n} D_{t_k}(t_i)u(q_i) \\ &= \sum_{i=1}^{k} q_i + \sum_{i=k+1}^{n} D_{t_k}(t_i)q_i \end{aligned}$$

Thus, we don't discount the past at all, since that has happened. But we do discount the future, and we do so according to our discount functions $D_t$.

In what follows, we will consider various constraints on your discount functions $D_{t_1}, \ldots, D_{t_n}$, and we will consider which, if any, are rational requirements. Then we will lay out Ahmed's argument that Parfit's account gives rise to discount functions that violate those rational requirements, and we will respond to it. Finally, we will ask what lessons this holds for assigning weights to future selves in our Aggregate Utility Solution to the problem of choosing for changing selves.

## 13.4 Constraints on discounting functions

The first constraint is known as *delay consistency*. This says that, for any time, while the amount that I discount the future at that time will increase as we move further from that time into the future, the *rate* at which I discount it should remain the same—that is, the amount by which my discount increases over a period of time should be a function only of the length of the period; it should not depend on the starting point of the period. So, for instance, if I am currently indifferent between one block of chocolate next

Monday and two blocks next Tuesday, then I must be indifferent between one block of chocolate next Friday and two blocks next Saturday, and I must be indifferent between one block of chocolate on 12th February 2048 and two blocks on 13th February 2048. In symbols:

> **Delay Consistency** For any time $t$, any times $t_1, t_2$, and any interval $a > 0$,
> $$\frac{D_t(t_1 + a)}{D_t(t_1)} = \frac{D_t(t_2 + a)}{D_t(t_2)}$$

Our second constraint is known as *time consistency*. It says that the extent to which my discount increases over a period of time should be the same from the vantage point of any time. Thus, if on Monday I am indifferent between one block of chocolate on Thursday and two on Friday, then on Tuesday I should be indifferent between one block of chocolate on Thursday and two on Friday, and similarly for Wednesday and Thursday. In symbols:

> **Time Consistency** For any times $t < t' < t_1$, and any interval $a > 0$,
> $$\frac{D_t(t_1 + a)}{D_t(t_1)} = \frac{D_{t'}(t_1 + a)}{D_{t'}(t_1)}$$

The third constraint is known as *stationarity*. It says that your discount function should 'look the same' from any point in time. Thus, if on Monday I am indifferent between one block of chocolate on Monday and two on Tuesday, then on Tuesday I should be indifferent between one block of chocolate on Tuesday and two on Wednesday, and so on. In symbols:

> **Stationarity** For any times $t_1, t_2$ and any interval $a > 0$,
> $$D_{t_1}(t_1 + a) = D_{t_2}(t_2 + a).$$

Now, it turns out that these three constraints are intimately connected, as the following result shows:

**Lemma 13.4.1**

(I) *If Stationarity, then (Delay Consistency ⇔ Time Consistency).*

Our fourth and fifth constraints say that if we discount the future, we should do so in a very specific way, namely, by what economists call exponential discounting. The first says that, at each time, we should discount the future exponentially; the second adds to this that the particular degree of exponential discounting at each time should be the same:

> **Time-Dependent Exponential Discounting** For any time $t$, there is $\beta_t$ such that, for any later time $t'$,
>
> $$D_t(t') = \beta_t^{t'-t}$$
>
> **Time-Independent Exponential Discounting** There is $\beta$ such that, for any time $t$ and any later time $t'$,
>
> $$D_t(t') = \beta^{t'-t}$$

Now, while Delay Consistency, Stationarity, and Time Consistency are intuitively plausible, most people have no intuitions concerning Time-Dependent or -Independent Exponential Discounting, either for or against. However, Delay Consistency is in fact equivalent to Time-Dependent Exponential Discounting, while Delay Consistency and Stationarity together are equivalent to Time-Independent Exponential Discounting (and, since Stationarity and Time Consistency are equivalent to Stationarity and Delay Consistency, they are jointly equivalent to Time-Independent Exponential Discounting as well). Those constraints narrow down the possible forms of discounting function to just the exponential ones:

**Lemma 13.4.2**

(II) *Delay Consistency* ⇔ *Time-Dependent Exponential Discounting.*

(III) *Delay Consistency* + *Stationarity* ⇔ *Time-Independent Exponential Discounting.*

(IV) *Time Consistency* + *Stationarity* ⇔ *Time-Independent Exponential Discounting.*

All proofs in this chapter can be found in the Appendix.

Our sixth constraint is not stated explicitly as a constraint on the form of your discount functions, or the way the values they assign at different times relate to one another, but rather as a constraint on the choices they would lead you to make. It says that it is irrational to have discount functions that would leave you vulnerable to certain exploitation. We say that a series of discount functions at the different times throughout your life is *exploitable* if there are two different times in your life and decision problems you might face at those times such that there are choices you could make that would leave you always at least as well off and sometimes better off than you will be as a result of the choices your discount functions will in fact lead you to make at those times. This notion will be familiar to those who know the Dutch Book arguments for Probabilism, for instance (Ramsey, 1931; Hájek, 2008; Vineberg, 2016). Those attempt to show that credences that are not probabilities are exploitable in the sense that there is a choice you could make—namely, abstain from betting—that will leave you always at least as well off and sometimes better off than the choices that are sanctioned by your individual credences—namely, accepting bets that your individual credences consider fair or favourable.

> **Unexploitability** The sequence $D_{t_1}, D_{t_2}, \ldots, D_{t_n}$ of discount functions through an agent's life is unexploitable.

Now, as Strotz (1955) showed, if your discount functions satisfy Stationarity but violate Delay Consistency, then they are exploitable, whereas if they satisfy Stationarity and Delay Consistency, then they are unexploitable. And furthermore if they violate Time Consistency, then they are exploitable, whether or not they satisfy Stationarity.

**Lemma 13.4.3**

(V) *If Stationarity, then (Unexploitability* $\Leftrightarrow$ *Delay Consistency).*

(VI) *Unexploitability* $\Rightarrow$ *Time Consistency.*

So, if you satisfy Stationarity, then you should satisfy Delay Consistency on pain of exploitation. And, whether you satisfy Stationarity or not, you should satisfy Time Consistency on pain of exploitation. Again, the proof is provided in the Appendix to this chapter. But it will be useful to sketch the proof strategy here—in particular, the proof that you are unexploitable only if you are time-consistent. The point is this: if your discount functions violate Time Consistency, then there are times $t < t'$ such that your discount rate at $t$ between two later times, $t_1$ and $t_2$, is different from your discount rate at $t'$ between $t_1$ and $t_2$. We can then use this to concoct a series of states of the world—$s_1$, $s_2$, $s_3$. These differ only in the quantities of chocolate you receive at $t_1$ and $t_2$. Because of your discount rates at $t$, you will prefer $s_2$ to $s_1$ at $t$; and because of your discount rate at $t'$, you will prefer $s_3$ to $s_2$ at $t'$. But, and herein lies the rub, $s_3$ will deliver you less chocolate at both of those later two times than $s_1$ will. Thus, you would have been better off had you chosen $s_1$ to begin with and then stuck with it.

## 13.5 Ahmed's objection to Parfit

At this point, we have all the ingredients we need to mount Ahmed's critique of Parfit's defence of temporal discounting. Parfit's claim is that it is rational for me to discount the utility of receiving a quantity of chocolate more the further in the future it lies because I am less strongly connected to my far future selves than to my near future selves, and it is rational to give less consideration or care to selves to which I am less strongly connected. Ahmed objects that my discounting functions must satisfy Stationarity as a matter of rationality; but, if Parfit is right, they will not satisfy Delay Consistency. Therefore, they will leave me vulnerable to exploitation.

First, the requirement of Stationarity. Ahmed argues that this is simply a requirement of rationality:

> non-stationarity is normatively unsatisfactory for reasons that do not apply to Parfit's original idea. For it makes Alice's present evaluation of a *future* delay depend not only on the futurity and length of that delay but also on what date it is *now*. More specifically, her rate of time-preference at any time is a function of the date $t$ [...] But why should her *past* longevity have normative bearing on her *present* concern for her *future* self? (Ahmed, 2018)

I will object to this below, so I have nothing more to say here.

Second, the failure of Delay Consistency. According to Parfit's proposal, the extent to which I discount goods enjoyed by my future selves should track the extent to which I am connected to those future selves. The more strongly connected I am, the more I should care about them, and the more their utility at that time will contribute to my current overall utility for the state in question—thus, the less I will discount them. Now, as Ahmed points out, for Parfit, the degree of my connectedness with a future self is determined by the extent to which traces of the psychological states of my current self are preserved in my latter self. Thus, the more that I share memories, character traits, experiences, thoughts, and beliefs with my future self—that is, the more the traces of those states in me currently exist in my future self—the more strongly connected I am to them. Ahmed then points out that, typically, these sorts of states can be grouped by how quickly they decay—that is, how soon after they first appear the traces of them disappear. Thus, my memory of scratching my nose two minutes ago is already fading and will likely be gone within the next two or three minutes, whereas my memory of a message I received this morning from a close friend telling me when her second baby is due will likely stay with me for years. The problem for Parfit's account is that, because of this, the rate at which I currently discount a future time will decrease as we move further into the future. To see this, we'll make two comparisons. First, consider the degree of connection between me now and me in one minute—which, for Parfit, tracks $D_{\text{now}}(\text{now + one minute})$—and the degree of connection between me now and me in an hour—which, for Parfit, tracks $D_{\text{now}}(\text{now + one hour})$. The first will be much greater than the second. After one minute, I'll retain lots of the short-term memories of what I'm doing now, and I'll be having experiences that are quite similar to the experiences I'm having now. After an hour, on the other hand, most of those memories will have decayed and my experiences will be quite different. Thus,

$$\frac{D_{\text{now}}(\text{now + one hour})}{D_{\text{now}}(\text{now + one minute})}$$

will be much much lower than 1. Second, consider the degree of connection between me now and me in one year and one minute—which tracks $D_{\text{now}}(\text{now + one minute + one year})$—and the degree of connection between me now and me in a year and an hour—which tracks $D_{\text{now}}(\text{now + one hour + one year})$. The first will only be slightly greater than the second. In a year and a minute, all the short-term, fast-decaying mental states I have currently will have decayed, and what connection there is will be due to long-term memories and character traits and enduring beliefs and other stable, long-lasting mental states. But very few if any of these will decay in the period between a year and a minute from now and a year and an hour from now. Thus,

$$\frac{D_{\text{now}}(\text{now + one hour + one year})}{D_{\text{now}}(\text{now + one minute + one year})}$$

will be quite close to 1. So

$$\frac{D_{\text{now}}(\text{now + one hour})}{D_{\text{now}}(\text{now + one minute})} < \frac{D_{\text{now}}(\text{now + one hour + one year})}{D_{\text{now}}(\text{now + one minute + one year})}$$

And this violates Delay Consistency.

Thus, Ahmed's argument runs as follows:

(A1) Stationarity is a requirement of rationality.

(A2) All Parfitian discount functions violate Delay Consistency.

(A3) If you satisfy Stationarity and violate Delay Consistency, then you are exploitable.

(A4) If you are exploitable, then you are irrational.

Therefore

(AC) All Parfitian discount functions are irrational.

The argument is valid. If a Parfitian discount function violates Stationarity, it is irrational by (A1). If it satisfies Stationarity, then it is irrational by (A2-A4).

Before we move on, it is interesting to note that economists—and, following them, Ahmed—feel the need to give such an argument for the irrationality of certain discounting functions. It reveals that they don't endorse the Platonist argument from above that preferring near-future goods to far-future goods is irrational because a difference in the times at which goods are distributed is not a legitimate reason for preference. This reveals, I think, how thoroughly Humean such thinkers are. They hold that we cannot criticise the rationality of preferences other than by pointing to their propensity to lose us those things that we claim to value. No objective constraints on what it is reasonable to value can be adduced.

## 13.6 Amending Ahmed's objection

In my treatment of Ahmed's argument, I want to begin by denying (A1). I think Stationarity is not a requirement of rationality. And I think that it will be violated often by agents whose discount functions are linked to degrees of connectedness in the way Parfit envisages.

Let's start by considering Ahmed's argument for Stationarity. If Alice violates it, Ahmed notes, she makes her "present evaluation of a *future* delay depend not only on the futurity and length of that delay but also on what date it is *now*". But what is so strange about that? Of course, the way Ahmed puts it makes it sound strange. It sounds as though Stationarity requires that Alice's valuation depends on the date alone in some irrational fetististic way: she discounts the future differently on 1st March 2019 and 1st September 2019 just because they are those dates and she has some irrational, possibly superstitious attitudes to these particular dates. But of course that needn't be the case. Alice might, on 1st March 2019, discount utilities at 2nd March 2019 differently from how, on 1st September 2019, she discounts utilities at 2nd September 2019. And she might do this not because she assigns any special significance to those dates—she might not even know of them under this description—but rather because she knows, on 1st September 2019, that she is more strongly connected to her future self on the following day than, on 1st March 2019, she will be to her future self the day after that. Perhaps

1st March is just an ordinary day, nothing remarkable or noteworthy, no major incidents. Her memories from that day will already have faded by the next day, and so her connectedness to her next-day self will be weak. On the other hand, 1st September might mark Alice's first day in a new job, the day her country passes legislation for which she's campaigned her whole life, the day her first child is born, or the day her best friend falls in love. In that case, the experience of that day will endure and remain present to her mind at least into the following day and likely much longer, and so her connectedness to her next-day self will be quite strong. So Parfit's account neatly illustrates why Stationarity is not rationally required.

Of course, you might object that Parfit's account sanctions irrational discount functions, and thus cannot be used to furnish examples of rational violations of Stationarity. But recall that the rational requirement of Stationarity is a premise in Ahmed's objection to Parfit's account. So the conclusion of that argument cannot be cited against it unless there is an independent motivation for it. And if there is, we don't need Ahmed's argument.

So, on Parfit's account, there will be violations of both Stationarity and Delay Consistency. Now, there are exploitable discount functions that violate both principles. That is why we cannot strengthen Lemma 13.4.3. So it is compatible with what we have learned so far that Parfit's account gives rise only to discount functions that are not exploitable. But that is a vain hope. While Stationarity is not a rational requirement, it is often rationally permissible. Above, we imagined that 1st March 2019 and 1st September 2019 are very different days for Alice. But we could equally imagine that all of Alice's days are very similar—each is just an ordinary day, nothing remarkable, nothing to write home about. Thus, for each day, when the following day comes, there has been the same attrition of memories and experiences and beliefs. And thus, Alice satisfies Stationarity. If we combine this with Ahmed's reason for thinking that Parfitian discount functions will violate Delay Consistency, we can infer that Alice's discount functions are exploitable. If we accept premise (A4) of Ahmed's argument—exploitability entails irrationality—they are therefore irrational. Thus, we might reformulate Ahmed's argument as follows:

(A1′) Some Parfitian discount functions satisfy Stationarity while violating Delay Consistency.

(A3) If you satisfy Stationarity and violate Delay Consistency, then you are exploitable.

(A4) If you are exploitable, then you are irrational.

Therefore,

(AC′) Some Parfitian discount functions are irrational.

Again, the argument is valid.

## 13.7 Exploitability does not imply irrationality

I reject (A1), and so I reject the first version of Ahmed's argument. I accept (A1′), but I still reject this second version of the argument. The problem is the move from exploitability to irrationality in premise (A4). To see what goes wrong in that inference, it is worth considering an alternative proof of Unexploitability ⇒ Time Consistency—that is, Lemma 13.4.3(VI)—which is what supports the previous premise (A3).

### 13.7.1 Time Consistency and Unexploitability

If I violate Time Consistency—which I do if I satisfy Stationarity and violate Delay Consistency—then there are $t < t' < t_1 < t_2$ such that

$$\frac{D_t(t_2)}{D_t(t_1)} \neq \frac{D_{t'}(t_2)}{D_{t'}(t_1)}$$

In our first proof that I am thereby exploitable, we describe decision problems I might face at $t$ and $t'$, and we show that there are choices I could make in response to these problems that would leave me at least as well off at all times (including $t$, $t'$, $t_1$, and $t_2$) *and better off at* $t_1$ *and* $t_2$ than the choices that my discount functions $D_t$ and $D_{t'}$ will in fact lead me to make. In our alternative proof, we describe decision problems I might face at $t$ and $t'$, and we show that there are choices I could make in response to these problems that would leave me at least as well off at all times (including $t$, $t'$, $t_1$, and $t_2$) *and better off at $t$ and $t'$* than the choices that my discount functions $D_t$ and $D_{t'}$ will in fact lead me to make.

To see this, note that, if

$$\frac{D_t(t_2)}{D_t(t_1)} \neq \frac{D_{t'}(t_2)}{D_{t'}(t_1)}$$

then there are states of the world $s_1$ and $s_2$ such that I prefer $s_1$ to $s_2$ at $t$, but I prefer $s_2$ to $s_1$ at $t'$. That is, my discount functions lead me to switch my preferences over $s_1$ and $s_2$ between $t$ and $t'$.[82] And it is easy to show that anyone

[82] Suppose

$$\frac{D_t(t_2)}{D_t(t_1)} = r < r' = \frac{D_{t'}(t_2)}{D_{t'}(t_1)}$$

who switches their preferences between two states of the world from one time to another is exploitable, whatever their reason for doing so, whether it is because of their discount functions or something else completely. After all, at $t$, I will pay a little to be in state $s_1$ rather than $s_2$, and at $t'$, I will pay a little to be in $s_2$ rather than $s_1$. Thus, I will end up in state $s_2$, which I could have chosen for free at $t$ and retained for free at $t'$, but I will have paid a little at $t$ and again at $t'$ to achieve that.

Here's an example to illustrate: a friend has a spare ticket to see *Hamilton* this evening and a spare ticket to see *In the Heights* this evening. In the morning, I prefer *Hamilton*, but in the afternoon I change my mind and prefer *In the Heights*. As a result, in the morning, there's some amount of money I'll pay to receive the *Hamilton* ticket rather than receiving the *In the Heights* ticket for free, while in the afternoon, there is then some amount of money I'll pay to switch to the *In the Heights* ticket rather than retaining the *Hamilton* ticket at no further cost. In this case, if I'd just chosen the *In the Heights* tickets for free in the morning and stuck with that decision in the afternoon for free, I'd have been as well off at all times and better off in the morning and afternoon than I in fact am.

So, according to Ahmed and many in the economics literature, temporal discounting that satisfies Stationarity and violates Delay Consistency—thereby violating Time Consistency—is irrational because it leads you to switch your preferences over time, and such switching makes you exploitable. Thus, if Parfitian temporal discounting is irrational, and irrational for this reason, any decision theory that permits preference changes is unacceptable, because such preference changes are irrational.

### 13.7.2 Changing preferences and changing selves

Now, consider our Aggregate Utility Solution to the problem of choosing for changing selves. According to this, we set the utilities that we use in decision-making—the utilities that determine our decision-making preferences—

And let $\varepsilon = \frac{r'-r}{2}$. Then let:

- $s_1 = \langle r+\varepsilon, 0\rangle$
- $s_2 = \langle 0, 1\rangle$

Since $r < r+\varepsilon$,

$$s_2 = \langle 0,1\rangle \prec_t \langle r+\varepsilon, 0\rangle = s_1$$

But $r+\varepsilon < r'$, so

$$s_2 = \langle 0,1\rangle \succ_{t'} \langle r+\varepsilon, 0\rangle = s_1$$

as required. □

by weighting the local utilities set by our various past, present, and future selves. Now, in many cases, if, from one time to another, I change the weights I place on the utilities set by my past, present, and future selves, I will end up switching my decision-making preferences between two or more outcomes. And, as we have just seen, that will render me exploitable.

For instance, suppose I always give more weight to my current self than to any past or future self. Then, as time passes, the self that is my current self changes and with it the weightings I give to the various selves change. As a result, my decision-making preferences will also come to change. And that will leave me exploitable.

Suppose, for example, that Aneri has decided to become a police officer, but her values have yet to socialise to this role—she's only a couple of days into the training. She assigns more weight to her current self than to her future selves, and thus prefers a particular future in which she has some autonomy at work to one in which she doesn't. Later, however, once she has socialised her values more, and she assigns more weight to her current values than her past values, she switches her preferences over these two futures—she comes to prefer the future in which she has less autonomy at work. Though her weightings have changed in a very natural and principled way, they have rendered her exploitable in exactly the way that Parfitian discount functions will sometimes render you exploitable. If Ahmed is right, and exploitability entails irrationality, then the Aggregate Utility Solution will often render people irrational.

### 13.7.3 Equal weightings

In response, we might say that you are rationally required to retain the same weightings for your various selves throughout your lifetime. But on what basis might we fix these static weights? The only principled way to do it seems to be to assign the same weight to each self at every time in your life. However, if we do this, the Aggregate Utility Solution to the problem of choosing for changing selves becomes a sort of interpersonal egalitarian utilitarianism, and it thereby inherits many of the problems of that ethical doctrine.

One of these is that the decision-maker becomes alienated from their decisions (Railton, 1984; Williams & Smart, 1973). I am maximally well connected to my current self. I am barely connected at all to my future self in fifty years' time or to my past self twenty years previously—I share few memories, experiences, and beliefs with those selves, and often I have few character traits or values in common with them. If decision theory tells me that I must assign the same weights to them all, and use the resulting weighted average to make my decisions, I will feel that the dictates of that theory confront me "as an alien set of demands, distant and disconnected from [my] actual concerns", which is how Railton describes the analogous worry about utilitarianism (Railton, 1984, 135). Of course, in a sense, they are not distant or disconnected from *my* actual concerns, since these distant future and far past selves are *my* distant future selves and *my* far past selves. But that will be little solace to my current self, who does not recognise himself in them. Now, in the ethical case, there is a temptation to say, as Railton puts it:

> to have a morality is to make normative judgments from a moral point of view and be guided by them, and [...] by its nature a moral point of view must exclude considerations that lack universality. [So] any genuinely moral way of going about life would seem liable to produce the sorts of alienation mentioned above. Thus it would be a conceptual confusion to ask that we never be required by morality to go beyond a personal point of view, since to fail ever to look at things from an impersonal (or nonpersonal) point of view would be to fail ever to be distinctively moral—not immoralism, perhaps, but amoralism. (Railton, 1984, 138)

And, in response to Bernard Williams' example of a partly-fictionalised Paul Gauguin, who abandons his family to go to paint in Tahiti on the grounds that it better accords with his actual concerns than do the demands of morality that bind him to his home and family in France, readers are often moved to say that such a demand for authenticity, the antithesis of alienation, is sheer self-indulgence, and much closer to immoralism than the amoralism that Railton generously ascribes (Williams, 1981, 23). But, whatever your view of this response, the same cannot be said in the decision-theoretic case. There is neither conceptual confusion, nor immorality, nor amorality in requiring that our decision theory should not require us to go beyond the point of view of the decision-maker—namely, should not require us to give equal consideration to selves very remote from the self who is making the decision and selves to whom the decision-making self is closely connected.

### 13.7.4 On the alleged badness of exploitability

In any case, though, I think we should reject the inference from exploitability to irrationality.[83] To see why, it will be helpful to begin by thinking about another argument that trades on this alleged implication, namely, the so-called Dutch Book or sure loss argument for Probabilism. According to Probabilism (see Footnote 55 above). The Dutch Book argument for Probabilism starts with an assumption about which bets a single credence in a single proposition will lead you to make.[84] It then proceeds to show that, if your credences don't hang together as the probability calculus prescribes, then there is a series of bets in a series of propositions such that (i) your credence in each of these propositions will lead you to take the corresponding bet, but (ii) when taken together the bets result in certain loss—that is, the total amount your credences will lead you to pay out to participate in these bets will exceed the total amount you can win from them in any situation. Thus, just as your preferences at different times are exploitable if they change, your credences at a time are exploitable if they are not probabilistic. In the case of the Dutch Book argument, the option of not participating in any of these bets leaves you better off in all worlds than the option of participating in all of them, which is the option your credences will lead you to take. In the case of the switching preferences, where you prefer *Hamilton* in the morning and *In the Heights* in the afternoon, the option of taking the *In the Heights* tickets for free in the morning and sticking with them for free in the afternoon leaves you at least as well off at all times and sometimes better off than the option of paying to choose the *Hamilton* tickets in the morning and then paying again to switch to the *In the Heights* tickets in the afternoon. Thus, in both cases, we criticise an element of your mental state—your credences or your preferences—by constructing a particular decision problem and showing that it leads you to choose badly in that decision problem.

However, a natural response in both cases is to point out that I have no reason to believe that I will face the particular decision problem that witnesses my exploitability and that is thought to reveal my irrationality. I have no reason to think that a friend will have *Hamilton* tickets or *In the Heights* tickets, nor that, if they do, they would offer one for free and the other at a price I would find acceptable in the morning, and then the option

[83] For an alternative argument against this inference, and one that I also find compelling, see Sarah Moss' treatment of changing your mind (Moss, 2015).

[84] According to the Dutch Book argument, if you have credence $p$ in proposition $X$, then it will (or should) lead you to pay $pS$ utiles for a bet that will gain you $S$ utiles if $X$ is true and 0 utiles if $X$ is false.

to switch for a price I would find acceptable in the afternoon. That is a very specific pair of decision problems, one in the morning, one in the afternoon. Why think I will encounter it? I might encounter it, of course, and if I do, I will choose in a way that will surely be to my detriment. But why is this worse than, say, encountering a decision problem in which I will take a bet that I will in fact lose, but which I could have won had the world been different?

One way to put my point is this: it is irrational to choose one option over another if the first is guaranteed to leave you worse off than the second—that's just the principle of dominance, which is central to decision theory. But, while having credences that violate Probabilism might lead you to make such a choice—you choose to participate in all of the Dutch Book bets when you could have opted not to participate in none of them—having those credences isn't itself such a choice. That is, the Dutch Book argument does not show that there are alternative credences I might have had that would guide me better in the face of all decision problems. And in fact we can see that there never could be.

Suppose I'm 90% confident that Andi is less than 5ft tall, but only 40% confident that they are under 6ft tall. Then I violate the probability axioms, which require that I'm never more confident in a proposition than I am in another it entails. Now, because of my first credence, I'll buy from you for £9 a bet where you pay me £10 if Andi is under 5ft and £0 otherwise.[85] And I'll sell to you for £4 a bet where I pay you £10 if Andi is under 6ft and £0 otherwise. Then there are three cases. In all of them, I have paid you £9 for the first bet and received £4 from you for the second. So I start off £5 down. First case: if Andi is under 5ft, I pay you £10 and you pay me £10, so I end up £5 down overall. Second case: if Andi is between 5ft and 6ft, I pay you £10 and you pay me £0, so I end up £15 down overall. Finally, third case: if Andi is taller than 6ft, neither of us pay the other anything, so I end up £5 down overall. So I'm exploitable—the option to abstain from both bets would have left me neither up nor down, but my credences lead me to accept both bets, and taken together they leave me £5 down. But now pick any other credences in those two propositions. Then we can easily find some decision problem where my non-probabilistic credences will outperform them—that is, my credences will lead me to choose an option that ends up leaving me better off than the option that the alternative credences would

[85] Here we're assuming, for the purposes of illustration, that my utility is linear in money and that I'm risk-neutral.

have led me to choose.[86] For instance, if those alternative credences are 70% confident that Andi is less than 5ft tall, then my credences will lead me to accept to pay £9 for a bet that pays out £10 if Andi is less than 5ft tall and £0 otherwise, while the alternative credences will not; and, if Andi is indeed less than 5ft tall, that bet will leave me £1 up. And so on. The point is this: my non-probabilistic credences may lead me to choose a dominated option, but they are not themselves a dominated option—there are no alternative credences that are guaranteed to lead me to make better choices than my actual credences will.

The same, I claim, is true for changing preferences. Yes, if my preferences change then there is some decision problem I might face in which they will lead me to choose a dominated option. But while there are preferences I might have had that would not lead me to choose so poorly in that pair of decision problems, and indeed would never lead me to choose dominated options in any pair of decision problems, there are none that will always outperform mine. I might have preferred *Hamilton* in the morning and then retained that preference in the afternoon. Then I would pay in the morning for the *Hamilton* ticket and then stick with that choice for free in the afternoon. If I'd done that, I'd have done better by the lights of these alternative preferences and my actual morning preferences than my actual morning preferences lead me to do. But not so by the lights of my actual afternoon preferences. Or I might have preferred *In the Heights* in the morning and then retained that preference in the afternoon. Now, these preferences will outperform my actual preferences in the decision problem I actually faced—they'll lead me to choose the *In the Heights* ticket for free in the morning and then retain it for free in the afternoon. But they will perform worse in other decision problems. For instance, perhaps I am offered the choice between the tickets in the morning with no further choice to switch in the afternoon. Then I will actually pick *Hamilton*, but these alternative preferences will pick *In the Heights*. But that latter choice is worse by the lights of my actual morning preferences.

Now, there are ways to respond to these problems in the case of the Dutch Book argument. But, as we will see, there are no analogous moves in the case of changing preferences.

First, it is often said that it is unnecessary and indeed misleading to for-

[86] By the so-called Converse Dutch Book Theorem, if the alternative credences are themselves probabilistic, then there will be no decision problem in the face of which those credences will lead me to choose a dominated option. But there are certainly decision problems in the face of which those alternative credences will lead me to choose an option that leaves me worse off in some world.

mulate Dutch Book arguments in the dramatic mode, with a devious bookie offering you a series of bets guaranteed to lose you money. According to this response, nobody is claiming that being vulnerable to a Dutch Book renders you irrational because it will in fact lead you to make poor decisions. Rather, your vulnerability to a Dutch Book reveals an inconsistency in your attitudes, and such inconsistency is irrational (Armendt, 1993; Christensen, 1996). Reformulated in this way, the argument is sometimes called the *depragmatised Dutch Book argument*. Thus, according to this account, if I am 90% confident that Andi is under 5ft and 40% confident that they are under 6ft, then I have inconsistent attitudes to propositions about Andi's height—the possible Dutch Book against me reveals that, and it renders me irrational.

Now, that may save the Dutch Book argument, but it will not work in the case of changing preferences. After all, in the case of the theatre tickets, we already know that my preferences in the morning are inconsistent with my preferences in the afternoon. They are opposite to each other, for one thing. They cannot both be satisfied. We hardly need to establish that they are exploitable to discover that they are inconsistent. Our question, instead, is whether the fact that they are inconsistent renders them irrational. And the strategy of the depragmatised Dutch Book argument can't help us there.

Let's turn, then, to the second response to my objection to the move from exploitability to irrationality. We saw above that, while non-probabilistic credences lead to bad choices in the face of some decision problems, there are no alternative credences that outperform them in all. So non-probabilistic credences are not themselves a dominated option—there's no alternative that is guaranteed to serve you better. But perhaps there is some alternative does not dominate yours, but will nonetheless outperform you *on average over all the decision problems you might face*. Here's the idea, which draws on technical results by Mark Schervish (1989) and a suggestive presentation by Ben Levinstein (2017). The motivating question is this: how are we to calculate the pragmatic utility of having a particular credence in a particular proposition given a particular way the world might be? Roughly, the idea is this: for any bet on that proposition that you might face, and any way the world might be, the utility of your credence is the utility of the payout of that bet at that world if your credence would lead you to take the bet, and zero utility if it wouldn't. To give the utility of the credence given just the world, we average over all the different bets that you might face—given there are continuum-many different bets determined by their different payouts, there are many different ways to take the average, each determined by a measure over the unit interval. Now, we needn't go into the details here, but Schervish shows that any utility function for credences that is determined in this way is from a particular family of functions—it is a *strictly proper scoring rule*. And indeed he shows that all strictly proper scoring rules can be formed in this way. Having done that, we can then piggyback on a well-known theorem due variously to de Finetti (1974), Savage (1971), and Predd et al. (2009): If we take the utility of a set of credences to be the sum of the utilities of the individual credences, and if we measure the utility of an individual credence in the way just described as the average payout of the choices it leads you to make—and thus using a strictly proper scoring rule, as Schervish shows—then, if your credences are not probabilistic, then there are alternative credences such that, any every world, the utility of the alternatives is higher than the utility of yours. Thus, if your credences violate probabilism, then there is an alternative set of credences over the same propositions that will serve you better on average when it comes to making decisions—it is guaranteed that they will, on average, leave you better off than your current non-probabilistic credences. And this, I think, gives us good reason to have probabilistic credences.

Might we be able to give an analogous sort of argument in the case of changing preferences? I think we can't. The problem is that, in the credal case, we assume that there is a fixed utility function that supplies the utilities of the payouts of bets; and we use this to calculate the average utility of the bets a set of credences leads us into, and that then figures in the dominance argument. In the changing preferences cases, we do not have that single perspective from which to judge how well a decision has gone—as we saw above, different choices will look better whether I take the *Hamilton*-preferring morning perspective, or the *In the Heights*-favouring afternoon perspective.

I conclude, therefore, that exploitability does not entail irrationality. The mere fact that there is some decision problem in which my attitudes—whether credences or preferences—will lead me to choose irrationally is not sufficient to render those attitudes irrational. Something else is required. In the credal case, we can find that something else, due to Schervish's ingenuity. But in the case of switching preferences, we cannot.

This is good news for the Aggregate Utility Solution to the problem of choosing for changing selves. After all, as we saw above, that solution is likely to lead to a great deal of changing preferences as agents change the weights they assign to their various selves. It is also good news for the Parfitian discounter. And indeed these two are closely linked. According to Parfit, it is reasonable to give different weights to future selves in line with the different degree of connection between them and your current self.

Ahmed considered the implications of this in the very narrow case in which states of the world are specified by the amount of a given quantity—in our stock example, chocolate—that you receive at various future times. For his purpose, this was unobjectionable—he wished to show that, even in this sort of context, Parfit's account gives rise to irrational preferences. But now that the threat of irrationality has receded, we can see that Parfit's account applies quite generally and can be used by agents to set the weights they assign to future selves.

## 13.8 Appendix: Proofs of Lemmas 13.4.1, 13.4.2, and 13.4.3

**Lemma 13.4.1**

(I) *If Stationarity, then (Delay Consistency $\Leftrightarrow$ Time Consistency).*

*Proof.* Suppose Stationarity. Then we begin by showing that a violation of Delay Consistency gives rise to a violation of Time Consistency. Suppose there are $t, t_1, t_2, a$ such that

$$\frac{D_t(t_1 + a)}{D_t(t_1)} \neq \frac{D_t(t_2 + a)}{D_t(t_2)}$$

Then, if we let $t' = t + t_1 - t_2$, then, by Stationarity,

- $D_{t'}(t_1 + a) = D_t(t_2 + a)$
- $D_{t'}(t_1) = D_t(t_2)$

So

$$\frac{D_t(t_1 + a)}{D_t(t_1)} \neq \frac{D_t(t_2 + a)}{D_t(t_2)} = \frac{D_{t'}(t_1 + a)}{D_{t'}(t_1)}$$

as required.

Next, we show that a violation of Time Consistency gives rise to a violation of Delay Consistency. Suppose

$$\frac{D_t(t_1 + a)}{D_t(t_1)} \neq \frac{D_{t'}(t_1 + a)}{D_{t'}(t_1)}$$

Then, if we let $t_2 = t_1 + t - t'$, then, by Stationarity,

- $D_t(t_2 + a) = D_{t'}(t_1 + a)$

- $D_t(t_2) = D_{t'}(t_1)$

So
$$\frac{D_t(t_2+a)}{D_t(t_2)} = \frac{D_{t'}(t_1+a)}{D_{t'}(t_1)} \neq \frac{D_t(t_1+a)}{D_t(t_1)}$$
as required. □

**Lemma 13.4.2**

(II) *Delay Consistency ⇔ Time-Dependent Exponential Discounting.*

(III) *Delay Consistency + Stationarity ⇔ Time-Independent Discounting.*

(IV) *Time Consistency + Stationarity ⇔ Time-Independent Discounting.*

*Proof.* We begin by proving (II). First, we know that $D_t(t) = 1$. Second, if $D_t$ satisfies Delay Consistency, then for any interval $y$, there is a constant $K_y$ such that for any time $x$, $\frac{D_t(x+y)}{D_t(x)} = K_y$. Then set $x = t$, so that $D_t(t+y) = D_t(t)K_y = K_y$. So now we have that $K_y = D_t(t+y)$. So we have, for any $x, y$,
$$D_t(x+y) = D_t(x)D_t(y+t).$$

Thus,
$$\log D_t(x+y) = \log D_t(x)D_t(y+t) = \log D_t(x) + \log D_t(y+t).$$

Now, by tweaking Cauchy's characterisation of the functional equation $F(x+y) = F(x) + F(y)$, we can characterise the functional equation $F(x+y) = F(x) + F(y+t)$ (for fixed $t$). Cauchy showed that, if $F(x+y) = F(x) + F(y)$, then there is $c$ such that $F(x) = cx$. We can show that if, $F(x+y) = F(x) + F(y+t)$, then there is $c$ such that $F(x) = c(x-t)$. Thus, there is $c$ such that $\log D_t(x) = c(x-t)$. And so $D_t(x) = e^{c(x-t)}$. And so there is $\beta_t$ such that $D_t(t') = \beta_t^{t'-t}$, as required for (II).

Next, we prove (III). By Delay Consistency, for $t_1, t_2$, there are $\beta_{t_1}, \beta_{t_2}$ such that $D_{t_1}(t') = \beta_{t_1}^{t'-t_1}$ and $D_{t_2}(t') = \beta_{t_2}^{t'-t_2}$. Then, for all $a$,
$$\beta_{t_1}^{a} = \beta_{t_1}^{(t_1+a)-t_1} = D_{t_1}(t_1+a) = D_{t_2}(t_2+a) = \beta_{t_2}^{(t_2+a)-t_2} = \beta_{t_2}^{a}$$

So $\beta_{t_1} = \beta_{t_2}$, as required.

(IV) then follows from (II) and (III) and Lemma 13.4.1. □

**Lemma 13.4.3**

(V) *If Stationarity, then (Unexploitability* $\Leftrightarrow$ *Delay Consistency).*

(VI) *Unexploitability* $\Rightarrow$ *Time Consistency.*

*Proof.* We first prove (V). And we begin here by establishing the right-to-left direction. Suppose your discount functions obey Stationarity and obey Delay Consistency. Then, by Lemma 13.4.2, there is $\beta$ such that $D_t(t') = \beta^{t'-t}$. Now, suppose $t_i < t_j$ are times and $s_1$ and $s_2$ are states that only differ in the quantity of chocolate you receive after $t_i$ and $t_j$. Thus,

$$\begin{aligned} s_1 &= \langle t_1, q_1; \ldots, t_i, q_i; \ldots; t_j, q_j; t_{j+1}, q_{j+1}; \ldots; t_n, q_n \rangle \\ s_2 &= \langle t_1, q_1; \ldots, t_i, q_i; \ldots; t_j, q_j; t_{j+1}, q'_{j+1}; \ldots; t_n, q'_n \rangle \end{aligned}$$

Then

$$\begin{aligned} u_{t_i}(s_1) &= \textstyle\sum_{k=1}^{i} q_k + \sum_{k=i+1}^{j} \beta^{t_k - t_i} q_k + \sum_{k=j+1}^{n} \beta^{t_k - t_i} q_k \\ u_{t_i}(s_2) &= \textstyle\sum_{k=1}^{i} q_k + \sum_{k=i+1}^{j} \beta^{t_k - t_i} q_k + \sum_{k=j+1}^{n} \beta^{t_k - t_i} q'_k \\ u_{t_j}(s_1) &= \textstyle\sum_{k=1}^{i} q_k + \sum_{k=i+1}^{j} q_k + \sum_{k=j+1}^{n} \beta^{t_k - t_j} q_k \\ u_{t_j}(s_2) &= \textstyle\sum_{k=1}^{i} q_k + \sum_{k=i+1}^{j} q_k + \sum_{k=j+1}^{n} \beta^{t_k - t_j} q'_k \end{aligned}$$

So

$$\beta^{t_i - t_j}[u_{t_i}(s_1) - u_{t_i}(s_2)] = [u_{t_j}(s_1) - u_{t_j}(s_2)]$$

So, if you would choose $s_1$ over $s_2$ at $t_i$, you would do the same at $t_j$, and vice versa. Thus, if your discount functions recommend one decision at one time, they will recommend the same decision at a later time. So, if you are exploitable, then there is some option that you will choose in the light of one of your decision functions that is guaranteed to be worse than some other option you might have chosen. But that isn't the case. If

$$\begin{aligned} s &= \langle t_1, q_1; \ldots; t_n, q_n \rangle \\ s' &= \langle t_1, q'_1; \ldots; t_n, q'_n \rangle \end{aligned}$$

and

(i) $u(q_i) \leq u(q'_i)$ for all $1 \leq i \leq n$, and

(ii) $u(q_i) < u(q'_i)$ for some $1 \leq i \leq n$,

then $u_{t_k}(s) < u_{t_k}(s')$. Thus, your sequence of discount functions are not exploitable.

Next, we show the left-to-right direction. So we suppose that your discount functions obey Stationarity and disobey Delay Consistency. Then they disobey Time Consistency. And thus, there are $t < t' < t_1 < t_2$ such that

$$\frac{D_{t'}(t_2)}{D_{t'}(t_1)} = r' < r = \frac{D_t(t_2)}{D_t(t_1)}$$

Now we will describe three states $s_1, s_2, s_3$ such that $D_t$ prefers $s_2$ to $s_1$, and $D_{t'}$ prefers $s_3$ to $s_2$, but $s_1$ has at least as high utility at all times and higher utility at some times than $s_3$. Thus, we can offer you the following choices: first, $s_1$ or $s_2$; second, stick with current choice or switch to $s_3$. And we can see that, if you'd chosen $s_1$ initially and then stuck with your current choice, you'd be guaranteed to do at least as well and sometimes better than if you'd chosen $s_2$ and then $s_3$, which is what your discount function mandates. In $s_1$, $s_2$, $s_3$, you only receive a quantity of chocolate at two times, $t_1$ and $t_2$. Thus, we represent these states by pairs $\langle q_1, q_2 \rangle$ — this represents the state in which your receive $q_1$ blocks of chocolate at $t_1$ and $q_2$ blocks at $t_2$. Now, we first pick $\varepsilon < \frac{r-r'}{3}$. Then we define these three states as follows:

$$\begin{aligned} s_1 &= \langle 1, 1 \rangle \\ s_2 &= \langle 1 - rr', 1 + r' + \varepsilon \rangle \\ s_3 &= \langle 1 - r'\varepsilon, 1 + r' - r + 3\varepsilon \rangle \end{aligned}$$

Now, first, it is clear that $1 - r'\varepsilon < 1$ and $1 + r' - r + 3\varepsilon < 1$. And, what's more:

$$\begin{aligned} \frac{1}{D_t(t_1)} u_t(s_1) &= 1 + r \\ &< 1 + r + r\varepsilon \\ &= (1 - rr') + r(1 + r' + \varepsilon) \\ &= \frac{1}{D_t(t_1)} u_t(s_2) \end{aligned}$$

So $u_t(s_1) < u_t(s_2)$. And

$$\begin{aligned} \frac{1}{D_{t'}(t_1)} u_t(s_2) &= (1 - rr') + r'(1 + r' + \varepsilon) \\ &= 1 - rr' + r' + r'r' + r'\varepsilon \\ &< 1 - rr' + r' + r'r' + 2r'\varepsilon \\ &= (1 - r'\varepsilon) + r'(1 + r' - r + 3\varepsilon) \\ &= \frac{1}{D_{t'}(t_1)} u_t(s_3) \end{aligned}$$

So $u_{t'}(s_2) < u_{t'}(s_3)$. As required for (V).

(VI) then follows immediately from the same proof. $\square$

# Chapter 14

# The nearer the dearer

In the previous chapter, we noted that you will be exploitable if you ever switch your preferences between two options. And we noted that, if your weightings change from one time to another, your decision-making preferences will in all likelihood switch as well. We noted that one way to avoid such exploitable switching is to have static weights—if for each self in a history there is a fixed weight that we apply to it at any time, then our decision-making utilities will never change and our preferences will never switch. But which static weights should we assign? The only really principled way to do this would be to assign equal weights—the same weight for all selves at all times. In the previous chapter, we noted one issue with this: if you are constrained to impose such weights, you might well feel alienated from your decisions, just as it is often said you might feel alienated from your decisions if you were constrained to choose in line with a utilitarian theory on which each person's utility is weighted equally in the calculation of aggregate utility. In the end, we concluded that being exploitable does not indicate irrationality, and so the need to constrain weights in the way suggested disappeared and with it the threat of alienation. However, in order to motivate the consideration we wish to develop in this chapter, it is worth thinking about another reason not to impose the requirement of equal weights. We might call it *the Stoic objection*.

## 14.1 The Stoic Objection

When we first introduced the example of Aneri in the first chapter of this book, we briefly mentioned the empirical literature on 'socialised values' (Bardi et al., 2014). It is common to observe that people 'grow into' their situation and environment—their values change so that they better fit with the environment in which they find themselves; they change so that they assign among their highest utilities to the history they in fact inhabit. For instance, those who become police officers grow to have a utility function that assigns higher value to experiences they are likely to have in that job; those who study economics grow to have values that align with that discipline.

Now, prior to careful empirical study, it isn't clear whether the time spent in the profession causes individuals to change their values or whether there is a common cause that both leads them to take up this profession and to change their values in this particular way—perhaps those with latent conformist values are attracted to police work. However, the empirical work that we now have suggests that it's the former—the experience of training to be a police officer genuinely causes the change; if the person hadn't become a police officer, they wouldn't have changed their values in this way.

The important upshot of this empirical work for our purposes is that, in many cases, you can choose your future utilities. You can do this by putting yourself into a situation that will socialise your values in the desired way. But this raises the following problem if you assign equal weights to all selves. If every past, present, and future self at a given state of the world contributes equally to the decision-making utility at that world, then the Aggregate Utility Solution will often require you to pick whichever option will lead you to have utilities that assign highest value to the world in which you'll end up. So, for instance, it might require you to choose whichever career best socialises your values to match the state of the world in which you have that career. Similarly, it will recommend any decision you can take that will furnish you with values that are satisfied most by the society you expect yourself to live in.

To see this in action, let's look at the simplest case of all in which I have just three selves—one past self (at time $t_0$), my current self (at time $t_1$), and one future self (at time $t_2$). I must choose between two options $a$ and $b$. Option $a$ leads to state $s_a$ for sure, while option $b$ leads to $s_b$ for sure. In both states, the possible world is @, namely, the actual world. But the future utilities I assign to this world are different in the two states. I abuse notation and write $U^a_i(@)$ rather than $U^a_{s_a,i}$ for the utility I assign to the situation $a$ & @ at time $t_i$ in state $s_a$, where I choose $a$; and similarly, I write $U^b_i(@)$ rather then $U^b_{s_b,i}$ for the utility I assign to the situation $b$ & @ at time $t_i$ in state $s_b$, where I choose $b$. My utilities are as follows:

| $U^a_0(@)$ | $U^a_1(@)$ | $U^a_2(@)$ | $U^b_0(@)$ | $U^b_1(@)$ | $U^b_2(@)$ |
|---|---|---|---|---|---|
| $m$ | $m$ | $m$ | $m$ | $m$ | $n$ |

where $n > m$. Thus, if I pick option $a$, my utility in the actual world will remain exactly the same (at $m$), whereas if I pick $b$, I will socialise my values so that my utility in the actual world be greater than it was (at $n$). Then, if I assign equal weight to the three local utilities, the expected decision-making utility of $a$ is $\frac{m+m+m}{3}$, while the expected decision-making utility of $b$ is $\frac{m+m+n}{3}$. So I am required to choose $b$ over $a$. That is, I am required to choose to change my utilities so that they assign greater value to the world in which I find myself. Thus, those in more authoritarian societies should choose to value that life, if they can, while those in a misogynist society will do best by socialising their values so that they assign higher utility to that, and so on. Thus, the Aggregate Utility Solution with equal weights will agree with the Stoic that, where you can't beat them, you should join them; where you can't bend your environment to match your values, you should bend your values to match your environment.

Indeed, it is worse than that—it doesn't even recommend joining them only conditional on being unable to beat them. If the benefits of socialising your values are great enough, they will outweigh the benefits of bending the world to your current values. For instance, suppose this time that your actions can either change the world or change your utilities. Thus, suppose again that I must choose between two options $a$ and $b$. Again, option $a$ leads to state $s_a$ for sure, while option $b$ leads to $s_b$ for sure. But this time, the worlds in states $s_a$ and $s_b$ are different. In state $s_a$, the world is $w_a$, in which current oppressions continue, while in state $s_b$, the world is $w_b$, which is the less oppressive world that you could bring about as a result of your activism. Here are your utilities

| $U^a_0(w_a)$ | $U^a_1(w_a)$ | $U^a_2(w_a)$ | $U^b_0(w_b)$ | $U^b_1(w_b)$ | $U^b_2(w_b)$ |
|---|---|---|---|---|---|
| $m$ | $m$ | $k$ | $n$ | $n$ | $n$ |

where $m < n < k$. So, if you pick $a$, then the levels of oppression remain the same, but you socialise your values to them; if you pick $b$, then the levels of oppression decrease, but your values do not change. Then, again if we weight each self equally, the expected decision-making utility of $a$ is $\frac{1}{3}m + \frac{1}{3}m + \frac{1}{3}k = \frac{2m+k}{3}$, and the expected decision-making utility of $b$ is $\frac{1}{3}n + \frac{1}{3}n + \frac{1}{3}n = n$. Thus, if $k > 3n - 2m$, then $a$ is the rational option, and it's better to join them than to beat them. If your activism will only make modest gains (so that $n$ isn't much greater than $m$), or if you will successfully socialise your values (so that $k$ is significantly larger than $m$ and $n$), then you should choose to socialise.

In many cases, however, this conclusion is unpalatable, while in others it is monstrous. Whether or not I can make any significant dent in the oppression suffered in my society, it is intolerable to come to value that oppression; and it is certainly intolerable to prefer changing my preferences to match that society when I might instead have changed the society to match my preferences better.

Now, you might think we could simply avoid such extreme cases by demanding that any self with an impermissible utility function—such as a future self who has come to value the oppression around them—is assigned zero weight in the calculation of the decision-making utilities. And indeed we might wish to do that. But it is worth noting that there are less extreme cases that such a constraint will not address. While choosing to bend your values to match an oppressive society is monstrous, bending to match an environment that is not oppressive but simply very different from the one that you currently value most will not seem monstrous, but it might seem inauthentic, weak, and revealing of a lack of true commitment. We might naturally ask of such an agent: to what extent were those really your values if you were prepared to jettison them so easily to obtain values that would be more easily satisfied? Such individuals we often accuse of having 'sold out'. And our disapproval seems to grow with the distance between your current values and the values you choose to adopt in order to ensure that you have values that are better satisfied.

Notice that this is similar to the reason for which we rejected the Unchanging Utility Solution in Chapter 3. We rejected that putative solution because it recommends choosing whatever will result in having utilities at a future time that are best satisfied at the time at which you have them. In both cases, our worry is that the resulting decision theory makes it too easy to abandon your current values.

## 14.2 Proximity considerations

The foregoing suggests that it is often reasonable to assign greater weight to your current self than to past and future selves, and to assign greater weight to selves with values that more closely resemble those of your current self. Thus, we might imagine a measure of distance between utility functions, and we might propose that, other things being equal, a self should receive greater weight the closer its utility function lies to the utility function of my current self. The measures of distance that we characterised in Chapter 9 would be natural candidates for this role.

In fact, I think this proposal needs to be refined a little, for not all of our preferences and utilities are born equal when we face a particular decision problem. Utilities are defined on possible states of the world, and while some of those states of the world are relevant to the decision I'm facing, some are not; some are genuine possibilities given the options between which I'm choosing, while some are not. We might imagine two of my future selves: the first assigns the same utilities as my current self to the outlandish possibilities that are not relevant to my decision, but assigns very different utilities to the possibilities that are; while the second assigns the same utilities that I assign to the relevant possibilities, but diverges from me dramatically on the irrelevant possibilities. We surely wish to assign greater weight to the utilities of the second than to the first. Thus, I propose that it is the distance not between the full utility function of the future self and the full utility function of the current self, but rather between the relevant portion of the utility functions that determines the weight assigned.

What, then, is the relevant portion of a utility function? The natural answer it is the portion that assigns utilities to the states of the world to which you assign positive probability at the time of the decision, given the acts that are available to you in the decision problem. If you assign zero probability to a state of the world, it isn't clear why the distance between your future self's and your current self's utilities in that state should make any difference to your current decision. For instance, I assign no probability at all to a state of the world in which George Eliot is the Prime Minister of the UK during 2020-2024. And so, however highly or lowly any future self values this state of affairs is irrelevant to the weight I will assign to them when I set my decision-making utilities.

Now, recall: we motivated this consideration—the suggestion that you should assign weights to other selves in part on the basis of the proximity of their (relevant) utilities to your current (relevant) utilities—by noting how we judge people who conspire to change their values so that the values they end up with are better satisfied by the world. But it is also a consequence of the Parfitian line of thought that we have been pursuing in the second part of this book. On that line of thought, we care more about selves to which we have a stronger connection, and we assign greater weight to those selves about whom we care more. In Chapter 12, we considered the connection that is forged between two selves when one benefits from the actions of the other, particularly their sacrifices—and we explained why it is therefore permissible and perhaps sometimes mandatory to give weight to past selves. And in Chapter 13, we turned to the connection that exists between two selves that share a great deal of their experiences, beliefs, and memories—the sort of connection that typically diminishes the further apart those selves are in time, and thus gives rise to temporal discounting. In this chapter, we consider the connection that exists between two selves who share a great deal of their values, or at least their relevant values, and thus have utility functions (or portions of them) that lie not too far apart. This is part of what Parfit wishes to illustrate by his example of the Russian nobleman, which I paraphrase here (Parfit, 1984, 327).

> **Russian Nobles** Pyotr, a nineteenth-century Russian count, is due to inherit vast estates upon the death of his father. He is a socialist, and he fervently wishes to give the lands away to the peasants who currently farm them. So, he signs a legal document to the effect that, when his father dies, the ownership of his lands will transfer directly to the workers. And he ensures that the document can only be voided by his wife, Anna, the countess. Now, Pyotr has seen too many young socialists like himself lose their ideals and become bourgeois as they get older, and he fears the same will happen to him. So, he asks Anna to promise that if he later asks her to void the document, she will refuse. Anna agrees and makes the promise to her husband. As the years pass, Pyotr's fears are realised and he becomes bourgeois. When his father dies and the lands are destined to transfer to the peasants, in line with the legal document, Pyotr asks Anna to void it to allow him to become the owner of the lands. What should Anna do?

Parfit invites us to agree with him that Anna should not void the document, as her later bourgeois husband requests. But if that's right, we face a puzzle. We usually think that, if you make a promise to me, then while you incur an obligation to fulfil that promise, I can nonetheless cancel that obligation if I release you from the promise. Why, then, can the later bourgeois husband not release his wife from the obligation she incurred by making the promise to his younger socialist self? Only, Parfit thinks, because the bourgeois husband is in some ethically important sense a different person from the socialist husband, and thus unable to release Anna from the obligation, much as I am unable to release my father from a promise made to my mother. While it might be strictly true that the socialist and bourgeois husband are both selves that belong to the same person, there is another sense in which the socialist husband does not survive as the bourgeois husband. Parfit uses this conclusion to bolster his case that it is often not the relation of numerical personal identity between selves that is important in ethics—rather, it is the relation of connectedness. His famous fission and fusion cases show that the strength of connectedness between two selves is partly determined by the number and importance of shared memories, beliefs, etc (Parfit, 1971, 1976, 1984). The case of Pyotr and Anna, on the other hand, shows that it is also partly determined by the proximity of the values of one self to the values of the other self. Thus, following the thesis of this part of our book, which suggests that you consider assigning weights to your various selves based on the degree of your connectedness to them, this conclusion suggests that you should assign greater weight to selves whose relevant values lie closer to the relevant values of your current self.

## 14.3 Aneri's career

With this consideration in hand, we can think again about two of the examples from the start of the book: Aneri's decision whether to become a police officer or a conservation offer; and Fernando's decision whether or not to bind his future self at the point of his retirement to donating a certain portion of his pension payments to effective charities.

If you recall, Aneri has hitherto valued and currently values the sort of life that being a conservation officer would offer her—some autonomy, but also the ability to work as part of a team towards an important goal. She doesn't assign as great value to the life of a police officer. While she sees that others might reasonably value the conformism that it would require, she doesn't. However, she's read some psychological studies and these suggest that, in fact, if she becomes a police officer, her values will change and she will come to value that life more than she will value the life of a conservation officer if she chooses that instead. How should she choose?

Structurally, Aneri's decision is similar to the example above in which you have the choice either to bend the world to your utilities and reduce the oppression it contains, or bend your utilities to the world and make peace with that oppression. Here are Aneri's utilities:

| $U_0^{\text{Pol}}(\text{Pol})$ | $U_1^{\text{Pol}}(\text{Pol})$ | $U_2^{\text{Pol}}(\text{Pol})$ | $U_0^{\text{Con}}(\text{Con})$ | $U_1^{\text{Con}}(\text{Con})$ | $U_2^{\text{Con}}(\text{Con})$ |
|---|---|---|---|---|---|
| $m$ | $m$ | $k$ | $n$ | $n$ | $n$ |

where $m < n < k$. Thus, if she becomes a conservation officer, her utility in that state remains unchanged, whereas if she becomes a police officer, her utility in that state increases. Here are her expected decision-making utilities for becoming a police officer and a conservation officer (we abuse our notation for the weights in the same way we do for utilities and write, for instance, $\alpha_1^{\text{Pol}}$ for $\alpha_{\text{Pol},1}$):

$$\begin{aligned}
V(\text{Police}) &= \alpha_0^{\text{Pol}} U_0^{\text{Pol}}(\text{Police}) + \alpha_1^{\text{Pol}} U_1^{\text{Pol}}(\text{Police}) + \alpha_2^{\text{Pol}} U_2^{a}(\text{Police}) \\
&= (\alpha_0^{\text{Pol}} + \alpha_1^{\text{Pol}})m + \alpha_2^{\text{Pol}} k \\
V(\text{Con}) &= \alpha_0^{\text{Con}} U_0^{\text{Con}}(\text{Con}) + \alpha_1^{\text{Con}} U_1^{\text{Con}}(\text{Con}) + \alpha_2^{\text{Con}} U_2^{a}(\text{Con}) \\
&= \alpha_0^{\text{Con}} n + \alpha_1^{\text{Con}} n + \alpha_2^{\text{Con}} n = n
\end{aligned}$$

Now, on the basis of the considerations we explored in Chapters 12 and 13, Aneri might give positive weight to her past self (at $t_0$) because of benefits she's received from their actions; and on the basis of the considerations proposed in Chapter 13, she might assign less weight to her past and future selves (at $t_0$ and $t_2$, respectively) merely because they are distant from her in time, and thus share fewer mental states with her. But, as it turns out, it really only matters how much weight she assigns to her future self as a police officer. For it is easy to see that

$$V(\text{Pol}) > V(\text{Con}) \Longleftrightarrow \alpha_2^{\text{Pol}} > \frac{n-m}{k-m}$$

Let's suppose, for instance, that $m = 2$, $n = 6$ and $k = 18$. Then Aneri will choose to be a police officer if she assigns a weight greater than $\frac{6-2}{18-2} = \frac{1}{4}$ to her future utilities. Anything less and the leap in the utility that she assigns to conformity as a police officer is not sufficient to outweigh her disapproval of that life in the past and present. One upshot of this is that the rise in utility must be quite substantial, and discounting of the future self on the basis of distance in time and distance in values must be quite small in order to push Aneri towards the choice of which her current self would disapprove. Of course, in a more realistic model of the decision, most careers are chosen nearer the beginning of your life, and so the number of future selves who will be considered is much greater than the number of past and present selves. And so the difference between the weightings each receives and the weightings that the past and present selves receive might be greater while still pushing Aneri to be a police officer.

## 14.4 Fernando's pension

What about Fernando? Fernando must make a decision now about how his pension savings will be distributed when he retires. If he opts in to the scheme he is considering, he will receive 90% of those savings, while the remaining 10% will be donated to effective charities; if he opts out, he will receive 100% of those savings. Currently, Fernando wishes to opt in; but he knows that future Fernando, standing at the threshold of retirement, will wish to opt out. What should he do? His decision is similar to the one that Pyotr faces in the Russian Nobles case, when he first decides to have the document drawn up disbursing his lands to the peasants who work them. But there is no second party in this case, and no promises are made.

Structurally, Fernando's case is quite different from the cases we've been considering so far. In the first case we presented as part of the Stoicism Objection, I had control over my utilities, but not over the world; in the second case, and in Aneri's case, the agent concerned had control over their utilities and over the world. In this case, Fernando has no control over his future values, but he does have some control over the world.

Fernando has two options—opt in and opt out. Here are his utilities:

| $U_0^{\text{In}}(\text{In})$ | $U_1^{\text{In}}(\text{In})$ | $U_2^{\text{In}}(\text{In})$ | $U_0^{\text{Out}}(\text{Out})$ | $U_1^{\text{Out}}(\text{Out})$ | $U_2^{\text{Out}}(\text{Out})$ |
|---|---|---|---|---|---|
| $m$ | $m$ | $k$ | $n$ | $n$ | $n$ |

where $m > n$ and $l > k$. So in the past and the present, he prefers to opt in, whereas in the future he prefers to opt out. Then his decision-making utility for opting in is

$$\begin{aligned} U(\text{In}) &= \alpha_0^{\text{In}} U_0^{\text{In}}(\text{In}) + \alpha_1^{\text{In}} U_1^{\text{In}}(\text{In}) + \alpha_2^{\text{In}} U_2^{\text{In}}(\text{In}) \\ &= \left(\alpha_0^{\text{In}} + \alpha_1^{\text{In}}\right) m + \alpha_2^{\text{In}} n \end{aligned}$$

and for opting out

$$\begin{aligned} U(\text{Out}) &= \alpha_0^{\text{Out}} U_0^{\text{Out}}(\text{Out}) + \alpha_1^{\text{Out}} U_1^{\text{Out}}(\text{Out}) + \alpha_2^{\text{Out}} U_2^{\text{Out}}(\text{Out}) \\ &= \left(\alpha_0^{\text{Out}} + \alpha_1^{\text{Out}}\right) k + \alpha_2^{\text{Out}} l \end{aligned}$$

What Fernando chooses depends entirely on how much less weight he assigns to his future self (at time $t_2$). Since the values of that future self lie some distance from him—he is altruistic, while his future self is not—he likely assigns lower weight to that self than to his current self. And that suggests that he should choose to opt in and bind himself to making the donation in the future. And this, I take it, is what we take to be the right option.

Before we conclude this chapter, it is worth noting an appealing feature of the account we've been developing—a feature that Fernando's case reveals. Notice that there is another way to account for our intuition that Fernando should opt in. We might simply refer to the conclusion of the previous chapter and say that, when he makes the decision now, he is assigning less weight to his future, retirement-aged self, but not because the values of that self lie far from the values of his current self; rather, he assigns them less weight because that self lies far in the future, and thus shares few of his memories, beliefs, experiences, and so on. However, if that's the case, we would expect the weight to remain the same regardless of the values of $m$, $n$, $k$, and $l$—if the weight is determined entirely by connection of cognitive and experiential mental states, then the extent to which future Fernando values keeping his whole pension pot is irrelevant to the weight. And if that's so then there will be some $l$ large enough that the utility of opting in is swamped by the utility of opting out. And that seems wrong. Fernando will surely not be moved to acquiesce to his future self's wishes just because his future self has more extreme utilities, and indeed extreme in the opposite direction to his current self. Happily, the current proposal can accommodate this. As Fernando's future self's utility for not donating the 10% grows greater and greater, their utilities move further and further from Fernando's current utilities in that state of the world; and thus the weight assigned to that future self diminishes as well. The more extreme Fernando's future self becomes, the less weight they are assigned. These then balance out to ensure that Fernando opts in.

# Chapter 15

# I'll be glad I did it—so, I'll do it

We start with two examples:

> **Evening activities** I have a free evening ahead of me. Should I go for a run or should I stay home and watch a movie?

I reason as follows:

(P1) If I go for a run, I'll be glad I did it.

Therefore,

(C) I'll go for a run.

This seems pretty good reasoning. Compare that reasoning to the case of Deborah from the beginning of the book:

> **Deborah** has decided to have a baby, but she needs to decide when to try to become pregnant: now, or in three months' time. Currently, she has a virus, and she knows that, when people become pregnant while carrying this virus, their child will have an extremely high chance of developing a very aggressive cancer around the age of forty. However, if she becomes pregnant in three months' time, once her body is rid of the virus, there will be no risk to her child. Currently, she values having the child with the prospect of aggressive cancer very much less than she values having the child without. However, if she becomes pregnant now and has a child with that prospect, she will, most likely, form a bond with them so strong that she would value having that particular child, with its tragic prognosis, more than having

> any other child, including the one without that prognosis that she would have had if she had waited three months. After all, the alternative child would have been a different child, created from different gametes; they would not be the child with whom Deborah has formed the bond. When should Deborah try to become pregnant?

Deborah reasons as follows:

(P1′) If I become pregnant now, I'll be glad I did it.

Therefore,

(C′) I'll become pregnant now.

This seems pretty bad reasoning. Elizabeth Harman (2009) considers related cases and tries to identify the difference—what is it that makes one good reasoning and the other bad? I'll try to do the same here by appealing to the Aggregate Utility Solution to the problem of choosing for changing selves. I'll come to rather a different conclusion from Harman's.

## 15.1 What is 'I'll be glad I did it' reasoning?

Before we can start on this, we need to spell out the reasoning a little more carefully. The first stumbling block is that there are two readings of the conclusion, and correspondingly two forms of reasoning that might lead to them. Let's consider my plans for the evening. I say:

(P1) If I go for a run, I'll be glad I did it.

Therefore,

(C) I'll go for a run.

When I conclude that I'll go for a run, do I conclude that going for a run is mandatory, or merely that it is permissible? Let's suppose I want the stronger conclusion—I wish to conclude that going for a run is mandatory. Then we can see straight away that I must be suppressing a premise. After all, for all I've said, it might be that, if I watch the movie, I'll be glad I did that too. So being glad I chose an option can't be sufficient to make it mandatory. So if I wish to establish the stronger conclusion, I must really mean:

(P1) If I go for a run, I'll be glad I did it.

(P2) If I watch a movie, I'll wish I hadn't.

Therefore,

($C^+$) It's mandatory that I go for a run.

On the other hand, perhaps I wish only to establish the weaker conclusion, namely, that going for a run is a permissible option. In that case, prima facie at least, it might be sufficient that I'll be glad I did it. That is, I reason:

(P1) If I go for a run, I'll be glad I did it.

Therefore,

($C^-$) It's permissible that I go for a run.

Next, let's spell out what I mean when I say that I'll be glad I did something. I take it I mean that, after having done it, I assign greater utility to the world that has resulted from my choice than I assign to the world that would have resulted had I chosen the alternative. So, to establish the stronger conclusion, I reason:

(P1) If I go for a run, then afterwards I'll assign higher utility to having gone for a run than to having watched a movie.

(P2) If I watch a movie, then afterwards I won't assign higher utility to having watched a movie than to having gone for a run.

Therefore,

($C^+$) It's mandatory that I go for a run.

And to establish the weaker conclusion, I reason:

(P1) If I go for a run, then afterwards I'll assign higher utility to having gone for a run than to having watched a movie.

Therefore,

($C^-$) It's permissible that I go for a run.

We might even spell these out in more formal terms. In the sorts of situation Harman considers, there are two options, $a$ and $b$, and two states, $s_a$ and $s_b$, such that $a$ leads to $s_a$ for sure, and $b$ leads to $s_b$ for sure. So if I choose $a$, then the world will be $w_a$, my utility at time $t_1$ will be $U_1^a$ and my utility at time $t_2$ will be $U_2^a$, where $t_1$ is the present time at which I'm making the decision and $t_2$ is the future time after I've been for the run or watched the movie. The reasoning to the stronger conclusion then runs:

(P1) $U^a_2(w_a) > U^a_2(w_b)$

(P2) $U^b_2(w_a) \not< U^b_2(w_b)$

Therefore,

($C^+$) It's mandatory that I choose *a*.

While the reasoning to the weaker conclusion runs:

(P1) $U^a_2(w_a) > U^a_2(w_b)$

Therefore,

($C^+$) It's permissible that I choose *a*.

In our example, Deborah seems to use the second sort of reasoning—to reach the weaker conclusion—but not the first. She holds that, if she were to become pregnant now, she would later value having done so more than she would value waiting three months; but she also holds that, if she were to become pregnant in three months, she would later value having done so more than she would value having not waited. So, at most, Deborah concludes that it is permissible for her to become pregnant now—she doesn't conclude that it's mandatory.

## 15.2 Two problems with the reasoning

With these preliminaries out of the way, let's now turn to the reasoning itself, with a particular focus on the permissibility version. If we adopt the Aggregate Utility Solution to the problem of choosing for changing selves, it's pretty clear that both sorts of reasoning are invalid. There are two reasons for this: first, 'I'll be glad I did it' reasoning pays no attention to your current utilities; second, it pays attention to the wrong comparisons between future utilities. Of course, the mere conflict between that form of reasoning and the Aggregate Utility Solution does not tell for or against either. But, as we will see, when we use the Aggregate Utility Solution to diagnose the flaw in 'I'll be glad I did it' reasoning, the criticism strikes us as valid.

The following example illustrates the first problem in the case of the permissibility reasoning:

> **Book Choice 1** I'm choosing which of two books to read—one is a spy thriller, the other a romantic novel. I've read both before, and I know that, whichever I read, I'll get so caught up in it that,

> by the end, I'll assign a certain high utility to having read that book and a certain low utility to having read the other. Currently, I'd strongly prefer to read the romantic novel. Which should I read?

In this example, if I read the spy thriller, I'll be glad I did it—that is,

$$U_2^{\text{Spy}}(\text{Spy}) > U_2^{\text{Spy}}(\text{Rom})$$

But the expected decision-making utility of the romantic novel is

$$\text{EU}(\text{Rom}) = \alpha_1^{\text{Rom}} U_1^{\text{Rom}}(\text{Rom}) + \alpha_2^{\text{Rom}} U_2^{\text{Rom}}(\text{Rom})$$

while the expected decision-making utility of the spy thriller is

$$U(\text{Spy}) = \alpha_1^{\text{Spy}} U_1^{\text{Spy}}(\text{Spy}) + \alpha_2^{\text{Spy}} U_2^{\text{Spy}}(\text{Spy})$$

Now, we know

$$U_2^{\text{Rom}}(\text{Rom}) = U_2^{\text{Spy}}(\text{Spy}) > U_1^{\text{Rom}}(\text{Rom}) > U_1^{\text{Spy}}(\text{Spy})$$

And, since $U_2^{\text{Rom}}$ lies closer to $U_1^{\text{Rom}}$ than $U_2^{\text{Spy}}$ lies to $U_1^{\text{Spy}}$, we have $\alpha_2^{\text{Rom}} \geq \alpha_2^{\text{Spy}}$, and thus $U(\text{Rom}) > U(\text{Spy})$. And thus, the romantic novel is mandatory and the spy thriller impermissible, even though, if I choose the spy novel, I'll be glad I did it. The point is this: While I'll be glad I read the spy thriller if I do, the same is true of the romantic novel. Thus, you might think that they are symmetric from the point of view of my future self. But they are not symmetric from the point of view of my current self. My current self prefers the romantic novel. And this breaks the symmetry in its favour.

The next example illustrates the second problem in the case of the permissibility reasoning:

> **Book Choice 2** Again, I'm choosing which of two books to read—one is harrowing and bleak; the other frothy and uplifting. I've read both before, and I know that, as in the previous example, I'll get so caught up in whichever I choose that, after finishing it, I'll end up assigning higher utility to having read that one than I'll assign to having read the other one. But while I'll get very caught up in the bleak book, it will really sink my spirits, while the frothy book will really lift them. So, after reading the harrowing book, I'll assign a low utility to having read that one but an even lower utility to having read the frothy one, while after reading

> the frothy one, I'll assign a high utility to reading that and a lower utility to having read the harrowing one. Currently, I'm indifferent between the two. Which should I read?

So, if I read the harrowing book, I'll be glad I did it—that is,

$$U_2^{\text{Bleak}}(\text{Bleak}) > U_2^{\text{Bleak}}(\text{Froth})$$

The decision-making utility of choosing the frothy novel is

$$U(\text{Froth}) = \alpha_1^{\text{Froth}} U_1^{\text{Froth}}(\text{Froth}) + \alpha_2^{\text{Froth}} U_2^{\text{Froth}}(\text{Froth})$$

while the utility of the bleak novel is

$$U(\text{Bleak}) = \alpha_1^{\text{Bleak}} U_1^{\text{Bleak}}(\text{Bleak}) + \alpha_2^{\text{Bleak}} U_2^{\text{Bleak}}(\text{Bleak})$$

Now, since I'm currently indifferent between the two, we can assume that, whichever I choose, my future utilities lie the same distance from my current ones, and so:

- $\alpha_1^{\text{Froth}} = \alpha_1^{\text{Bleak}}$
- $\alpha_2^{\text{Froth}} = \alpha_2^{\text{Bleak}}$

But we also have:

- $U_1^{\text{Froth}}(\text{Froth}) = U_1^{\text{Bleak}}(\text{Bleak})$
- $U_2^{\text{Froth}}(\text{Froth}) > U_2^{\text{Bleak}}(\text{Bleak})$

Thus, $\text{EU}(\text{Froth}) > \text{EU}(\text{Bleak})$. So, the frothy novel is mandatory and the harrowing book is impermissible, even though, if I choose the harrowing book, I'll be glad I did it. The point is this: I'll be glad I read the bleak book if I do, and I'll be glad I read the froth if I do that. I'm currently indifferent between the two. So it initially looks as if my current and future preferences are symmetric between the two options. Surely in this sort of case, either is permissible. But when we look not at the comparisons between Froth and Bleak from the point of view of my future utilities having read the bleak book, nor from the point of view of my future utilities having read the frothy book, but the comparisons between Froth from the point of view of my frothy utilities having read the frothy novel and Bleak from the point of view of my bleak utilities having read the bleak novel, the symmetry is broken in favour of the froth.

In sum, then, 'I'll be glad I did it' reasoning is invalid. How your preferences will be at a later time can never be enough to determine the choice you should make at an earlier time, since your preferences at that earlier time might outweigh them. And, even if, at the time of the decision, you are indifferent between the two options, the order in which you place the options at the later time is not sufficient to determine your choice either, since your decision-making utility for a state of the world takes into account your utilities at that state of the world *for the possible world it contains*, but not your utilities at that state of the world *for any other possible world*—that is, the quantities you use to make your decisions do not care how your future utilities in the two possible worlds compare; they care how the future utility you'll have if you choose one way in the world that results from you choosing that way compares with the future utility you'll have if you choose the other in the world that results from you choosing that way.

This latter point is interesting, for it reveals that the Aggregate Utility Solution has a consequence that is revealed by the following example:

> **Book Choice 3** This time I'm choosing between a tale of two friends and a tale of a large group of friends. I've read both before, and I know something about how my future utilities will look depending on which I choose. First: if I read the tale of two friends, it will make me happy by reminding me of my closest friendship; but, having thought of that, I'll wish I'd read about an even more inclusive friendship, so I'll assign high utility to having read the book I did, but even higher utility to having read the other one. Second: if I read the tale of the large friendship group, it will get me down because my own large friendship group is scattered to the winds; but it will make me glad I didn't read the book about the two friends, since that would have brought me even lower, so I'll assign low utility to having read the book I did, but even lower utility to having read the other one. That is, whichever I choose to read, I'll end up preferring having read the tale of the large friendship group. Currently, I'm indifferent between them. Which should I choose?

According to the Aggregate Utility Solution, I should choose the tale of two friends even though I'm currently indifferent and, whichever I choose, I'll come prefer the tale of the large group. After all, we have:

- $\alpha_1^{\text{Duo}} = \alpha_1^{\text{Group}}$
- $\alpha_2^{\text{Duo}} = \alpha_2^{\text{Group}}$

- $U_1^{\text{Duo}}(\text{Duo}) = U_1^{\text{Group}}(\text{Group})$
- $U_2^{\text{Duo}}(\text{Duo}) > U_2^{\text{Group}}(\text{Group})$

That is, what is crucial for the decision is not whether, having chosen one, I prefer the other; it is whether having chosen one I assign it higher utility than I would assign to the other had I chosen that.

This consequence might look paradoxical at first blush. In certain situations, the Aggregate Utility Solution demands not only that we choose an option that will result in a possible world that we disprefer to the possible world that would have resulted had we chosen the other option; it sometimes demands that we choose an option such that *whichever option we picked*, we would have come to disprefer the world that results from the option we did in fact pick. To dispel the air of paradox, it is useful to distinguish between an option such that, once it is chosen, you come to think you should have chosen an alternative option, and an option such that, once it is chosen, you come to disprefer the world it has created to the world that the other would have created. A rule that leads you to choose the first sort of option is paradoxical, and it is what decision theorists try to rule out by demanding that decisions are ratifiable (Jeffrey, 1983, 2004). But it is the second sort of option that the Aggregate Utility Solution sometimes entails you should choose. And when we think through the examples in which it does entail this, they do not seem problematic. It is perhaps best illustrated by a case in which only taste is involved.

> **Taste pills** The scientists at the local Sense Perception Lab have developed two pills. Pill A makes me love lemon sorbet, but love dark chocolate ice cream even more; Pill B makes me hate lemon sorbet, and hate dark chocolate ice cream only slightly less. I am offered two menus when I dine at their cafeteria: Pill A for starter and lemon sorbet for dessert; or Pill B to start and dark chocolate ice cream for dessert. Which menu should you pick?

Whichever I choose, I'll prefer dark chocolate ice cream to lemon sorbet. But if I choose the first, I'll love lemon sorbet and get that, while in the second, I'll hate dark chocolate ice cream and get that. So it seems obvious I should choose the first menu with Pill A and lemon sorbet.

## 15.3 Deborah's choice

Having diagnosed the general problem with 'I'll be glad I did it' reasoning, let's now return to Deborah's case. Harman is particularly keen to understand what goes wrong with this reasoning, since she thinks it underlies faulty reasoning in the disability rights literature. I disagree with her analysis of that literature, but I hope that Deborah's example is one on which we can agree—it is bad reasoning. And the diagnosis in this case is the same as in the case of Book Choice 1 above, since the two cases are structurally identical. If Deborah becomes pregnant now, she'll assign the same utility to the situation that results—in which she has the child with the tragic prognosis—as she will assign if she becomes pregnant in three months' time to the situation that results from that—in which she has the child without the prognosis. Thus, at the later time, Deborah's utilities are the same for the two options. But at the earlier time, the current time at which the decision is to be made, her utilities are not the same. She prefers to have the child without the prognosis. And that breaks the symmetry and implies that she should have the child at the later time, the child without the prognosis. So we can agree with Harman that Deborah should wait before becoming pregnant.

Interestingly, though, we disagree with Harman about the first case—whether to go for a run or to stay home and watch a movie. Without a number of further premises, this reasoning is bad. For one thing, we must add that I do not prefer watching the movie now. Or, if I do, we must add that there is some longer period in the future where I value having gone for the run. And we must also add that the utility I will assign to having gone for a run after I do so will be greater than the utility I will assign to having watched a movie after I do that. Indeed, it's fair to say that the conclusion of our investigation here is that 'I'll be glad I did it' reasoning is always deeply mistaken. For it turns on a comparison that is irrelevant to the expected utility calculation. 'I'll be glad I did it' reasoning asks you to compare your future utility, should you choose one option, for the world you create by choosing that option and your future utility, should you choose that same option, in the world you would have created by choosing the alternative. That is, when we are considering the permissibility of $a$, it asks us to compare $U_2^a(w_a)$ and $U_2^a(w_b)$. What is relevant, however, is the comparison between your future utility, should you choose one option, for the world you created by choosing that option what you did and your future utility, should you choose the alternative, for the world you create by choosing that instead. That is, when we are considering the permissibility of $a$, it is relevant to compare $U_2^a(w_a)$ and $U_2^b(w_b)$.

## 15.4 Self-frustrating choices

Our diagnosis of the failure of 'I'll be glad I did it' reasoning also helps to dispel any mystery around what we might call 'I'll regret I did it' cases. Consider, for instance, the example of Cheragh from the beginning of the book:

> **Cheragh** is deciding whether or not to write the great novel that has been gestating in her imagination for five years. But she faces a problem. If she writes it, she knows she will come to have higher literary standards than she currently has. She also knows that while her own novel would live up to her current standards, it will not live up to these higher ones. So, if she writes it, she'll wish she'd never bothered. On the other hand, if she doesn't write it, she'll retain the same literary standards she has now, and she'll know her novel would have attained those standards. So, if she doesn't write it, she'll wish she had. Should Cheragh write her book?

This is structurally similar to an example that Bykvist (2006) describes in which we must choose whether or not to marry, but know that whichever way we choose we'll come to have preferences that favour the alternative. In 'I'll be glad I did it' cases, the decisions are what Hare & Hedden (2016) call *self-reinforcing*—make the choice and you'll come to think it was best. In 'I'll regret I did it' cases, they are what Hare and Hedden call *self-frustrating*—make the choice and you'll come to think it was worst.

The lesson from the discussion of 'I'll be glad I did it' reasoning is that self-reinforcing and self-frustrating cases do not pose any particular problems, and the reasons are the same in each case. When we calculate the expected utility of an option, we never include in that calculation the utility that, as a result of choosing that option, you will come to assign to any world that will not result from choosing that option. So it is of no relevance to your decision whether the preferences you will come to have as a result of choosing one option favour the world you have thereby brought about, or favour some other world. That fact never enters the calculation in question.

## Chapter 16

# The road ahead

In this book, I have presented an account of how we should choose when we recognise that our values have changed in the past or might change in the future, whether as a result of our actions, or as part of our development, or because of some external influence. In this final chapter, I wish to do two things. First, I will give a time-lapse recapitulation of the ground we have covered the book, building up the Aggregate Utility Solution to the problem of choosing selves one step at a time. Second, I will sketch avenues for future exploration.

## 16.1 The Aggregate Utility Solution summarised

The question of the book can be put as follows: we should make our decisions on the basis of our beliefs and our values; but our values change over the course of our life; so which values should we appeal to when we make our decisions? We begin, I suggest, by conceiving of such decisions as judgment aggregation problems, or problems of collective decision-making. You, the person making the decision, may be an enduring entity that exists equally on your fifth birthday and your fiftieth, but you are also a corporate entity composed of parts that we call selves—just as the University of Bristol is an enduring entity that existed equally in 1925 and 1976 and in 2018, but also a corporate entity that comprises the individuals who have belonged to it, who belong to it now, and who will belong to it in the future. When one of those selves makes a decision at the time at which they exist, they make their decision on behalf of this corporate entity that is me. They are, while they exist, the Chief Executive Officer of the corporation of which they are a part—the corporation Richard Pettigrew. This role passes from one self to another as the latter succeeds the former.

Granted this, the problem of choosing for changing selves becomes the following problem: how should a single self decide on behalf of this corporate entity to which they belong? Our answer comes in three parts, each building on the previous one.

First, we have to decide at what level we should aggregate the doxastic and conative attitudes of the various past, present, and future selves—at the level of preferences, or the level of expected utilities, or the level of utilities and credences? In Chapter 6, we argued that we should aggregate credences and utilities separately to give group credences and group utilities, and then combine those to give the group expected utilities and finally the group's preferences.

Second: we have to say how to aggregate credences and utilities. In Chapters 5, 7, and 9, we argued that the group's credences will be just the credences of the self who is making the decisions—that is, the current self—while the utility that the group assigns to a state of the world will be a weighted average of utilities that the various past, present, and future selves assign in that state to the way the world is in that state.

Third, we have to say how we are to determine the weights that we assign in this weighted average to the various selves. Here, we sought to apply Derek Parfit's approach. Parfit held that, in many cases in ethics, where philosophers have appealed to the relation of personal identity between selves at different times, it is really the relation of connectedness between selves that is important. In the second part of the book, we explored the suggestion that the weight your current self assigns to one of your past or future selves should be determined by the degree of connectedness between them. In particular, we explored three different ways in which two selves might be connected and thus three considerations that might guide you when you set your weights: the connection formed when one self benefits from the sacrifices of the other; the connection formed by shared doxastic, cognitive, and experiential states, such as memories, beliefs, cognitive architecture, and conceptual scheme; and finally the connection formed by sharing values in common.

And so we have come to endorse the view that, when you face a decision problem, you should choose an option that maximises your expected decision-making utilities when those are calculated by the lights of your current credences. That is, you should pick $a$ for which the following is maximal:

$$\begin{aligned}\mathrm{EU}(a) &= \sum_s P^a_G(s) U^a_G(s) \\ &= \sum_s P_p(s||a) \sum_{i=1}^{n} \alpha^s_i U^s_i(a \,\&\, w^s)\end{aligned}$$

Here, a state of the world $s$ specifies not only a possible world $w^s$, but also the utility functions $U^s_1, \ldots, U^s_n$ that you have at each time in that world. And your decision-making utility function $U^s_G$ is a weighted average of those utility functions (with weights $\alpha^s_1, \ldots, \alpha^s_n$). Each $\alpha^s_i$ might be determined in part by whether your self at $t_i$ made sacrifices from which you benefitted, in part by the degree of psychological connectedness between your current self and your self at $t_i$, and in part by the extent to which these two selves subscribe to similar values. This, I claim, is how we should choose for changing selves.

## 16.2 Questions for future selves

I think the account of choosing for changing selves that I have proposed in this book is correct, but I do not claim that it is complete. In the interests of space, I have made some simplifying assumptions that must be lifted in order to provide a complete account, and I feel sure that there I have not enumerated all of the factors that contribute to the connectedness between two selves, and that therefore may figure in a reasonable assignment of weights to selves. In this section, I will identify some of these gaps and point to how they might be filled.

In Part I, we made a number of simplifying assumptions concerning the decision-theoretic framework in which the Aggregate Utility Solution is stated. There are three main ones. First, we have assumed that the individual in question is risk-neutral; second, we have assumed that their credences and utilities at a given time can be represented faithfully by a single real-valued credence function and a single real-valued utility function; third, we have assumed that each state of the world contains the same number of selves. Let's take these in turn.

### 16.2.1 Sensitivity to risk

I offer you the following choice: £20 for sure, or a fair coin toss that will give you £100 if the coin lands heads and £0 if it lands tails. You choose the sure thing over the gamble. Intuitively, this is rationally permissible, and yet you know that the expected monetary value of the sure thing (£20) is less than the expected monetary value of the gamble (£50). This looks like a form of risk aversion—you favour the certainty of receiving a smaller amount over a gamble that might leave you much better off, but might also leave you empty-handed. The orthodox decision theorist traditionally responds to this by noting that rationality compels us to maximise expected *utility*, not expected *monetary value*. So, providing your utility is a concave function of money—so that money has diminishing marginal value, and the utility of £20 is greater than the average of the utility of £100 and the utility of £0—your decision will be rational by the lights of orthodox decision theory. According to this account, you are risk averse just in case your utility function is a concave function of money.

However, we cannot accommodate all intuitively rational risk-sensitive preferences within orthodox expected utility theory in the same way. Consider, for instance, the following decision problem, formulated by the French economist Maurice Allais (1953) in what has come to be known as the *Allais paradox*:

> **Allais** An urn sits before you. It contains 100 balls, numbered from 1 to 100. A ball will be drawn. You are offered a pair of choices whose payouts depend on the number of the ball drawn. The first is between options *A* and *B*; the second between options *C* and *D*. The following table specifies their payouts:

| | *A* | *B* | *C* | *D* |
|---|---|---|---|---|
| 1-89 | £1m | £1m | £0m | £0m |
| 90 | £0m | £1m | £0m | £1m |
| 91-100 | £5m | £1m | £5m | £1m |

> Thus:
>
> - *A* gives 89% chance of £1m, 1% chance of £0m, 10% chance of £5m.
> - *B* gives 100% chance of £1m.
> - *C* gives 90% chance of £0m, 10% chance of £5m.
> - *D* gives 89% chance of £0m, 11% chance of £1m.

Now, suppose you have utility function $U$. Then

$$\begin{aligned} & \mathrm{EU}(A) - \mathrm{EU}(B) \\ = & \ (0.89u(£1\mathrm{m}) + 0.01u(£0\mathrm{m}) + 0.1u(£5\mathrm{m})) - u(£1\mathrm{m}) \\ = & \ 0.01u(£0\mathrm{m}) - 0.11u(£1\mathrm{m}) + 0.1u(£5\mathrm{m}) \\ = & \ (0.9u(£0\mathrm{m}) + 0.1u(£5\mathrm{m})) - (0.89u(£0\mathrm{m}) + 0.11u(£1\mathrm{m})) \\ = & \ \mathrm{EU}(C) - \mathrm{EU}(D) \end{aligned}$$

Thus, according to expected utility theory, a rational individual must order $A, B$ in the same way as $C, D$. They must either prefer $A$ to $B$ and $C$ to $D$ or $B$ to $A$ and $D$ to $C$. However, many people report preferring $B$ to $A$, but $C$ to $D$. They are risk averse, and so prefer the certain £1m offered by $B$; but they are not so risk averse that they are not prepared to take the tiny extra risk offered by $C$ in order to open up the possibility of £5m. As we have just seen, these preferences cannot be accommodated in expected utility theory.

There are a number of ways to modify orthodox decision theory so that it can accommodate these preferences. On some, we continue to represent you as having just a credence function and a utility function, but ensure that either the credence function or the utility function encode your attitudes to risk. For instance, if you are risk-averse, we might capture that by skewing your credences so that they are pessimistic—that is, we represent you as assigning lower credence to the best outcome of a gamble than your evidence recommends, or we represent you as assigning higher credence to the worst outcomes (Machina & Schmeidler, 1992). Or, again if you are risk averse, we might capture that by skewing your utilities so that we exaggerate the utility of the worst outcomes in a gamble (Pettigrew, 2015a; Stefánsson & Bradley, ta). But the most perspicuous account of risk-sensitive individuals comes from Lara Buchak (2013) and recommends that we augment our representation of an individual so that it includes not only credences and utilities but also a risk function. Where orthodox expected utility theory says that you should maximise your expected utility—which is a combination of your credences and utilities—Buchak's risk-weighted expected utility theory says that you should maximise your risk-weighted expected utility—which is a combination of credences, utilities, and the outputs of the risk function.

Now, if we assume not that your past, present, and future selves are the sort of individuals that expected utility theory treats but rather that they are the sort that risk-weighted expected utility treats, a series of questions arise. First:

(Q1) At what level should we aggregate the diverse attitudes of your different selves to give the decision-making attitudes at the current time?

Should we aggregate their preferences? Their risk-weighted expected utilities? Their risk-weighted credences and their utilities separately, afterwards combining them to give the group's risk-weighted expected utility? Or their risk functions, their credences, and their utilities separately, afterwards combining all three as risk-weighted expected utility theory dictates?

I think the arguments of Chapter 6 tell in favour of the latter. But then the second question arises:

(Q2) How should we aggregate each of the components? As before, I think we should take your current credences to be the decision-making credences, and a weighted average of your local utilities to be the decision-making utilities, but I am less sure how we should determine the decision-making risk function. Perhaps it should be a weighted average of the individual risk functions. Does the minimal mutilation argument from Chapter 9 apply when we wish to aggregate risk functions?

I leave these questions for future work.

### 16.2.2 Imprecise credences and imprecise utilities

From the start, we have assumed that your doxastic state is represented by your credence function, which assigns to each state of the world and each act a single real number that measures the strength of your belief in that state of the world under the supposition that the act is performed. And, in Chapter 8, we explained how to make sense of the claim that our values can be represented by a single utility function, which assigns to each state and each act a single real number that measures the strength of your preference for being in that state having performed that act—though in this case, we noted that any positive linear transformation of this utility function would represent the utilities equally well.

Both of these are strong assumptions. They entail, for instance, that you either more confident that Germany will no longer be a member of the EU in 2040 than that France will no longer be a member, or less confident, or exactly as confident. And you might well feel that they simply don't make any of these judgments. It also entails that you are more, less, or exactly as confident that you will get the job for which you're applying if you put in ten more minutes on your CV as if you put in ten more minutes on your application form. But again, you often want to say that you don't make judgments as fine-grained as this.

Similar problems arise for the assumption that my values can be represented using single precise numerical values. In Chapter 8, we explained how to make sense of this assumption. First, we identify the outcome that you think is best—call it $o_{\text{best}}$—and the outcome that you think is worst—call it $o_{\text{worst}}$. Then pick two real numbers, $a$ and $b$—these are going to be the two ends of the scale on which you measure utilities. Now, given any outcome $o$—say the outcome in which I take an umbrella and it rains, or the outcome in which I vote Labour in a General Election and the Tories win—let $p_o$ be the probability such that you are indifferent between receiving outcome $o$ for sure, on the one hand, and entering into a gamble that gives you $o_{\text{best}}$ with probability $p_o$ and $o_{\text{worst}}$ with probability $1 - p_o$. Then your utility for $o$ measured on this scale is $(1 - p_o)a + p_o b$.

However, this account of how to represent utilities numerically assumes that there is such a probability $p_o$ that plays the role required of it. But suppose that there is not. There may be many probabilities $p$ such that you prefer $o$ for sure to the gamble between $o_{\text{best}}$ and $o_{\text{worst}}$ with probability $p$, and many for which you prefer the gamble to the sure thing, but none for which you are indifferent. Perhaps there is a set of such probabilities $p$ for which you have no preferences at all—you neither prefer the gamble with probability $p$ to the sure thing for these, nor prefer the sure thing to the gamble, nor are indifferent. In this case, we cannot make sense of a precise numerical representation of our values.

Decision theorists have responded to this concern by proposing to represent your beliefs at a given time not by a single credence function $P$ but by a set of them $\mathbf{P}$, and your values at a given time not by a single utility function $U$ but by a set of them $\mathbf{U}$. We'll call these sets your *credal representor* and your *utility representor*, respectively. Roughly speaking, the idea is this: if you make a particular credal judgment, then all of the credence functions in your credal representor $\mathbf{P}$ should also make that judgment. Thus, for instance, if you judge that, given I vote for a Labour candidate it's more likely that Labour will win than that the Tories will win, then, for every $P$ in your credal representor $\mathbf{P}$, $P(\text{Labour win}||\text{RP vote Labour}) > P(\text{Tories win}||\text{RP vote Labour})$. On the other hand, if, on the assumption of my vote for Labour, you are neither more nor less nor exactly as confident that Labour will win than that the Tories will win, then there will be $P_1$ in $\mathbf{P}$ that orders them one way, $P_2$ in $\mathbf{P}$ that orders them the other way, and $P_3$ in $\mathbf{P}$ that renders them equal. Thus, while each credence function in the credal representor orders the two states one way or the other or makes them equal, conditional on the act, the representor does not, for the representor only orders them a particular way if *every* credence function it contains orders them that way.

And the same goes for your utilities: if you prefer the outcome in which I vote Labour and Labour win to the outcome in which I vote Labour and the Tories win, then all of the utility functions $U$ in $\mathbf{U}$ should reflect that; if, on the hand, you neither prefer, disprefer, nor are indifferent between those two outcomes, there will be at least one utility function in $\mathbf{U}$ that prefers the first to the second, at least one that prefers the second to the first, and at least one that is indifferent between them. Again, the utility representor takes a stand on a judgment of value if, and only if, every utility function it contains takes that same stand.

If we represent each of our selves as having credal and utility representors, the question then arises:

(Q3) How should we aggregate the imprecise attitudes of our past, present, and future selves in order to give the decision-making attitudes for the current self?

The problem need not arise for the credal attitudes, since we can again take the credal attitudes of the current self to be the decision-making attitudes. Thus, no non-trivial aggregation is required. Your decision-making credal representor is your current self's credal representor. But we wish to aggregate the utilities. Your decision-making utility representor should be an aggregate of the local utility representors of your past, present, and future selves? How are we to effect this aggregation? Here are two suggestions.

First, we might pursue a minimal mutilation strategy of the sort that we used in Chapter 9 when we argued that, when the values of each of your different selves are represented by a single utility function, we should aggregate those values using a weighted average of those utility functions. The idea is this: first, define a measure of distance between the representations of the values of the different selves; second, identify a feature that an aggregate might have—for instance, it is a weighted average of the utility functions it is aggregating—and show that (i) if it does not have that feature, there is an alternative that does have the feature that is closer to every single one of the representations it is aggregating, and (ii) if it does have that feature, there is no such alternative. To do this, we need a natural notion of distance between sets of utility functions. In Chapter 9, we narrowed down the vast menagerie of possible distance functions between individual utility functions by giving principled reasons for favouring the standard Euclidean distance. Perhaps we can do likewise to narrow down the vast array of different ways in which we might measure the distance between two sets of utilities. Some examples of these possibilities: we might measure the distance between two utility representors $\mathbf{U}$ and $\mathbf{U}'$ to be the minimum distance between utility functions that they contain, though that has the consequence that any set that overlaps all of the sets to be aggregated will lie zero distance from them all, but not all such sets look like legitimate aggregates; or we might measure it to be the maximum distance from one point in $\mathbf{U}$ to the closest point in $\mathbf{U}'$, giving us the well-known *Hausdorff distance*; or we might measure it to be the average distance from points in $\mathbf{U}$ to points in $\mathbf{U}'$ given some suitable definition of average for such things; and so on.

To introduce our second suggestion, let me explain why decision theorists prefer to represent your values as a *set* of utility functions each of which assigns to a given outcome a *single precise numerical value*, rather than as a *single* utility function that assigns to an outcome a *set* or *range of numerical values*.[87] That is, when we notice that we cannot identify a single numerical value that measures my utility in outcome in which I vote Labour and they win, why don't we retain the idea that I am represented by a single utility function, but rather than taking it to assign *single numbers* to each outcome, as we have done before, rather take it to assign *sets* or *ranges* of numbers to each outcome; why instead embrace the idea that I am represented by a set of utility functions each of which assigns a single number to an outcome? The answer is that representing your values in the former way can miss out important features of your values that we can capture by representing them in the latter way.

For instance, suppose you make the following judgments over just four outcomes, $o_1, \ldots, o_4$:

$$o_1 \prec o_2 \prec o_3 \prec o_4.$$

Then, using the strategy from Chapter 8, we want to say that (i) your utility in $o_1$ is $a$ (the bottom of your utility scale) while your utility in $o_4$ is $b$ (the top of that scale), and (ii) your utility in $o_3$ is higher than your utility in $o_2$, and both lie between $a$ and $b$. But you say nothing else. Suppose we set $a = 0$ and $b = 1$ and we represent your values by a set-valued utility function. Then it would look like this:

| | $o_1$ | $o_2$ | $o_3$ | $o_4$ |
|---|---|---|---|---|
| $U$ | 0 | (0,1) | (0,1) | 1 |

Now suppose that, by contrast, I have the following preferences: $o_1 \prec o_2 \sim o_3 \prec o_4$. Then my set-valued utility function would also be as follows:

[87] This second suggestion arose from a discussion with Jason Konek in which he proposed doing something similar for credal representors in a different context.

| | $o_1$ | $o_2$ | $o_3$ | $o_4$ |
|---|---|---|---|---|
| $U$ | 0 | $(0,1)$ | $(0,1)$ | 1 |

Thus, we would have the same utility functions, but different values. I am indifferent between $o_2$ and $o_3$, and thus will never pay to have one over the other; you prefer $o_3$ to $o_2$, thus there will always be some amount (perhaps very small) of money you'll pay to receive outcome $o_3$ if the alternative is $o_2$. On the other hand, if I represent your values as a set of utility functions, we will represent me and you differently:

- $\mathbf{U}_{\text{you}} = \{U : 0 = U(o_1) < U(o_2) < U(o_3) < U(o_4) = 1\}$
- $\mathbf{U}_{\text{me}} = \{U : 0 = U(o_1) < U(o_2) = U(o_3) < U(o_4) = 1\}$

Thus, your set will contain

| | $o_1$ | $o_2$ | $o_3$ | $o_4$ |
|---|---|---|---|---|
| $U$ | 0 | 0.2 | 0.4 | 1 |

while mine won't; and my set will contain

| | $o_1$ | $o_2$ | $o_3$ | $o_4$ |
|---|---|---|---|---|
| $U'$ | 0 | 0.3 | 0.3 | 1 |

while yours won't. The point is that, by representing your values using a set of point-valued utility functions, I can represent certain comparative judgments you make—one outcome is better than another—that I can't represent using just the ranges of the utilities that you assign to them individually.

Now, when we represent your different selves at different times, it seems natural to represent each self using a set of utility functions. But, just as there are comparisons between the utilities an individual assigns to different outcomes that require me to represent that individual using a set of point-valued utility functions rather than a single set-valued utility function, so there are comparisons between the utilities that different selves assign to a single outcome that require me to represent a sequence of selves using a set of sequences of precise utility functions rather than a sequence of sets of precise utility functions.

Suppose, for instance, that you and I assign utilities only to three outcomes, $o_1, o_2, o_3$. My current self orders them as follows: $o_1 \prec o_2 \prec o_3$. So does my future self. And so do your current and future selves. However, the value I assign to $o_2$ increases between the current time and the future time, while the value you assign remains constant. If we represent each of us as a sequence of sets of utility functions—one set to represent our current values and one set to represent our future values—then we will represent us as the same: both of us will have utility representor $\mathbf{U} = \{U : U(o_1) < U(o_2) < U(o_3)\}$ at the current time and also at the future time. So this representation misses something out—it fails to represent the difference between us. So we do better to represent each of us as a set of *sequences* of precise utility functions, rather than as a sequence of sets of precise utility functions. Thus, we represent me by the following set (where we set $a = 0$ and $b = 1$):

$$\begin{aligned}\mathbf{U}_{\text{me}} = \{\langle U_1, U_2\rangle :& \\ & 0 = U_i(o_1) < U_i(o_2) < U_i(o_3) = 1 \text{ for } i = 1, 2 \\ & \qquad\qquad\qquad \& \ U_1(o_2) < U_2(o_2)\}\end{aligned}$$

And we represent you by this different set:

$$\begin{aligned}\mathbf{U}_{\text{you}} = \{\langle U_1, U_2\rangle :& \\ & 0 = U_i(o_1) < U_i(o_2) < U_i(o_3) = 1 \text{ for } i = 1, 2 \\ & \qquad\qquad\qquad \& \ U_1(o_2) = U_2(o_2)\}\end{aligned}$$

Now that we have this representation of the evolution of my values over the course of my life, we can see how to aggregate the values of my different selves using the technique we already have. Suppose we represent the values of my different selves by a set $\mathbf{U}^*$ of sequences of point-valued utility functions—each sequence contains a utility function for each time. Then for each such sequence $\langle U_1, \ldots, U_n\rangle$, we can simply aggregate the utility functions in that using weighted averaging, and take the aggregate values to be those represented by the set of these weighted averages. Thus, if $\mathbf{U}^*$ is the set of sequences, the aggregate is

$$\mathbf{U}^+ = \left\{\sum_{i=1}^{n} \alpha_i U_i : \langle U_1, \ldots, U_n\rangle \in \mathbf{U}^*\right\}$$

These, then, are two possible ways to aggregate the values of my past, present, and future selves when the values of some of those selves are not represented by a single utility function. Which of them we should choose I leave for future work.

# Bibliography

Aczél, J., & Wagner, C. G. (1980). A characterization of weighted arithmetic means. *SIAM Journal on Algebraic Discrete Methods*, *1*(3), 259–260.

Ahmed, A. (2018). Rationality and Future Discounting. *Topoi*.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica*, *21*(4), 503–546.

Armendt, B. (1993). Dutch Books, Additivity, and Utility Theory. *Philosophical Topics*, *21*(1).

Arneson, R. J. (1982). The Principle of Fairness and Free-Rider Problems. *Ethics*, *92*(4), 616–633.

Arrow, K. J. (1950). A Difficulty in the Concept of Social Welfare. *Journal of Political Economy*, *58*(4), 328–346.

Arrow, K. J. (1951). *Social Choice and Individual Values*. New York: Wiley.

Arrow, K. J. (1977). Extended sympathy and the possibility of social choice. *American Economic Review*, *67*(1), 219–225.

Bardi, A., Buchanan, K. E., Goodwin, R., Slabu, L., & Robinson, M. (2014). Value stability and change during self-chosen life transitions: Self-selection versus socialization effects. *Journal of Personality and Social Psychology*, *106*(1), 131–146.

Becker, G. S. (1998). *Accounting for Tastes*. Cambridge, Mass.: Harvard University Press.

Berntson, D., & Isaacs, Y. (2013). A New Prospect for Epistemic Aggregation. *Episteme*, *10*(3), 269–281.

Bognar, G., & Hirose, I. (2014). *The Ethics of Health Care Rationing: An Introduction*. Oxford: Routledge.

Brandt, R. B. (1992). *Morality, Utilitarianism, and Rights*. Cambridge University Press.

Bricker, P. (1980). Prudence. *Journal of Philosophy*, *77*(7), 381–401.

Briggs, R. (2009). Distorted Reflection. *Philosophical Review*, *118*(1), 59–85.

Briggs, R. (2015). Transformative Experience and Interpersonal Utility Comparisons. *Res Philosophica*, *92*(2), 189–216.

Briggs, R. A. (2017). Normative Theories of Rational Choice: Expected Utility. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Briggs, R. A., & Pettigrew, R. (2018). An accuracy-dominance argument for conditionalization. *Noûs*.

Buchak, L. (2013). *Risk and Rationality*. Oxford University Press.

Bykvist, K. (2003). The moral relevance of past preferences. In H. Dyke (Ed.) *Time and Ethics: Essays at the Intersection*. Dordrecht.

Bykvist, K. (2006). Prudence for changing selves. *Utilitas*, *18*(3), 264–283.

Carel, H., Kidd, I. J., & Pettigrew, R. (2016). Illness as Transformative Experience. *The Lancet*, *388*(10050), 1152–53.

Christensen, D. (1996). Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers. *The Journal of Philosophy*, *93*(9), 450–479.

Cohen, S. (2013). A defense of the (almost) equal weight view. In J. Lackey, & D. Christensen (Eds.) *The Epistemology of Disagreement: New Essays*, (pp. 98–120). Oxford: Oxford University Press.

D'Agostino, M., & Dardanoni, V. (2009). What's so special about Euclidean distance? A characterization with applications to mobility and spatial voting. *Social Choice and Welfare*, *33*(2), 211–233.

D'Agostino, M., & Sinigaglia, C. (2010). Epistemic Accuracy and Subjective Probability. In M. Suárez, M. Dorato, & M. Rédei (Eds.) *EPSA Epistemology and Methodology of Science: Launch of the European Philosophy of Science Association*, (pp. 95–105). Springer Netherlands.

de Finetti, B. (1974). *Theory of Probability*, vol. I. New York: John Wiley & Sons.

Dietrich, F., & List, C. (2015). Probabilistic Opinion Pooling. In A. Hájek, & C. R. Hitchcock (Eds.) *Oxford Handbook of Philosophy and Probability*. Oxford: Oxford University Press.

Dougherty, T., Horowitz, S., & Sliwa, P. (2015). Expecting the Unexpected. *Res Philosophica*, *92*(2), 301–321.

Doyle, J. R. (2013). Survey of time preference, delay discounting models. *Judgment and Decision Making*, *8*(2), 116–135.

Easwaran, K., Fenton-Glynn, L., Hitchcock, C., & Velasco, J. D. (2016). Updating on the Credences of Others: Disagreement, Agreement, and Synergy. *Philosophers' Imprint*, *16*(11), 1–39.

Elga, A. (2010). Subjective Probabilities Should Be Sharp. *Philosophers' Imprint*, *10*(5), 1–11.

Elkin, L., & Wheeler, G. (2016). Resolving Peer Disagreements Through Imprecise Probabilities. *Noûs, doi: 10.1111/nous.12143.*

Eriksson, L., & Hájek, A. (2007). What are degrees of belief? *Studia Logica*, *86*(2), 183–213.

Fishburn, P. C. (1977). Condorcet social choice functions. *SIAM Journal of Applied Mathematics*, *33*, 469–489.

Fleming, M. (1952). A cardinal concept of welfare. *The Quarterly Journal of Economics*, *66*, 366–384.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, *40*(2), 351–401.

Gaertner, W. (2009). *A Primer in Social Choice Theory*. Oxford: Oxford University Press.

Genest, C. (1984). A characterization theorem for externally Bayesian groups. *Annals of Statistics*, *12*(3), 1100–1105.

Genest, C., & Wagner, C. (1987). Further evidence against independence preservation in expert judgement synthesis. *Aequationes Mathematicae*, *32*(1), 74–86.

Genest, C., & Zidek, J. V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*, *1*(1), 114–135.

Gibbard, A., & Harper, W. (1978). Counterfactuals and Two Kinds of Expected Utility. In E. F. M. Clifford Alan Hooker, James L. Leach (Ed.) *Foundations and Applications of Decision Theory*, vol. 13a of *University of Western Ontario Series in Philosophy of Science*, (pp. 125–162). D. Reidel.

Goldman, A. (1976). Discrimination and Perceptual Knowledge. *Journal of Philosophy*, *73*, 771–91.

Greaves, H., & Wallace, D. (2006). Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind*, *115*(459), 607–632.

Griffin, J. (1986). *Well-Being: Its Meaning, Measurement and Moral Importance*. Oxford: Clarendon Press.

Hájek, A. (2008). Dutch Book Arguments. In P. Anand, P. Pattanaik, & C. Puppe (Eds.) *The Oxford Handbook of Corporate Social Responsibility*. Oxford: Oxford University Press.

Hall, N. (1994). Correcting the Guide to Objective Chance. *Mind*, *103*, 505–518.

Hall, N. (2004). Two Mistakes About Credence and Chance. *Australasian Journal of Philosophy*, *82*(1), 93 – 111.

Hammond, P. J. (1991). Interpersonal Comparisons of Utility: Why and How They Are and Should Be Made. In J. Elster, & J. Roemer (Eds.) *Interpersonal Comparisons of Utility*. London: Cambridge University Press.

Hare, C., & Hedden, B. (2016). Self-Reinforcing and Self-Frustrating Decisions. *Noûs*, *50*(3), 604–628.

Hare, R. M. (1989). Prudence and past preference: Reply to Wlodzimierz Rabinowicz. *Theoria*, *55*(3), 152–58.

Harman, E. (2009). 'I'll be glad I did it' reasoning and the significance of future desires. *Philosophical Perspectives*, *23*(1), 177–189.

Harman, E. (2015). Transformative Experience and Reliance on Moral Testimony. *Res Philosophica*, *92*(2), 323–339.

Harsanyi, J. (1977a). Morality and the Theory of Rational Behavior. *Social Research*, *44*(4), 632–56.

Harsanyi, J. (1977b). *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Hart, H. L. A. (1955). Are there any natural rights? *Philosophical Review*, *64*, 175–91.

Hedden, B. (2015). Does MITE Make Right? In R. Shafer-Landau (Ed.) *Oxford Studies in Metaethics*, vol. 11. Oxford University Press.

Hicks, A. (ta). Moral Uncertainty and Value Comparison. In R. Shafer-Landau (Ed.) *Oxford Studies in Metaethics*, vol. 13. Oxford: Oxford University Press.

Hild, M. (2001). Stable Aggregation of Preferences. Social science working paper 1112, California Institute of Technology.

Ismael, J. (2008). Raid! Dissolving the Big, Bad Bug. *Noûs*, *42*(2), 292–307.

Ismael, J. (2013). In Defense of IP: A response to Pettigrew. *Noûs*.

Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy*, *83*(5), 291–295.

Jeffrey, R. (1992). *Probability and the Art of Judgment*. New York: Cambridge University Press.

Jeffrey, R. (2004). *Subjective Probability: The Real Thing*. New York: Cambridge University Press.

Jeffrey, R. C. (1983). *The Logic of Decision*. Chicago and London: University of Chicago Press, 2$^{\text{nd}}$ ed.

Joyce, J. M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, *65*(4), 575–603.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge: Cambridge University Press.

Joyce, J. M. (2009). Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In F. Huber, & C. Schmidt-Petri (Eds.) *Degrees of Belief*. Springer.

Joyce, J. M. (2010). A Defense of Imprecise Credences in Inference and Decision Making. *Philosophical Perspectives*, *24*, 281–322.

Kemeny, J. (1959). Mathematics without numbers. *Daedalus*, *88*, 571–591.

Konieczny, S., & Grégoire, E. (2006). Logic-based approaches to information fusion. *Information Fusion*, *7*, 4–18.

Konieczny, S., & Pino-Pérez, R. (1998). On the logic of merging. In *Proceedings of KR'98*, (pp. 488–498).

Konieczny, S., & Pino-Pérez, R. (1999). Merging with integrity constraints. In *Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'99)*, (pp. 233–244).

Kopec, M., & Titelbaum, M. G. (2016). The Uniqueness Thesis. *Philosophy Compass*, *11*(4), 189–200.

Laddaga, R. (1977). Lehrer and the consensus proposal. *Synthese*, *36*, 473–77.

Lehrer, K., & Wagner, C. (1983). Probability amalgamation and the independence issue: A reply to Laddaga. *Synthese*, *55*(3), 339–346.

Leitgeb, H. (2016). Imaging all the People. *Episteme*.

Leitgeb, H., & Pettigrew, R. (2010). An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science*, *77*, 236–272.

Levinstein, B. A. (2012). Leitgeb and Pettigrew on Accuracy and Updating. *Philosophy of Science*, *79*(3), 413–424.

Levinstein, B. A. (2017). A Pragmatist's Guide to Epistemic Utility. *Philosophy of Science*, *84*(4), 613–638.

Lewis, D. (1980). A Subjectivist's Guide to Objective Chance. In R. C. Jeffrey (Ed.) *Studies in Inductive Logic and Probability*, vol. II. Berkeley: University of California Press.

Lewis, D. (1999). Why Conditionalize? In *Papers in Metaphysics and Epistemology*. Cambridge, UK: Cambridge University Press.

Lewis, D. (59). Causal Decision Theory. *Australasian Journal of Philosophy*, (pp. 5–30).

Lockhart, T. (2000). *Moral Uncertainty and its Consequences*. Oxford University Press.

MacAskill, W. (2016). Normative Uncertainty as a Voting Problem. *Mind*, *125*(500), 967–1004.

Machina, M. J., & Schmeidler, D. (1992). A More Robust Definition of Subjective Probability. *Econometrica*, *60*(4), 745–80.

Madansky, A. (1964). Externally Bayesian Groups. Memorandum rm-4141-pr, The RAND Corporation.

Miller, M. K., & Osherson, D. (2009). Methods for distance-based judgment aggregation. *Social Choice and Welfare*, *32*(575-601).

Mongin, P. (1995). Consistent Bayesian Aggregation. *Journal of Economic Theory*, *66*(2), 313–351.

Moss, R. H., & Schneider, S. H. (2000). Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment reporting. In R. Pachauri, T. Taniguchi, & K. Tanaka (Eds.) *Guidance Papers on the Cross Cutting Issues of the Third Assessment Panel of the IPCC*, (pp. 33–51). Geneva: World Meteorological Organization.

Moss, S. (2013). Epistemology Formalized. *Philosophical Review*, *122*(1), 1–43.

Moss, S. (2015). Credal Dilemmas. *Noûs*, *49*(4), 665–683.

Moss, S. (2018). *Probabilistic Knowledge*. Oxford University Press.

Nagel, T. (1978). *The Possibility of Altruism*. Princeton University Press.

Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.

Okasha, S. (2016). On the Interpretation of Decision Theory. *Economics and Philosophy*, *32*, 409–433.

Papineau, D. (2012). Can we be harmed after we are dead? *Journal of Evaluation in Clinical Practice*, *18*(5), 1091–94.

Parfit, D. (1971). Personal Identity. *Philosophical Review*, *80*, 3–27.

Parfit, D. (1976). Lewis, Perry, and what matters. In A. Rorty (Ed.) *The Identities of Persons*. Berkeley: University of California Press.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Pasternak, A. (2014). Voluntary Benefits from Wrongdoing. *Journal of Applied Philosophy*, *31*(4), 377–391.

Paul, L. A. (2014a). *Transformative Experience*. Oxford: Oxford University Press.

Paul, L. A. (2015a). Transformative Choices: Discussion and Replies. *Res Philosophica*, *92*(2), 473–545.

Paul, L. A. (2015b). What You Can't Expect When You're Expecting. *Res Philosophica*, *92*(2).

Paul, S. K. (2014b). Diachronic Incontinence is a Problem in Moral Philosophy. *Inquiry: An Interdisciplinary Journal of Philosophy*, *57*(3), 337–355.

Pettigrew, R. (2014). What chance-credence norms should not be. *Noûs*.

Pettigrew, R. (2015a). Risk, rationality, and expected utility theory. *Canadian Journal of Philosophy*, *45*(5-6), 798–826. Unpublished manuscript.

Pettigrew, R. (2015b). Transformative Experience and Decision Theory. *Philosophy and Phenomenological Research*, *91*(3), 766–774.

Pettigrew, R. (2016a). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Pettigrew, R. (2016b). Book Review of L. A. Paul's *Transformative Experience*. *Mind*, *125*(499), 927–935.

Pettigrew, R. (2017). Aggregating incoherent agents who disagree. *Synthese*.

Pettigrew, R. (taa). On the Accuracy of Group Credences. In T. S. Gendler, & J. Hawthorne (Eds.) *Oxford Studies in Epistemology*, vol. 6. Oxford: Oxford University Press.

Pettigrew, R. (tab). Transformative experience and the knowledge norms for action: Moss on Paul's challenge to decision theory. In J. Schwenkler, & E. Lambert (Eds.) *Transformative Experience*. Oxford: Oxford University Press.

Pettit, P., & List, C. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press.

Pigozzi, G. (2006). Belief merging and the discursive dilemma: an argument-based approach to paradoxes of judgment aggregation. *Synthese*, *152*, 285–298.

Plato (1997). *Plato: Complete Works*. Indianapolis: Hackett.

Predd, J., Seiringer, R., Lieb, E. H., Osherson, D., Poor, V., & Kulkarni, S. (2009). Probabilistic Coherence and Proper Scoring Rules. *IEEE Transactions of Information Theory*, *55*(10), 4786–4792.

Predd, J. B., Osherson, D., Kulkarni, S., & Poor, H. V. (2008). Aggregating Probabilistic Forecasts from Incoherent and Abstaining Experts. *Decision Analysis*, *5*(4), 177–189.

Quiggin, J. (1993). *Generalized Expected Utility Theory: The Rank-Dependent Model*. Kluwer Academic Publishers.

Raiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Reading: Addison-Wesley.

Railton, P. (1984). Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs*, *13*(2), 134–171.

Ramsey, F. P. (1931). Truth and Probability. *The Foundations of Mathematics and Other Logical Essays*, (pp. 156–198).

Rawls, J. (1975). *A Theory of Justice (revised edition)*. New York: Oxford University Press.

Rinard, S. (2015). A Decision Theory for Imprecise Probabilities. *Philosophers' Imprint*, *15*(7), 1–16.

Ross, J. (2006). Rejecting Ethical Deflationism. *Ethics*, *116*(4), 742–768.

Russell, J. S., Hawthorne, J., & Buchak, L. (2015). Groupthink. *Philosophical Studies*, *172*, 1287–1309.

Saari, D. G., & Merlin, V. R. (2000). A geometric examination of Kemeny's rule. *Social Choice and Welfare*, *17*, 403–438.

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Econometrica*, *15*(60), 243–253.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons.

Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Schervish, M. J. (1989). A general method for comparing probability assessors. *The Annals of Statistics*, *17*, 1856–1879.

Seidenfeld, T. (2004). A contrast between two decision rules for use with (convex) sets of probabilities: Gamma-maximin versus E-admissibility. *Synthese*, *140*, 69–88.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2010). Coherent choice functions under uncertainty. *Synthese*, *172*, 157–176.

Sen, A. (2017). *Collective Choice and Social Welfare*. Penguin, expanded edition ed.

Sepielli, A. (2009). What to do when you don't know what to do. In R. Shafer-Landau (Ed.) *Oxford Studies in Metaethics*, vol. 4, (pp. 5–28). Oxford: Oxford University Press.

Shimony, A. (1955). Coherence and the Axioms of Confirmation. *Journal of Symbolic Logic*, *20*, 1–28.

Staffel, J. (2015). Disagreement and Epistemic Utility-Based Compromise. *Journal of Philosophical Logic*.

Stalnaker, R. C. (1970). Probability and Conditionals. *Philosophy of Science*, *37*, 64–80.

Stefánsson, H. O., & Bradley, R. (ta). What is Risk Aversion? *British Journal for the Philosophy of Science*.

Stigler, G. J., & Becker, G. S. (1977). De Gustibus Non Est Disputandum. *The American Economic Review*, *67*(2), 76–90.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, *23*(3), 165–180.

Talbott, W. J. (1991). Two Principles of Bayesian Epistemology. *Philosophical Studies*, *62*(2), 135–150.

Thoma, J. (ta). Temptation and Prefence-Based Instrumental Rationality. In J. Bermudez (Ed.) *Self-Control, Decision Theory, and Rationality*. Cambridge University Press.

Thoma, J., & Weisberg, J. (2017). Risk writ large. *Philosophical Studies*, *174*, 2369–2384.

Titelbaum, M. G., & Kopec, M. (ta). When Rational Reasoners Reason Differently. In M. Balcerak-Jackson, & B. Balcerak-Jackson (Eds.) *Reasoning: Essays on Theoretical and Practical Thinking*. Oxford University Press. Ms.

Tollesfen, D. P. (2015). *Groups as Agents*. Polity.

Ullmann-Margalit, E. (2006). Big Decisions: Opting, Converting, Drifting. *Royal Institute of Philosophy Supplement*, *81*(58), 157–172.

van Fraassen, B. C. (1984). Belief and the Will. *Journal of Philosophy*, *81*, 235–56.

van Fraassen, B. C. (1995). Belief and the Problem of Ulysses and the Sirens. *Philosophical Studies*, *77*(1), 7–37.

van Fraassen, B. C. (1999). Conditionalization, A New Argument For. *Topoi*, *18*(2), 93–96.

Vineberg, S. (2016). Dutch Book Arguments. In E. N. Zalta (Ed.) *Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

von Neumann, J., & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press, 2$^{\text{nd}}$ ed.

Wagner, C. (1984). Aggregating subjective probabilities: some limitative theorems. *Notre Dame Journal of Formal Logic*, *25*(3), 233–240.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford University Press.

Wedgwood, R. (2017). Must rational intentions maximize utility? *Philosophical Explorations*, *20*, 73–92.

Williams, B. (1981). *Moral Luck*. Cambridge, UK: Cambridge University Press.

Williams, B., & Smart, J. J. C. (1973). *Utilitarianism: For and Against*. Cambridge, UK: Cambridge University Press.

Young, H., & Levenglick, A. (1978). A consistent extension of Condorcet's election principle. *SIAM Journal of Applied Mathematics*, *35*, 285–300.

Young, H. P. (1974). An axiomatization of Borda's rule. *Journal of Economic Theory*, *9*, 52–53.