Choosing for Changing Selves

Richard Pettigrew

December 11, 2017

Contents

Chapter 1

The problem of choosing for changing selves

This book is about how we should make decisions; it's about how we should 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 agent 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 to her socialised values very slightly better than her current values fit with the conservation role. 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 will wish that she had. Should Cheragh write her book?[3]

---

[1] For related examples, see (Bardi et al., 2014; Bricker, 1980; Bykvist, 2006; Ullmann-Margalit, 2006; Paul, 2014a).

[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 always wanted to be a train driver (Hare, 1989); for a discussion of the normative force of past preferences, see (Bykvist, 2003).

[3] For a related example, see Bykvist's example of someone contemplating marriage (Bykvist, 2006).

Deborah has decided to have a baby, but she needs to decide when to try to become pregnant: now, or in two weeks' 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 two weeks' 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 two weeks. 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 a £6,000 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?

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 retires, 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?[5]

[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, 2015).

[5]For a related example, see Parfit's Russian nobleman case (Parfit, 1984).

What these examples share in common is that, for the agent 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.[6] Or, it might happen as a result of external factors, such as a change in the ideologies that dominate in the culture in which the agent lives, 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 to have this experience.[7] Alternatively, an agent'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's 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 — I take it this is the central insight of belief-desire psychology. 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

[6]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).

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

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?

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 **??**, we will consider how our solution to the problem of choosing for changing selves might be adapted to alternative theories.

Chapter 2

The economists' orthodoxy: expected utility theory

Let me introduce expected utility theory using an example.[8] I have been learning to drive for some time. My test is only four weeks away. Should I practise or not? Here's how orthodox decision theory would have us make this decision. There are two actions 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*). So I take each option in turn, and I evaluate it. That is, I assign it a number that is going to measure how good I think it is as a way of getting me what I want. Thus, for instance, $V(\textit{Practise})$ measures my subjective assessment of practising as a means to my ends; $V(\textit{Don't Practise})$ does the same for not practising.

Let's first consider *Practise*. To assign its value 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, *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.[9] And simi-

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

[9] As we'll see below, it's not quite right to call it *the* real number that measures this, but

larly 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, the value $V(\textit{Practise})$ 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*:

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

Expected utility theory then says that I am rationally required to pick whichever of the options has the highest value as a means to my ends; or, if they both have the same value, 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 amongst the options that have maximal subjective expected utility.

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 decision that an agent faces, and provides a formal model of that decision, which we might call the corresponding *decision problem*. Decision problems contain representations of what the agent 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, 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:

---

let's indulge in this fiction for the moment.

- $\mathcal{A}$ is the set of possible actions.

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

- $\mathcal{S}$ is the set of possible states of the world.

  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 $\mathcal{S} = \{\text{License}, \text{No License}\}$.

- 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 states that represent 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 $P^a(s)$.[10] 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.

  We assume that, for each a in $\mathcal{A}$, $P(-||a)$ (or $P^a(-)$) is a probability function. That is, an agent's credences in each state of the world, taken together, sum to 1; and an agent's 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.[11]

- 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; the states that represent 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.

[10] 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; Joyce, 1999).

[11] For arguments that credences with this property are required by rationality, see (Hájek, 2008; Joyce, 1998; Pettigrew, 2016a), but also chapter 8.

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.[12]

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 6. 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.[13] 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

[12]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 5, however, even this assumption won't ensure that my favoured solution to the problem of choosing for changing selves is also partition invariant.

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

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.[14] 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 may currently be twice as hot as Irkutsk according to the Celsius representation, but it won't be relative to the Fahrenheit representation. 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 6, I will consider it in more detail; and in chapter **??**, 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 value 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.

[14] 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 $\frac{5}{9}$.

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 values b at least as much as 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 $a \prec b$.[15]

There are two interpretations of decision theory. I will call them the *realist interpretation* and the *constructivist interpretation*.[16] 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 a value 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

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

[16] I borrow the terminology from (Buchak, 2013). Okasha (2016) uses 'mentalistic' instead of 'realist', and 'behaviouristic' instead of 'constructivist'.

were to have these credences and utilities and were to determine their value function and preference ordering on the basis of them via (EU1) and (EU2). The realist, on the other hand, says that the credences and utilities are fundamental and they determine the preference ordering via (EU1) and (EU2) from above. 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 decision. When new evidence arrives, we change our credences first, and that then often determines a change in our preferences. Or 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 value, 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.

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

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

- *Imprecise decision theory* represents the same features of an agent's state that expected utility theory, but represents them differently (Elga, 2010; Joyce, 2010; Moss, 2015; Rinard, 2015). It represents an agent's credal state as a *set* of probability functions, rather than a single one, and represents her values as a *set* of utility functions, rather than as a single one. The idea is that the true features of an agent's doxastic state are those shared by all probability functions in the set, while the true features of her conative state are those shared by all utility functions in the set. Thus, she thinks s 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 she

values s & a more than s' & a' iff $U(s \,\&\, a) > U(s' \,\&\, a')$ for all utility functions U in the set that represents her conative state. This allows us to represent agents who don't have a determinate opinion as to whether s is more or less likely than s' given a. The set of probability functions that represents such an agent's doxastic state will include P such that $P(s||a) < P(s'||a)$ as well as P' such that $P'(s||a) > P'(s'||a)$. And it allows us to represent agents who don't have a determinate opinion as to whether s & a is more or less valuable than s' & a'. The set of utility functions that represents such an agent's conative state will include U such that $U(a \,\&\, s) < U(a' \,\&\, s')$ as well as U such that $U'(a \,\&\, s) > U'(a' \,\&\, s')$.

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 the agent's doxastic state and for each utility function from the set of utility functions representing the agent's 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 all 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 **??**.

- *Risk-weighted expected utility theory* represents further features of an agent's state (Buchak, 2013; Quiggin, 1993). It represents her attitudes to risk. It does this using a function r, which transforms the agent's credences. 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 **??**.

Having said that, we offer an argument in favour of expected utility theory in chapter 8. 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.

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 value functions; 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 4.

Chapter 3

Existing solutions to our problem

First, we consider three alternative solutions to the problem of choosing for changing selves. I will call them the *ur-utility solution*, the *higher-order utility solution*, and the *objective utility solution*.

3.1 The Ur-Utility Solution

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*.[17]

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 citric

[17] Nagel (1978) proposes a solution along these lines to a related problem.

> acidity of the lemon sorbet. So, clearly, when you ask me on Monday which you should deliver to me on Friday, 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.

Of course, the case we have just presented seems like a case of choosing for changing selves. Indeed, it looks analogous to the example of Erik from the introduction, who has to decide whether or not to buy a 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 ur-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 ur-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 different elements of it give you what you value.

How do we calculate the utility of a state of the world, according to the ur-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. In the first state of the world, I receive lemon on Monday and dark chocolate on Friday; in the second, these are reversed. Then, according to the ur-utility solution, my utility for the two states will be different, even though they contain the same bundles of commodities and my values don't change. The point is that my utility is determined by, for each time, how much the world gives me that, 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 ur-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 ur-utility solution in the cases we have been considering. However, that 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 conversation 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 ur-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 ur-utility solution, the second following on from the only 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 ur-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 ur-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 that has been gestating in her head 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 might be quite rational for Chinara to embark on writing the novel. 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. But the ur-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.

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 ur-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 ur-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 ur-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 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 Ur-Utility Solution to the problem of choosing for changing selves is that it ignores the first consideration completely.

3.2 The Higher-Order Utility Solution

The second potential solution to our problem that I would like to consider is the higher-order 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.

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. Quite often, we have preferences concerning which preferences we have; we assign different values to having various alternative sets of values. On the higher-order utility solution, as the name suggests, we represent these as our higher-order utilities. Our first-order utilities represent the values that we have concerning different states the world might be in 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 higher-order 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 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 higher-order 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, I should choose to maximise the sum of my first- and second-order utilities. Either way, in the case described, I should choose to adopt a child.

There are at least two problems with this sort of solution. 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 higher-order 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 higher-order 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?

The second problem with the higher-order utilities solution is that it assumes 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 15C, but disbelieve that it's below 30C, then I might respond by throwing out my belief that it's below 15C, or abandoning my disbelief that it's below 30C, 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.

3.3 The Objective Utility Solution

We come now to the third putative solution to the problem of choosing for changing selves — the objective utility solution.[18] Consider again the examples from the beginning of the book — Aneri, Blandine, Cheragh, etc. — where it seems initially as if the agents' values change over time. According to the objectivist, while there is a sense in which it is true that the *subjective*

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

values or utilities that each of these agents have change over time, what really happens is that their opinions about what is the one true *objectively* correct set of 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, for the objectivist, the components of an agent's conative state — her subjective values, desires, likes and dislikes — are all simply shadows cast by her doxastic states — in particular, her credences concerning what is objectively valuable.

Recall Blandine, for instance. 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. 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 will 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.

To see how this might solve the problem of choosing for changing selves, let's work through an example.

> **Cruise** Recall Erik from the beginning of the book, where he was considering a deal that his pension scheme is offering. Consider his friend, Effie, who is contemplating an alternative deal from her pension scheme. On this alternative, if Effie pays an extra £50 into her pension scheme now, she will receive a £20,000 around-the-world cruise when she is 70 years old. Her situation differs from Erik's. Erik valued the white knuckle theme park ticket now, but was very confident that he wouldn't as an elderly man. Effie, by contrast, feels rather indifferent to the cruise now — she could take it or leave it. But she knows that, by the time she's 70, she will have come either to love the idea of a cruise or to loathe it. She just doesn't know which of these two directions her values will take — indeed, she thinks each is equally likely

— but she does know that, whether or not she accepts this offer will make no difference at all. On the objectivist view we are considering here, Effie's situation might be modelled as follows:

- The set $\mathcal{A}$ of acts contains *Offer* and *No Offer*.
- The set $\mathcal{S}$ of states of the world contains *Cruise* and *No Cruise*.
- Effie knows that, if she accepts the offer, she'll receive the cruise ticket, and if she doesn't, she won't. Thus, if P is her current credence function:

$$P(\textit{Cruise}||\textit{Offer}) = 1 = P(\textit{No Cruise}||\textit{No Offer})$$

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

$$OU(\textit{Cruise} \,\&\, \textit{Offer}) = 9 \quad \text{and} \quad OU(\textit{No Cruise} \,\&\, \textit{No Offer}) = 5$$

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

$$OU'(\textit{Cruise} \,\&\, \textit{Offer}) = 1 \quad \text{and} \quad OU'(\textit{No Cruise} \,\&\, \textit{No Offer}) = 5.$$

- 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 agent's subjective utility for a situation, such as *Cruise* & *Offer*, is her subjective expectation of its objective utility. Thus, for instance,

$$\begin{aligned} &U(\textit{Cruise} \,\&\, \textit{Offer}) = \\ &\quad P(OU \text{ gives objective utilities})OU(\textit{Cruise} \,\&\, \textit{Offer}) + \\ &\qquad P(OU' \text{ gives objective utilities})OU'(\textit{Cruise} \,\&\, \textit{Offer}) = \\ &\qquad\quad 0.5 \times 9 + 0.5 \times 9 = 5 \end{aligned}$$

And similarly for *No Cruise* & *No Offer*:

$$\begin{aligned} &U(\textit{No Cruise} \,\&\, \textit{No Offer}) = \\ &\quad P(OU \text{ gives objective utilities})OU(\textit{No Cruise} \,\&\, \textit{No Offer}) + \\ &\qquad P(OU' \text{ gives objective utilities})OU'(\textit{No Cruise} \,\&\, \textit{No Offer}) = \\ &\qquad\quad 0.5 \times 5 + 0.5 \times 5 = 5 \end{aligned}$$

So, $V(\textit{Offer}) = 5 = V(\textit{No Offer})$.

This gives the credences and utilities and the resulting values assigned by Effie's current self. Let's now consider the credences and utilities of her two possible 70 year old selves: the first has come to love cruises, the second to loathe them. 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 loves cruises, 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 loathes cruises, 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:

$$\begin{array}{lllllll} V_1(\textit{Offer}) & = & U_1(\textit{Cruise \& Offer}) & = & 0.9 \times 9 + 0.1 \times 1 & = & 8.2 \\ V_1(\textit{No Offer}) & = & U_1(\textit{No Cruise \& No Offer}) & = & 0.9 \times 1 + 0.1 \times 9 & = & 1.8 \end{array}$$

And vice versa for U_2 and V_2:

$$\begin{array}{lllllll} V_2(\textit{Offer}) & = & U_2(\textit{Cruise \& Offer}) & = & 0.9 \times 1 + 0.1 \times 9 & = & 1.8 \\ V_2(\textit{No Offer}) & = & U_2(\textit{No Cruise \& No Offer}) & = & 0.9 \times 9 + 0.1 \times 1 & = & 8.2 \end{array}$$

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 *Offer* or *No Offer* will make no difference to which she will be. So:

$$P(P_1 \text{ gives credences at } 70 || \textit{Offer}) = 0.5 = P(P_2 \text{ gives credences at } 70 || \textit{Offer})$$

and

$$P(P_1 \text{ gives credences at } 70 || \textit{No Offer}) = 0.5 = P(P_2 \text{ gives credences at } 70 || \textit{No Offer})$$

> Now, to which value function should Effie appeal when she chooses between *Offer* and *No Offer*? Should she appeal to V, the value function of her current self? Or should she appeal to some aggregate of the value functions of her two possible future selves? In fact, it turns out that it doesn't matter. Effie's current value for each option — *Offer* and *No Offer* — is equal to her expectation of her future value for that option:
>
> $$V(\textit{Offer}) = 5 = 0.5 \times 1.8 + 0.5 \times 8.2 =$$
> $$P(P_1 \text{ gives credences at } 70||\textit{Offer})V_1(\textit{Offer}) +$$
> $$P(P_2 \text{ gives credences at } 70||\textit{Offer})V_2(\textit{Offer})$$
>
> And similarly for $V(\textit{No Offer})$.

Thus, there is no dilemma: we don't need to choose between using current or future values to make our decision; we can use either and they will agree.

Indeed this is no fluke that arises from how we assigned the numbers. On the objectivist view, it will hold whenever our agent's credences satisfy two putative principles of rationality — the Reflection Principle and the Independence Principle:

> **Reflection Principle (RP)** Suppose P is our agent's current credence function; and suppose $P_1, \ldots, P_n$ are all our agent's possible future credence functions at some particular future time t. Then, for any act a in $\mathcal{A}$,[19]
>
> $$P^a(-) = \sum_{i=1}^{n} P^a(P_i \text{ gives credences at } t)P_i^a(-)$$
>
> **Independence Principle (IP)** For any act a in $\mathcal{A}$, state s in $\mathcal{S}$, and possible objective utility function OU,
>
> $$P^a(s|OU \text{ gives the objective utilities}) = P^a(s)$$

The Reflection Principle says that our current credences should be our expectations of our 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,

[19] Recall: $P^a(X) = P(X||a)$ is the agent's credence in X on the supposition of a.

and I think each equally likely, then I should be 65% confident now that it will rain on Saturday. And so on.

The Independence Principle says that, on the supposition of performing any act a, the state of the world s and the identity of the objective utility function should be independent. This simply says that, how the world is cannot affect the objective utility function, nor vice versa. Given these two, we have:[20]

Theorem 3.3.1 *Suppose P is our agent's current credence function; and suppose $P_1, \ldots, P_n$ are all our agent's possible future credence functions at some particular future time t. And suppose that she satisfies the Reflection Principle and the Independence Principle. Then, for any act a in $\mathcal{A}$,*

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

where

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

The problem with this solution is that, while it can account for Effie's case, it cannot account for others. Consider, for instance, Erik. Currently, he assigns high subjective utility to the experience of the theme park. But he knows that, when he is 90, he will assign very low subjective utility to the same experience. Remember: according to the objectivist, subjective utility is just expected objective utility. Thus, Erik knows that his expectation for the objective utility of the theme park experience will change between now and when he is ninety. What's more, he knows exactly how it will change. Thus, Erik knows exactly how his credences will change over that time. Thus, he must violate the Reflection Principle, since that entails that, if you are sure that your credence function at some point in the future is P', and nothing more, then P' should be your credence function now.

Of course, we might reject the Reflection Principle. After all, it is a controversial claim. 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 belief does not oblige me to set my current credences in line with my future ones; surely I am not

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

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.[21]

The problem is that, if we abandon the Reflection Principle, the problem of choosing for changing selves returns. Should Erik use his current credences concerning objective utilities and the expectations of that objective utility that arise from those credences to make his decision? Or should he use his expectation of his future credences and the expectations of objective utility that arose from those credences instead? If he abandons the Reflection Principle, these will be different.

Here's an argument that he should use his current credences. While the Reflection Principle is no doubt false in full generality, there is a weaker version that is plausible:

> **Weak Reflection Principle (WRP)** Suppose P is our agent's current credence function; and suppose $P_1, \ldots, P_n$ are all our agent's possible future credence functions at some particular future time t. Suppose, furthermore, that our agent knows that, at t, her credence functions will be rational, and she will have acquired them from her 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 your future credences in the psilocybin case seemed wrong because you thought that the way in which your future credences would be formed would be irrational; second, when you instead think that your future credences will be rationally formed from your current ones, you should defer to your future credences because they will be at least as well informed as your current ones. So the Weak Reflection Principle captures what is right about the Reflection Principle, but jettisons what is wrong.

Now, if Erik satisfies this weaker version of the Reflection Principle, he must then judge his future credence functions to be irrational, or least have some doubt about their rationality — after all, if he thought they were rational, he'd defer to them, and he doesn't. Thus, he would be better off using his current credences in his decision making, since he has no such doubts about them. But, of course, if he does this, he'll take up the offer

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

from his pension scheme and his ninety year old self will be presented with a ticket to a white knuckle theme park. And, intuitively, that is the wrong decision.

Thus, the objectivist seems to face a trilemma. She could accept the Reflection Principle, in which case she can solve the problem of choosing for changing selves, but only on the basis of an implausible principle. Or she could accept the Weak Reflection Principle, in which case she can also solve the problem of choosing for changing selves, but this time she solves it in a way that runs contrary to our intuitions — she must say that Erik should accept the deal from his pension scheme. Or we could abandon both. I close this section by considering the third option. In the end, I think this is the option that the objectivist should take. It turns out that the problem of choosing for changing selves will reappear on this option. But, as I will argue in chapter 5, the objectivist can solve this problem by adapting the judgment aggregation solution to the subjectivist's version of that problem that I will defend in this book.

Why might we reject even the Weak Reflection Principle? To motivate our answer, consider the following sort of example. For much of the evidence we receive, it seems that there is more than one rational response to it. Consider a possible major financial event in the future — a crash on the FTSE, 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 saying 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 rationally 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*.[22] 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 shuffled repeatedly between the different possible rational responses to the evidence.[23] But

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

[23] This would be an epistemic version of what Richard Kraut calls 'brute shuffling' in the

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. For instance, suppose my evidence warrants any credence between 0.44 and 0.47 in the proposition that the FTSE 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 retained my credence of 0.45 at the start, then I would violate even the Weak Reflection Principle.

For this reason, the objectivist might abandon the Weak Reflection Principle. And she might say that, in cases like Erik's, what happens between now and when he is ninety is just that his credences concerning the objective utilities shift from one part of the space of rationally permissible responses to his evidence to another part. Thus, Erik is rational now; and he is rational when he is ninety; so he violates the Reflection Principle and its weakened version, but that is not a problem. Now, if the objectivist takes this line, the problem of choosing for changing selves arises again. But we can respond to that using the judgment aggregation solution that I defend in this book. So it is to that we now turn.

3.4 Appendix: Proof of Theorem 3.3.1

Theorem 3.3.1 *Suppose P is our agent's current credence function; and suppose $P_1, \ldots, P_n$ are all our agent's possible future credence functions at some particular future time t. And suppose that she satisfies the Reflection Principle and the Independence Principle. Then, for any act a in $\mathcal{A}$,*

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

where

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

case of intentions (Paul, 2014b, 344).

Proof.

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

as required. □

Chapter 4

The Judgment Aggregation Solution I: which attitudes to aggregate?

We have seen three putative solutions to the problem of choosing for changing selves: the Ur-Utility Solution, the Higher-Order Utility Solution, and the Objective Utility Solution. The first and second don't work, and the third needs work. The problem remains open. In this chapter, I formulate my own favoured solution. In fact, we will really be describing 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 *judgment aggregation solution* for reasons that will quickly become obvious.

The problem of choosing for changing selves arises because my past, current, and future selves do not all share the same values. Or, at least, this might be the case. So, perhaps better, the problem arises because 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 entities 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.[24] 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.[25] 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.[26] And so on. What is clear from these examples is that, in judgment aggregation problems, the sorts of judgments to be aggregated 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 the 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 and utilities are on the lowest, most fundamental level; the value function is on the next level, determined by the credences and utilities; and the preference ordering is on the highest level,

[24]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).

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

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

determined by the value function. So we might aggregate our agents' attitudes by aggregating their preference orderings (section 4.1). Or we might aggregate our agents' value functions (section 4.2) — this is known as the *ex ante* approach. Or we might aggregate our agents' credences and aggregate our agents' utilities, separately, and combine them to give their aggregate value function and their aggregate preference ordering (section 4.3) — this is sometimes called the *ex post* approach.[27] 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. In our setting, 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.

4.1 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 *Arrow Impossibility Theorem*.[28]

[27] 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.

[28] 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,

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

- **Weak Pareto** This says that, if all agents 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 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

Social Choice and Individual Values (Arrow, 1951).

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. Now, as we will see, there are serious proposals that violate No Dictator. For instance, Parfit (1984) considers the possibility that, in cases where our values have changed in the past or might change in the future, we ought to simply decide on the basis of our current self's preferences. But, as we will see in the second half of this book, that has unpalatable consequences. And, of course, the

case of Erik from the beginning of the book illustrates why. Intuitively, Erik should not simply choose on the basis of his current utilities. A dictatorship in which my current self is the tyrant errs in exactly the opposite way to the Ur-Utility Solution. Recall: our objection to the Ur-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.

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	Crabb	Crabb
Gove	Fox	**May**
May	**Leadsom**	**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	Ranger	Ranger
Actor	Carpenter	**Librarian**
Librarian	**Police**	**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 acts, $a_1, \ldots, a_n$. 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.

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 $\frac{4+5+5}{3} \approx 4.667$, while Leadsom's is $\frac{5+4+3}{3} = 4$. Relative to the second set, May scores $\frac{1+5+2}{3} \approx 2.667$, while Leadsom scores $\frac{5+4+1}{3} \approx 3.333$. 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 $\frac{5+3}{2} = 4$. Supposing I am completely ignorant which of the two selves will come about, thinking 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 $0.5\frac{5+1}{2} + 0.5\frac{5+1}{2} = 3$, 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, whilst 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 how 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 12, while May will receive 11. 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 judgment aggregation 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 my 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 gives 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. 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 value 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 value 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

value functions on which they are based, must assign the same aggregate preference ordering to both groups. But it seems that we might want their aggregate ordering to be different. For instance, if we just take straight averages of the values 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.

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 4.2 makes the group value for a particular act depend only on the individual values for that act, and the weighted average ex post method that we consider in section 4.3 makes the group credences and the group utilities for a particular act (and thus the group value 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 8, 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.

4.2 Aggregating value functions

So aggregating preference orderings won't work. Let's now consider how we might aggregate value functions instead, 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 agents 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 agent i has a value 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 agent 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 value function V_G to be a weighted arithmetic average of the individuals' value functions.[29] That is, we take a set of non-negative real numbers $\alpha_1, \ldots, \alpha_n$ — one for each agent in the group — such that they sum to 1 — that is, $\alpha_1 + \ldots + \alpha_n = 1$. And we define V_G, the value 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.

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 value 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: $P^a_G(-) = \sum_{i=1}^{n} \alpha_i P^a_i(-)$. And the group's utility function, U_G, is a weighted sum of the individual utilities: $U^a_G(-) = \sum_{i=1}^{n} \beta_i U^a_i(-)$. 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.

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, tab). 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 agents' credences in that proposition (Aczél & Wagner, 1980);

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

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

To its detriment:

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

(v) when all agents 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 8, where we offer a direct argument in favour of linear pooling for credences and utilities.

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:

> **Date night 1** 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_1) or Silvio's (act a_2). 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. Isaak is 70% confident in s_1 and 30% confident in s_2, while Jeremy is 40% confident in s_1 and 60% 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_1 & s_1 Silvio's Zitti Chicken	a_1 & s_2 Silvio's Meatballs Veg	a_2 & s_1 Thai Garden Zitti Chicken	a_2 & s_2 Thai Garden Meatballs Veg
Isaak	10	8	2	4
Jeremy	7	3	9	2

> Now, let's look first to the weighted average version of the ex post method. On this, we assign a weight, α, to Isaak's credences and the rest, $1 - \alpha$, to Jeremy's. And we assign a weight, β to

Isaak's utilities and the rest, $1-\beta$, to Jeremy's. Then their aggregate credence in s_1 is $0.7\alpha + 0.4(1-\alpha)$, while their aggregate credence in s_2 is $0.3\alpha + 0.6(1-\alpha)$. And their aggregate utility in a_1 & s_1 is $10\beta + 7(1-\beta)$; their aggregate utility in a_1 & s_2 is $8\beta + 3(1-\beta)$; and so on. And so, for instance, their aggregate value for going to Silvio's (act a_1) is:

$$\begin{aligned}\text{ExPost}_{\alpha,\beta}(a_1) \\ = \underbrace{[0.7\alpha + 0.4(1-\alpha)]}_{\text{Aggregate credence in } s_1} \times \underbrace{[10\beta + 7(1-\beta)]}_{\text{Aggregate utility in } a_1 \text{ \& } s_1} + \\ \underbrace{[0.3\alpha + 0.6(1-\alpha)]}_{\text{Aggregate credence in } s_2} \times \underbrace{[8\beta + 3(1-\beta)]}_{\text{Aggregate utility in } a_1 \text{ \& } s_2}\end{aligned}$$

Next, consider the weighted average version of the ex ante method. Here, we assign a weight, γ, to Isaak's value function and the rest, $1-\gamma$, to Jeremy's. Thus, their aggregate value for going to Silvio's (act a_1) is:

$$\text{ExAnte}_\gamma(a_1) = \gamma \underbrace{[0.7 \times 10 + 0.3 \times 8]}_{\text{I's value for } a_1} + (1-\gamma) \underbrace{[0.4 \times 7 + 0.6 \times 3]}_{\text{J's value for } a_1}$$

Now, it turns out that, if we set $\alpha = \beta = 0.5$ — that is, if we weight Isaak and Jeremy equally when we apply the ex post method — there is no γ such that:

- $\text{ExPost}_{\alpha,\beta}(a_1) = \text{ExAnte}_\gamma(a_1)$
- $\text{ExPost}_{\alpha,\beta}(a_2) = \text{ExAnte}_\gamma(a_2)$

So the weighted average ex ante method and the weighted average ex post method are different and 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 from above, the Weak Pareto condition: 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 agent i, then $a \prec_G b$.

Now, it is clear that the weighted average version of ex ante aggregation satisfies this.[30] However, the corresponding version of ex post aggregation does not. Let's see this in an example:

> **Date night 2** Suppose Ingrid and Jakob are making the same decision as Isaak and Jeremy. But Ingrid is 10% confident in s_1 and 90% confident in s_2, while Jakob is exactly the opposite — that is, he is 90% confident in s_1 and only 10% confident in s_2. Their utilities are given as follows:

	a_1 & s_1 Silvio's Zitti Chicken	a_1 & s_2 Silvio's Meatballs Veg	a_2 & s_1 Thai Garden Zitti Chicken	a_2 & s_2 Thai Garden Meatballs Veg
Ingrid	3	2	10	1
Jakob	4	1	3	2

> Then both Ingrid and Jakob assign higher value to a_1 than to a_2 — Ingrid assigns 2.1 and 1.9, respectively, while Jakob assigns 3.7 and 2.9 — but the aggregate obtained by the weighted average ex post method with equal weights $\alpha = \beta = 0.5$ for both individuals assigns higher value to a_2 then to a_1 — it assigns credence 0.5 to s_1 and s_2, and utilities as follows, which gives values of 2.5 to a_1 and 4.5 to a_2:

	a_1 & s_1 Silvio's Zitti Chicken	a_1 & s_2 Silvio's Meatballs Veg	a_2 & s_1 Thai Garden Zitti Chicken	a_2 & s_2 Thai Garden Meatballs Veg
Group	3.5	1.5	7.5	1.5

Notice how the phenomenon arises in our example. Ingrid and Jakob agree that going to Silvio's is better than going to Thai Garden, but the reasons behind their judgments are different. Jakob 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. Ingrid, by contrast, thinks that going to Thai Garden is much better than going to Silvio's in state s_1, where

[30] 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 $a \prec_G b$.

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 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? Usually, the most intuitively secure desiderata in judgment aggregation are the so-called *unanimity preservation principles*. As mentioned above, these are the principles that demand that, whenever the agents 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 agents 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 agents 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 agents 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.[31]

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

[31] 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 agents 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 agents agree that two outcomes have equal utility, then the group should agree that those two outcomes 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 Φ is a property of an agent'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 unanimity preservation principle would be Independence Preservation. But we don't want to enforce this for every property Φ. 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 credence p'. So consider the instance of UPP with the following property:

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

Then the antecedent of UPP_Ψ is satisfied, since Kacee and Lonnie both have property Ψ. So, their aggregate must have that property. That is, $P_G(\mathit{Leave}) = p$ or $P_G(\mathit{Leave}) = p'$. 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 Φ.

Thus, the question arises: for which properties Φ should we endorse the corresponding unanimity preservation principle? What properties of judgments should be preserved by aggregates when they are shared by all agents 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 agent but where such a shared property only reveals superficial agreement; that is, where is it possible for both agents to have that property for very different reasons. An example:

> **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 here? The point is that an agent's doxastic state is not exhausted by their credences. It also contains their evidence, which gives at least part of the agent's reasons for having the credences they do have. The problem for unanimity preservation principles is that, when agents 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. 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.

Another example:

> **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).[32] 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.[33]

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 preference orderings may mask deeper disagreement on the reasons behind those preference orderings, just as in our examples, 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

[32] Suppose $0 \leq \alpha_1, \ldots, \alpha_n \leq 1$ 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 $\alpha_1, \ldots, \alpha_n$ is $r_1^{\alpha_1} \times \ldots \times r_n^{\alpha_n}$.

[33] For a discussion of this problem with geometric pooling, as well as an exploration of the possibly ways of avoiding it, see (Pettigrew, taa, Section 9).

see that all agents 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 agents 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 agent'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 is not itself representable as 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 agents 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.[34] That is, there's no obligation to consider your group as an agent, 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 agent. 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

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

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 agent or not — they are simply coherence constraints on preferences.

4.3 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 agents' 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 agent 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 Rain*, *2mm of Rain*, . . . , *10mm of Rain*, and *No Rain*? There is an enormous range of possible levels. It is important that the value I assign to 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.[35]

As we noted in chapter 2, many decision theories ensure that no such instability can arise by demanding that an agent's utility at a coarser level of

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

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:

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

where $P^b(-) = P(-||b)$ and $U^b(-) = U(- \ \& \ b)$, as usual. And, in general:[36]

> **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 agents 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.

(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 agents such that both agents satisfy Inter-Grain Coherence and such that the group preference between a and b 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:

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

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.

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.

Chapter 5

The Judgment Aggregation Solution II: the solution itself

The version of the judgment aggregation solution that I favour is an ex post method. We aggregate the credences of my past, present, and future selves; and we aggregate the utilities of my past, present, and future selves; and then we combine these in the usual way to give the aggregate value 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 that it violates Weak Pareto. And, as we will see, it does. But I will argue that we can solve this problem.

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 the agent's current utilities, but also her 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 the agent's utilities at each time, but also her 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 the agent. 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 the credence function that our agent has 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 the agent's credence at t_i in s that world w is actual on the supposition that act a is performed.

  $\widehat{P_{s,i}}$ is the proposition that says that the agent has credence function $P_{s,i}$ at time t_i.

  For simplicity, we will assume that each state contains the same number of moments $t_1, ..., t_n$. In chapter **??**, we will ask what happens when we relax this assumption.

- $U_{s,i}$ is the utility function that our agent has 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 the agent's utility at t_i in s for the outcome in which w is actual and a is performed.

  $\widehat{U_{s,i}}$ is the proposition that says that the agent has utility function $U_{s,i}$ at time t_i.

- These three components — the possible world and the sequences of credence and utility functions belonging to the agent's successive selves — determine the state. Thus,

$$s = w^s \;\&\; \widehat{U_{s,1}} \;\&\; \ldots \;\&\; \widehat{U_{s,n}} \;\&\; \widehat{P_{s,1}} \;\&\; \ldots \;\&\; \widehat{P_{s,n}}$$

Of course, we know from chapter 2 that it doesn't make a lot of sense to say that $U_{s,i}$ is the agent's utility function, since an agent's 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 the agent's utility function at t_i in s is that we've picked a particular scale on which to measure our agent's utilities at all times and in all states and $U_{s,i}$ gives the agent's utilities measured on that scale. Now, there are potential problems here, since it is often claimed that we cannot compare the scales on which the utilities of different agents are measured, and it has been argued that the same is true even for different selves of the same agents Briggs (2015). We'll address that concern in chapter 6.

- $V_{s,i}$ is the value function that our agent has 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:

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

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 the various selves of the agent within these states. How, then, do we propose to aggregate these attitudes to give the agent's overall attitude at a particular time — that is, the attitude that she will use to make her decisions?

The first thing to note is that, so long as we assume the Reflection Principle, the principle of credal rationality that we introduced in section 3.3 above, there is no need to aggregate the credences of my different past, present, and possible future selves. Recall: the reflection principle says that your current credences should be your expectation of your future credences at a given time. If you are unsure what evidence you will receive between now and tomorrow, and perhaps even unsure of how you will respond to a given piece of evidence, you will be unsure what your credence function will be tomorrow. Nonetheless, your current credences should line up with your expectation of those future credences. So, if I think it's 50% likely that tomorrow I'll think it's 90% likely to rain; and if I think it's 50% likely that

tomorrow I'll think it's 30% likely to rain; then today I should split the difference and think it's 60% likely to rain. The idea is that I should treat my future self as an expert, for they will have more collected more evidence than I have. And the way we defer to experts is to set our credences equal to our expectation of theirs. For instance, if I hear the weather forecast in the background, but I don't hear everything they say clearly, and I think it's 50% likely that they said that it is 90% likely to rain, and 50% likely that they said that it is 30% likely to rain, then I should think it's 60% likely to rain, since that's my expectation of how likely they think it is to rain, and they're the expert. Recall: the general version of the reflection principle says:

> **Weak Reflection Principle (WRP)** Suppose P_p is our agent's current credence function; and suppose $P_{s_1,i}, \ldots, P_{s_n,i}$ are all our agent's possible future credence functions at some particular future time t_i. Suppose, furthermore, that our agent knows that, at t_i, her credences will be rational, and she will have acquired them from her credences by a rational process. Then, for any act a in $\mathcal{A}$,
>
> $$P^a_p(-) = \sum_{i=1}^{n} P^a_p(P_i \text{ gives credences at } t) P^a_i(-)$$

But of course this means that my current credence function is just a weighted average of my future credence functions. So there is no need to aggregate my future credence functions to give the group credence function for the group of my future selves — we can simply appeal to my current credence function, which already acts as such an aggregate, and we will do this henceforth. Thus, we let $P_G = P_p$.

In contrast, there is no reflection principle for utilities — that is, rationality does not require that my current utilities are my expectations of my future utilities.[37] Indeed, in many cases, they are not — for instance, his is so in the examples of Aneri or Cheragh or Erik from the start of the book, where it is clear there is nothing irrational about these changes in utility. 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

[37] Cf. (Harman, 2009) and (Hedden, 2015b, Section 4.2).

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.

Let us turn now to see how Hild's instability objection plays out in our context and for my favoured version of the judgment aggregation 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). 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. $U_0^a(W_1) = 2$).
- I have utility 1 for not becoming Sophie's parent (i.e. $U_0^b(W_2) = 1$)

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 some to think it's 75% likely that she's an introvert and 25% likely that she's an extrovert (P_1), or 25% likely that she's an introvert and 75% 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. $P_0(\widehat{P_1}) = 0.5 = P_0(\widehat{P_2})$).

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 Weak 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, my credence that she's an introvert 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. $U_1^a(w_1) = 6$);
- my utility for being Sophie's parent and her being an extrovert (w_2) will decrease to -10 (i.e. $U_1^a(w_2) = -10$).

And, mutatis mutandis, if my credence that she's an extrovert 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. $U_2^a(w_1) = -10$);
- my utility for being Sophie's parent and her being an extrovert (w_2) will increase to 6 (i.e. $U_2^a(w_2) = 6$).

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

- my utility for being Sophie's parent will be $U_i^a(W_1) = 2$, for each $i = 1, 2$.

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, 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 25% likely that she's an extrovert and 75% likely that she's an introvert (P_{11}), or I will think it's certain (i.e. 100% likely) that she's an introvert (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 75% likely that she's an extrovert and 25% likely that she's an introvert (P_{21}), or I will think it's certain (i.e. 100% likely) that she's an extrovert (P_{12}). And let's suppose that, at the five year point:

- $P_1(\widehat{P_{11}}) = \frac{1}{3}$, $P_1(\widehat{P_{12}}) = \frac{2}{3}$.
- $P_2(\widehat{P_{21}}) = \frac{2}{3}$, $P_2(\widehat{P_{22}}) = \frac{1}{3}$.

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 Weak Reflection Principle.

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 $U_{12}^a(w_2) = 2$;
- If my credence function is P_{11}, then my utility function is U_{11}, where $U_{21}^a(w_1) = 2$ and $U_{21}^a(w_2) = 0$;
- If my credence function is P_{11}, then my utility function is U_{11}, where $U_{22}^a(w_1) = -10$ and $U_{22}^a(w_2) = 6$.

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:

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

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:

- $s_{111} = w_1$ & $\widehat{U_0}$ & $\widehat{U_1}$ & $\widehat{U_{11}}$ & $\widehat{P_1}$ & $\widehat{P_{11}}$
- $s_{112} = w_2$ & $\widehat{U_0}$ & $\widehat{U_1}$ & $\widehat{U_{11}}$ & $\widehat{P_1}$ & $\widehat{P_{11}}$
- $s_{121} = w_1$ & $\widehat{U_0}$ & $\widehat{U_2}$ & $\widehat{U_{12}}$ & $\widehat{P_2}$ & $\widehat{P_{12}}$
- $s_{122} = w_2$ & $\widehat{U_0}$ & $\widehat{U_2}$ & $\widehat{U_{12}}$ & $\widehat{P_2}$ & $\widehat{P_{12}}$
- $s_{211} = w_1$ & $\widehat{U_0}$ & $\widehat{U_1}$ & $\widehat{U_{21}}$ & $\widehat{P_1}$ & $\widehat{P_{21}}$
- $s_{212} = w_2$ & $\widehat{U_0}$ & $\widehat{U_1}$ & $\widehat{U_{21}}$ & $\widehat{P_1}$ & $\widehat{P_{21}}$
- $s_{221} = w_1$ & $\widehat{U_0}$ & $\widehat{U_2}$ & $\widehat{U_{22}}$ & $\widehat{P_2}$ & $\widehat{P_{22}}$
- $s_{222} = w_2$ & $\widehat{U_0}$ & $\widehat{U_2}$ & $\widehat{U_{22}}$ & $\widehat{P_2}$ & $\widehat{P_{22}}$

Then

$$V(a) = P^a(s_{111})U^a(s_{111}) + \ldots + P^a(s_{222})U^a(s_{222})$$

Now, suppose we define $P_a(s_{ijk})$ as follows:

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,

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

Thus, my aggregate value for adopting is 0.

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

$$U_0^a(W_2) = U_i^a(W_2) = U_{ij}^a(W_2) = 1.$$

So my aggregate value for not adopting is $V(b) = 1$. And so, on the coarse-grained version, 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.

Of course, this is just a single case. Perhaps it is an anomaly? Let me now investigate how Hild's instability objection plays out in the general case. We will assume that our agent's 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 $P_{s,i}(-) = P_p(-|E_{s,i})$. And similarly for $P_{S,i}$. 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 S for states in $\mathcal{S}_1$ and lower case variables 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

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

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

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

(iv) 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.

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

Now, consider an act a in $\mathcal{A}$. Let's first of all consider the value of a

relative to the coarse-grained set of states, $\mathcal{S}_1$, which we'll write V_1:

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

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

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

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

Proposition 5.0.1 *Suppose*

$$P^a(w^s | w^S \mathbin{\&} \bigwedge_{j=1}^{n} \widehat{U_{s,j}} \mathbin{\&} \bigwedge_{j=1}^{n} \widehat{E_{s,j}}) = P^a(w^s | w^S E_{s,i})$$

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

From Proposition 5.0.1, we have: 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 $V_1(a) = V_2(a)$. But usually there will be acts a for which $V_1(a)$ and $V_2(a)$ are different. And, if they are different, then there is a value, m, that lies 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 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. Thus, we see that the problem of the partition sensitivity of our favoured solution is widespread.

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.[38] Indeed, this move fits well with the theme of this section that we should begin by aggregating an agent's reasons for having the attitudes she has, and not the attitudes that she bases 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 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 5.0.1(i), if we calculate our expected utility relative to that level, fine-graining any further won't make any difference — that is, our value 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

[38]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 **??**. See (Thoma & Weisberg, 2017) for a discussion of this feature of Buchak's theory.

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 our agent cares about is specified.

5.1 The Objective Utility Solution Redux

In section 3.3, we considered the Objective Utility Solution to the problem of choosing for changing selves. According to that putative solution, while an agent's subjective utilities do change over time, this is only because her credences concerning the true objective utilities change — indeed, an agent's subjective utility for an outcome is simply her subjective expectation of its objective utility.

Now, as we saw above, if we assume that our agent satisfies the Reflection Principle, the problem of choosing for changing selves evaporates. However, that problem itself demonstrates the implausibility of the Reflection Principle in this context — if we adopt the principle, we must consider agents like Blandine and Erik irrational, which is an unpalatable consequence. Thus, the problem of choosing for changing selves remains. Now that I have formulated my own favoured judgment aggregation solution, I think we can adapt it to provide a solution for the objectivist.

On my favoured solution, the aggregate value of an act is given as follows:

$$\begin{aligned} V_G^{\text{subj}}(a) &= \sum_{s \in \mathcal{S}} P_G^a(s) U_G^a(s) \\ &= \sum_{s \in \mathcal{S}} P_p^a(s) \sum_{i=1}^{n} \alpha_{s,i} U_{s,i}^a(w^s) \end{aligned}$$

Thus, we specify each state in sufficient detail that it tells us the agent's utility function at each time in that state, and we take the value of an act for me to be my subjective expectation of the aggregate subjective utility for that act, where the aggregate subjective utility for an outcome is a weighted average of the individual utilities for that outcome.

The natural way to adapt this for the objectivist is as follows:

$$\begin{aligned} V_G^{\text{obj}}(a) &= \sum_{s \in \mathcal{S}} P_G^a(s) U_G^a(s) \\ &= \sum_{s \in \mathcal{S}} P_p^a(s) \sum_{OU} P_G^a(OU \textit{ gives the objective utilities}) OU^a(w^s) \\ &= \sum_{s \in \mathcal{S}} P_p^a(s) \sum_{OU} \left(\sum_{i=1}^{n} \alpha_{s,i} P_{s,i}^a(OU \textit{ gives the objective utilities}) \right) OU^a(w^s) \end{aligned}$$

Thus, we specify each state in sufficient detail that it tells us the agent's utility function at each time in that state, and we take the value of an act for me to be my subjective expectation of its aggregate subjective utility, where the aggregate subjective utility for an outcome for me is my subjective expectation of its objective utility from the point of view of a weighted average of the credences of my past, present, and future selves concerning the objective utility.

Chapter 6

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 given 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 (i) a possible world, which details how things are outside her; (ii) the times within that world; (iii) the utility functions that she has at each of these times. 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.

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.

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.[39] 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 agent'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 agent 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 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

[39] 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.

that, at t_0 and t_1, Aneri's utility for *Police* & w_1 is 3, while her utility for *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 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 that.

I trust that the discussion above of Aneri's decision sufficiently motivates giving these numerical representations of her values, and giving them

on the same scale. Without this, we cannot say, for instance, how much more utility Aneri must get from her future point of view as a police officer and how much weight that future 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 to adjudicate trade-offs. We need to know how much worse it is to suffer a year with kidney failure than a year with medically-managed diabetes in order to allocate scarce resources (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 whether Caro values free access to wilderness spaces more than Don values free access to libraries in order to know which should be our priority for government funding (Sen, 2017).

6.1 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 value difference between watching musical theatre and watching rugby is greater than the value difference 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 case 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 know 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 using a 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 agent 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_{worst} be the outcome this agent values least;
- Let o_{best} be the outcome she values most;
- Pick two real numbers, $a < b$;
- Let her utility for o_{worst} be a and her utility for o_{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_{worst} and o_{best} that she considers exactly as good as getting o for sure.

  In particular:

  - Let p_o be the probability such that our agent is indifferent between o for sure, on the one hand, and p_o chance of o_{best} and $1 - p_o$ chance of o_{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 agent 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 agent'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 means by which to *measure* the strength of an agent'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 assume 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*. 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.

6.2 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 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 well, of course, is that, when restricted just to the outcome friend pairs (o, Ben), the ordering or indifference judgments agree with Ben's personal ordering; and similarly for Camille. If she does manage that, then we can construct measures of utility for Ben and Camille that are bound 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_{best} with outcome o_{best} with chance p_o and i_{worst} with outcome o_{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.

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 agents, 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 agent is indifferent between o_1 and o_2;

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

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

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

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 agent'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 agents.

In the interpersonal case, in which the agents 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 Hieronymous, 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 of everything above them. So Hieronymous' utility for having a lemon ice cream increases as he imagines worse and worse possible outcomes. Similarly, consider Kimmy, Hieronymous' 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 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 agent 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.

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. 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 above, 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 they 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 that the moral theories will rank differently, but which involve no violations of rights. 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 agents 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 quite a lot more insight into the mind of my past selves than I have into your mind or the minds of even my 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 socialises her values so that she becomes 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 4. Let's consider the finest-grained outcomes first. Since these outcomes

specify everything the agent cares about, it is actually rather unlikely that we'll find some finest-grained outcome that we value to the same extent before and after our values change. For instance, Aneri's present self values being a police officer less than her future officer-self. But each finest-grained outcome will specify whether or not she is a police officer. So she will value each such outcome in which she is a police officer more in the future, and less in the past. Similarly, for me making my decision whether or not to adopt Sophie. Each finest-grained outcome specifies whether or not I adopt Sophie. Thus, I will value any outcome in which I do more in the future when I am Sophie's 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:

$$U_0^a(\textit{Choc}) = P_0^a(\textit{Choc \& PO}|\textit{Choc})U_0^a(\textit{Choc \& PO}) + P_0^a(\textit{Choc \& CO}|\textit{Choc})U_0^a(\textit{Choc \& CO})$$

And her future police-self sets that utility similarly:

$$U_1^a(Choc) = P_1^a(Choc\ \&\ PO|Choc)U_1^a(Choc\ \&\ PO) + P_1^a(Choc\ \&\ CO|Choc)U_1^a(Choc\ \&\ CO)$$

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:

$$\begin{aligned} U_1^a(Choc\ \&\ PO) &> U_0^a(Choc\ \&\ PO) \\ U_1^a(Choc\ \&\ CO) &< U_0^a(Choc\ \&\ CO) \end{aligned}$$

But Aneri's future police-self is also certain that she chose to be a police officer, and not a conservation officer, so $P_1^a(PO) = 1$ and $P_1^a(CO) = 0$. So

$$\begin{aligned} P_1^a(Choc\ \&\ PO|Choc) &= 1 \\ U_1^a(Choc\ \&\ CO|Choc) &= 0 \end{aligned}$$

And thus, $U_1(Choc) > U_0(Choc)$. 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. Instead, when she says that her value for chocolate ice cream and her value for vanilla ice cream haven't changed, she is 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 agent's utility function would be her worst outcome and the unit would be the change of utility between that and her best outcome. If two utility functions U, U' that both represent an agent's utilities share the same zero, one is a scaling of the other: that is, there is α such that $U' = \alpha U$. If they share the same unit, one is a translation of the other: that is, there is β 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.

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 agent's values change, this usually gives rise to changes in the utility they assign to *every* finest-grained outcome, and also to every coarse-grained outcome. If I adopt Sophie and I come to value being Sophie's parent more, my values in all finest-grained outcomes in which I am her parent increase. But just as I might think a future self could judge when her utilities haven't changed from those of some recent past self, so we might think that they could make some judgments about how much they have changed. Suppose I can say, for six outcomes o_1, o_2, o_3 and o_1', o_2', o_3', 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_3, and likewise for o_1', o_2', o_3', then:

$$\begin{aligned} U_0(o_1) - U_0(o_2) &= U_0(o_2) - U_0(o_3) \\ U_0(o_1') - U_0(o_2') &= U_0(o_2') - U_0(o_3') \end{aligned}$$

This then fixes the utilities of $U_1(o_3)$ and $U_1(o_3')$ 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. 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 is equal to the difference between her present utility in that outcome and her present utility in the finest-grained outcome in which she eats vanilla ice cream and she is a police officer. This, then, is our solution to the problem of intrapersonal/interself utility comparisons.

Chapter 7

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 *Transformative Experience* book (Paul, 2014a).[40] 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 judgment aggregation 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 judgment aggregation solution as well.

[40]Material in this chapter is adapted from (Pettigrew, 2015b, 2016b, tac).

7.1 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. But secondly, she takes it to apply to what I will call the *deliberative* understanding of decision theory, as opposed to the *evaluative*. 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 agent 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. Paul's challenge applies primarily to the deliberative understanding of decision theory.

7.2 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 decision theory introduced above, 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*). According to the realist, to choose whether or not to apply, I must determine whether I prefer applying to not applying — that is, whether $Apply \preceq Don't\ Apply$ or $Apply \sim Don't\ Apply$ or $Apply \succeq Don't\ Apply$. 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 — that is, I must know $P(Parent||Apply)$, $P(Child\text{-}free||Apply)$, and so on. And I must know my utilities for the different possible outcomes — that is, I must know $U(Apply\ \&\ Parent)$ and $U(Apply\ \&\ Child\text{-}free)$ and $U(Don't\ Apply\ \&\ Parent)$ and $U(Don't\ Apply\ \&\ Child\text{-}free)$ (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 $U(Apply\ \&\ Parent)$ 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. 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 preference ordering. And for many constructivists I can know that simply by observing how I choose between given sets of actions. 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 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* $U(Apply\ \&\ Parent)$ prior to my choice between *Apply* and

Don't Apply, but indeed that $U(\textit{Apply} \,\&\, \textit{Parent})$ is not even *determined* prior to that choice. If that is the case, then there is no way to make the choice rationally, even according to the realist-evaluative understanding. Similarly, if my preference between *Apply* and *Don't Apply* is not even defined prior to my choice between them, then I cannot make that choice rationally, even according to the constructivist. But of course this is not what Paul's argument establishes. The argument is explicitly epistemic.

7.3 The Fine-Graining Response

There is a natural response to Paul's challenge, and it is similarity between this natural 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 agent. 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 $P(\textit{Parent}||\textit{Apply}) = 0.12$. 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 the world, 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, but also about the value $U(\textit{Apply} \,\&\, \textit{Parent})$. 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 val-

ues that I might assign to being a parent are -12, 3, and 10. Then, while my original set of possible states of the world is $\mathcal{S} = \{\textit{Parent}, \textit{Child-free}\}$, my new expanded set of possible states of the world is

$$\mathcal{S}^* = \{\textit{Parent \& utility of being a parent is } -12,\\ \textit{Parent \& utility of being a parent is } 3,\\ \textit{Parent \& utility of being a parent is } 10,\\ \textit{Child-free}\}.$$

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:

$$P(\textit{Parent \& utility of being a parent is } -12||a),\\ P(\textit{Parent \& utility of being a parent is } 3||a),\\ P(\textit{Parent \& utility of being a parent is } 10||a),\\ P(\textit{Fail}||a),$$

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 agent, but also how things are inside the agent; 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 our favoured solution to the problem of choosing for changing selves. I'll consider two such objections: the first due to Paul herself (section 7.4), the second to Sarah Moss (section 7.5).

7.4 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.[41] 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

[41] 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 Moss considers.

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 $P(Parent||Apply) = 1$):

- $P(\textit{Parent \& utility of being a parent is } -12 \,||\, Apply) = 0.3$,
- $P(\textit{Parent \& utility of being a parent is } 3 \,||\, Apply) = 0.6$,
- $P(\textit{Parent \& utility of being a parent is } 10 \,||\, Apply) = 0.1$,

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 agent who made the decision to become a parent in this way would be "alienated" from that decision; the choice thus made would be "inauthentic":

> 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 agents. 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 agent 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 agent who makes the moral decision for themselves.

I have responded to Paul's concern elsewhere, where I argue that there is a crucial difference between these cases (Pettigrew, 2015b, 770). 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.

I mention Paul's Authenticity Reply here partly for the sake of completeness, but also because Moss' No Knowledge Reply to the Fine-Graining Response 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. Let's turn to Moss' reply now.

7.5 Moss' No Knowledge Reply

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, ms, 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 that 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, ms). 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 $\{P : P(\textit{Kenny is in Hamburg}) = 0.5\}$.

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*. 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 $\{P : P(\textit{Kenny is in Hamburg}) > 0.5\}$. 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, $\{P : P(\textit{Kenny is in Hamburg}) > 0.5\}$.

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 propositions. 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 proba-

bilistic 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 amongst 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 amongst living 35 year olds is quite different from the likelihood amongst all human beings. You know that I am male. And you can't rule out that the likelihood of death amongst living males is different from the likelihood amongst 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 amongst 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 amongst *all* existing parents. Then there are certainly relevant alternative references classes that I should consider: the class of all male parents; the class of all gay male parents; 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 agent 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 agent whose credal state is represented by **P** — in the language

introduced above, **P** is her representor.[42] Which actions are permissible for this agent? 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 agent 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 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

[42] For more on the correct decision theory for imprecise credences, see (Seidenfeld, 2004; Seidenfeld et al., 2010; Elga, 2010; Joyce, 2010; Rinard, 2015).

the question of what my statistical evidence allows me to know, we will have more to say below.

7.6 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 possible acts are
  - *Apply*,
  - *Don't Apply*;
- the possible states are
  - *Parent & utility of being a parent is* -12,
  - *Parent & utility of being a parent is* 3,
  - *Parent & utility of being a parent is* 10,

  - *Fail*;

- the doxastic states are my precise or imprecise 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, the new decision problem that we feed into the machinery of decision theory is this:

- the possible acts are

  - *Apply*,
  - *Don't Apply*,

  as before;

- the possible states are

  - *Succeed*,
  - *Fail*;

- the doxastic states are my precise or 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 the decision problem is different from that in the orthodox version, while the doxastic state remains the same.

In contrast, for Moss, the new decision problem is this:

- the possible acts are

  - *Apply*,

  - *Don't Apply*;

- the possible states are
  - *Parent & utility of being a parent is* -12,
  - *Parent & utility of being a parent is* 3,
  - *Parent & utility of being a parent is* 10,
  - *Fail*;

- the doxastic states are not my precise or imprecise 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 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.

7.7 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, granted that the Fine-Graining Response is wrong, 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 concern I wish to raise 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 7.3 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 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.

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 agent with unjustified credences should maximise expected utility with respect to those credences, that agent 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*.

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 amongst 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 amongst 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 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 amongst 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 agent 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 agent 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 amongst 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 amongst 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 dis-

tribution of utilities within them, but nonetheless 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 amongst 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. Consider the following Non-Probabilistic Knowledge Norm for Reasons. It says that a proposition p can count as your reason for performing an action just in case you know p. That can legitimately be said to set an ideal — 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. But that isn't the case for probabilistic knowledge. Sup-

pose I know that it is at least 50% likely to rain. And suppose I am deciding whether or not to take my umbrella. The higher the likelihood of rain, the higher my expected utility. If it's over 40% likely to rain, I maximise my utility by taking my umbrella. Thus, since I know it's at least 50% likely to rain, I should take it. But this piece of knowledge 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 that belief is justified but not yet knowledge. It isn't obvious that I do better by gaining evidence that turns my justified belief that it's at least 50% likely to rain into knowledge than by gaining evidence that justifies a belief that it's at least 90% likely to rain, but which doesn't make it knowledge. And if that is not the case, then knowledge isn't the goal at which we always aim.

Before we wrap up this chapter, I'd like to draw attention to one final point, which is apt to be neglected. On the orthodox version of decision theory, an agent 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 agent 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 agent 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. 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 $\mathbf{P}$ 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.

Chapter 8

Why aggregate using weighted averages?

In chapter 4, we considered a number of different ways in which we might aggregate the attitudes and judgments of a group of agents who boast credences, utilities, value 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 agent's credence and utility function to give their value function, and then aggregate the resulting value functions to give the aggregate value function of the whole group; and we considered the weighted average ex post method, on which we aggregate the agents' credences and utilities separately first to give the group credence function and the group utility function, and then combine those aggregates to give the aggregate value function. In each case, whenever the judgments were represented numerically — such as the Borda count representation of a preference ordering, or the value function, credence function or utility function representation of the agents' doxastic and conative attitudes, which measure the value of an act, the credence in a state, or the utility of a situation, respectively, on a cardinal scale — 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 takes the group score for a particular act to be its average Borda count across the agents in the group — and straight averages are just weighted averages in which each agent is assigned the same weight. The weighted average ex ante method takes the group value function to be a weighted average of the individual value 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 is a weighted average of the utilities assigned to that state 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 above, which says that your value for an act should be your expectation of the utility of that act, is itself a judgment aggregation norm, and indeed a weighted averaging or linear pooling norm at that. Let's see what I mean by this. Consider a standard decision problem, equipped with a set of states, $\mathcal{S}$, and a set of acts, $\mathcal{A}$. You know your utilities at each of the situations a & s, where a is an act in $\mathcal{A}$ and 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, and you need to set your value for each act. 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. And you do know what your value 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 value function coincide. Thus, if we consider your know-it-all counterpart at state s, then their value for 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_s, utility function U_s, value function V_s, and preference ordering $\preceq_s$. Your attitudes, on the other hand, are P, U, V, and $\preceq$. Then:

- $P_s(X) = 1$, if s is in X, and $P_s(X) = 0$, if s is not in X.
- $U_s(a \,\&\, s) = V_s(a) = U(a \,\&\, s)$

Now, suppose you are considering how to set your attitudes. You have set

your utility function U. But you have not yet set your credence function P, your value function V, nor your preference ordering $\preceq$. A natural thing to say is that your credences and your values are the aggregated attitudes of your know-it-all counterparts. Thus, your credences are obtained by aggregating the credences of your know-it-all counterparts, while your values are obtained by aggregating their values. Suppose, then, that we aggregate these all together using the method of weighted arithmetic averages; that is, by linear pooling. Then there are weights $0 \leq \langle \alpha_s \rangle_{s \in \mathcal{S}} \leq 1$, one 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 agents:

(a) Your credence in a proposition X is:

$$P(X) = \sum_{s \in \mathcal{S}} \alpha_s P_s(X) = \sum_{s \in X} \alpha_s$$

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

(b) Your value for a is

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

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, which say that my credence in a tautology should be 1, my credence in a contradiction should be 0, and 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.

8.1 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* average as opposed to a weighted *arithmetic* average, which we have been doing so far — this is known as *geometric pooling*. 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 not only our favoured approach to the problem of choosing for changing selves — granted that it is a judgment aggregation problem, and that the attitudes to be aggregated are the most fundamental ones, namely, the credences and utilities, we will have shown that these fundamental attitudes should be aggregated using weighted arithmetic averaging — we will also have justified the claim that a rational agent's credences are probabilities, and taking the value of an act for an agent to be their subjective expectation of its utility.

8.1.1 The Principle of Minimal Mutilation

The argument that we will present is based on what I will call *the principle of minimal mutilation*.[43] 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. The principle of minimal mutilation says, roughly, that the aggregate of the attitudes of a group of individuals should lie as close as

[43] See (Pettigrew, 2015a, tab) 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 consider by Predd et al. (2008). The claim that minimizing average or total distance from agent'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).

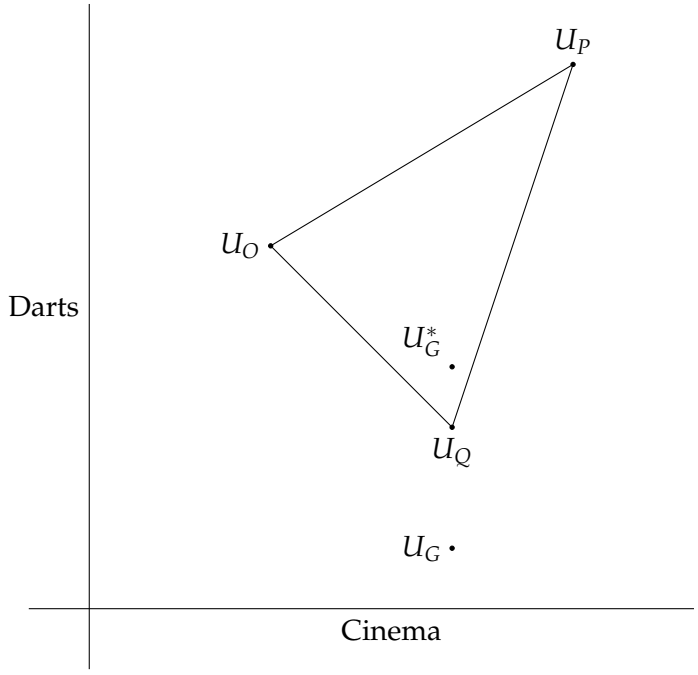

Figure 8.1: Here, we plot the utilities of Omar, Pepejn, Quentin, and the two putative aggregates on the Euclidean plane. The x-coordinate gives their utility for going to the cinema; the y-coordinate gives their utility for playing darts. The triangle formed by drawing lines between each of them contains all and only the weighted arithmetic averages of the three sets of attitudes.

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 8.1):

	Omar	Pepejn	Quentin
Cinema	3	8	6
Darts	6	9	3

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 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.

8.1.2 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. 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 measure together with any positive transformation of it.[44] 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

[44]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) $d'(\mathbf{a}, \mathbf{b}) = H(d(\mathbf{a}, \mathbf{b}))$.

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.[45] Throughout, we will assume that all of the agents between 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 agents assigns credences, or they may be situations to which they assign utilities, or they may be acts to which they assign values. Let's write the set of items towards which each of our agents has attitudes $X_1, \ldots, X_m$. Thus, to specify an agent, 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 agent to another, we are asking for a function $d_m : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty]$.

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 agents can depend upon. Given a pair of agents, **a** and **b**, the distance between them depends only on the following multiset:[46]

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

Formally:

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

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

[46] 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.

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 agents, 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 agents 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

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*.[47] 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)

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

(ii)

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

then

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

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

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:

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

	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.

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 aggregating 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_k, a_l and b_k, b_l are ordered in the same way;

(iii) $|a_i - a_j| = |a_k - a_l|$ and $|b_i - b_j| = |b_k - b_l|$;

Then

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

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:

$$d_n(\langle a_1, \ldots, a_n\rangle, \langle b_1, \ldots, b_n\rangle) = H\left(\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}\right)$$

where X is a continuous and strictly increasing function. That is, d_n is a positive transformation of Euclidean distance.

How does this help? Because of the following fact, which is illustrated in Figure 8.2:

Theorem 8.1.1 (Dominance Theorem) *Suppose we have a set of n agents,* $\mathbf{a}^1, \ldots, \mathbf{a}^n$, *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 agents. 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 $1 \leq k \leq m$,

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

where b_k is the attitude of agent $\mathbf{b}$ to item X_k, while a_k^i is the attitude of agent $\mathbf{a}^i$ to item X_k.

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 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. As a result, in order to persuade your audience that your particular favoured method of aggregation is the correct one, the sort

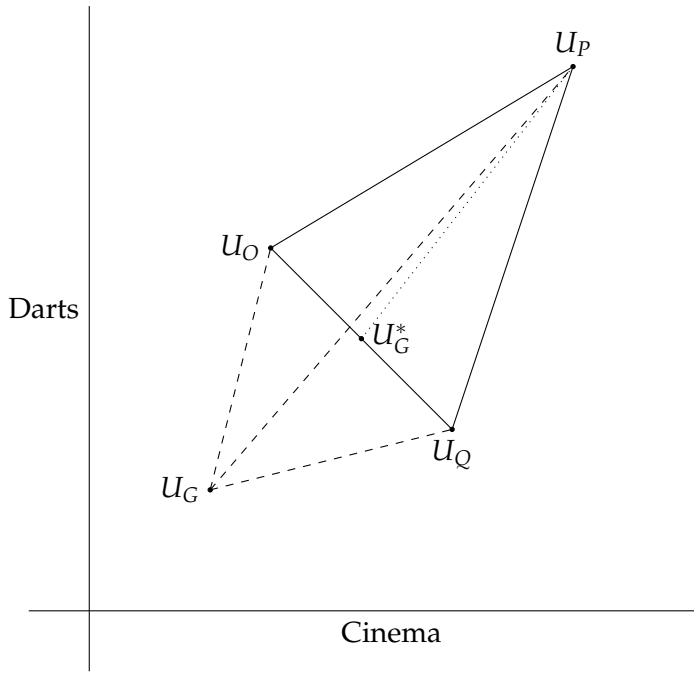

Figure 8.2: Again, we plot the utilities of Omar, Pepejn, and Quentin on the Euclidean plane (as U_O, U_P, U_Q, respectively). Consider the candidate aggregate utility function U_G, which assigns utility 2 to both Cinema and Darts. It is not a weighted average of the individual utility functions; all such weighted averages lie inside the triangle. However, as we can see, there is an alternative candidate aggregate, U_G^*, that is closer to each individual utility function than U_G is. The dashed lines show the distance from U_G to the various different agents; the dotted lines show the distance from U_G^* to those agents.

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.

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

Then 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 $1 \leq k \leq m$, $a_k^{i^*} = a_k^G$. Now, there are certainly linear pooling methods that violate this: if all of the weight is placed on one individual — so that $\alpha_1 = 1$ and $\alpha_2 = \ldots = \alpha_n = 0$, for instance — then that individual is a dictator. If that isn't the case, however — that is, if $0 < \alpha_1, \ldots, \alpha_n < 1$ — 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 agent by their average distance to the other agents in the group, so that we assign lower weight to outliers and higher weight to those who belong to clusters of agents who agree or nearly agree on a lot. Or, in the credal case at least, we might determine an agent'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.

8.3 Conditions on aggregate credences

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 features of no non-dictatorial linear pooling methods.

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 ourselve what it means for a probabilistic credence function P to render two propositions X and Y independent. Recall from above: *X and Y are probabilistically independent relative to P*, if $P(X|Y) = P(X)$; that is, if the credence in X does not change when we condition on Y. Equivalently, X and Y are probabilistically independent relative to P, if $P(XY) = P(X)P(Y)$; that is, if the probability of X and Y both occurring is the probability of X occurring weighted by the probability of Y occurring. 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 8.3.1 *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.*

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 whenever those propositions are probabilistically independent relative to her credence function. Thus, if your credence function is P, and if X and Y are probabilistically independent relative to P, then you judge X and Y to be independent. 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 agents may be unanimous in their judgment that Y is independent of X, while at the same time nearly all linear pools of their credences judge otherwise.

It seems to me that the mistake in this objection 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 is 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.[48]

[48]More precisely: There are two possible objective chance functions ch_1 and ch_2. If we let

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 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

---

H_i be the proposition that the coin will land heads on its i^{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, $P(H_2|H_1) = \frac{P(H_2H_1)}{P(H_1)} = \frac{P(C_{ch_1})ch_1(H_2H_1)+c(C_{ch_2})ch_2(H_2H_1)}{P(C_{ch_1})ch_1(H_2)+P(C_{ch_2})ch_2(H_2)} = \frac{(0.5\times0.1^2)+(0.5\times0.9^2)}{(0.5\times0.1)+(0.5\times0.9)} = \frac{0.82}{1} = 0.82 > 0.5 = P(H_2)$. 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.

them to be stochastically dependent — what I don't know is whether the dependence is positive or negative.[49]

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 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

[49] 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

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

judgments, even when they are made unanimously. Remember: we noted in section 4.2 that there will always be some features shared by all members of a group that cannot be shared with the group credence function.

Elkin & Wheeler (2016) try to dramatise the objection we are considering by presenting a Dutch Book argument against groups whose group credences fail to preserve independences shared by all members of the group. Their idea is this: Suppose that, relative to the credence function of each member of a group, propositions X and Y are probabilistically independent. And suppose that, relative to their group credence function P, X and Y are not probabilistically independent — that is, $P(XY) \neq P(X)P(Y)$. Then, according to Elkin and Wheeler, there are two ways in which we can calculate the price at which the group will be prepared to buy or sell a £1 bet on the proposition XY — that is, a bet that pays £1 if XY turns out to be true, and which pays £0 if XY is false. First, the group will be prepared to buy or sell a £1 bet on XY at $£P(XY)$, since that is the group credence in XY. Second, Elkin and Wheeler claim that the group should also be prepared to buy or sell a £1 bet on XY at $£P(X)P(Y)$, since $P(X)$ is the group credence in X, $P(Y)$ is the group credence in Y, and the group judges X and Y to be independent. But, by hypothesis, $£P(XY) \neq £P(X)P(Y)$, and if an agent has two different prices at which they are prepared to buy or sell bets on a given proposition, it is possible to Dutch Book them. Suppose that $P(X)P(Y) < P(XY)$. Then we simply sell them a £1 bet on XY at $£P(XY)$, which they consider a fair price. This will give the group a net gain of $£(1 - P(XY))$ if XY is true and a net gain of $-£P(XY)$ if XY is false. And then we buy from them a £1 bet on XY at $£P(X)P(Y)$, which is their other fair price. This will give the group a net gain of $£(P(X)P(Y) - 1)$ if XY is true and a net gain of $£P(X)P(Y)$ if XY is false. Thus, their total net gain if XY is true is $£(1 - P(XY)) + £(P(X)P(Y) - 1) < £0$. And their total net gain if XY is false is $-£P(XY) + £P(X)P(Y) < £0$. That is, the group is vulnerable to a series of bets, each of which it considers fair, but which collectively guarantee that it will lose money. And similarly with the sign reversed if $P(XY) < P(X)P(Y)$.

The problem with this argument is the same as the problem with the original objection. The fact that X and Y are probabilistically independent relative to each individual's credence function does not entail that each individual judges X and Y to be independent. And without that, we have no reason to think that the group should also judge X and Y independent, and thus no reason to think that the group should judge $£P(X)P(Y)$ a fair price for a £1 bet on XY. In sum: I conclude that it does not count against linear pooling that it does not preserve probabilistic independence.

So I think we shouldn't require our aggregation methods for credences to preserve probabilistic independence from individuals to aggregate. The next apparently desirable feature of aggregation methods for credences that I'll consider is closely related to this. Both concern conditional probabilities — the first concerned their relationship to judgments of independence; the second concerns their relationship to rules of updating. 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 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 α 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 — perhaps they both learn the output from some climatological instrument. 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 individual's 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 α 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. The following theorem makes this precise:

Theorem 8.3.2 ((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.

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, ms). So we have (1) and (2).[50]

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):

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

Theorem 8.3.3 ((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 α'.*

That is, to obtain the new weight, α', for the first individual (whose initial credence function is P_1), we take the old weight, α, 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 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.[51] 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

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

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 agent 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 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.[52] This is my response to Staffel's objection.

[52]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^{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 $ch_2(H_iH_j) = ch_2(H_i)ch_1(H_j)$

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, we have no reason to hope that our aggregation method preserves independence judgments, and linear pooling in fact interacts well with conditionalization.

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

Bibliography

Aczél, J., & Wagner, C. G. (1980). A characterization of weighted arithmetic means. *SIAM Journal on Algebraic Discrete Methods*, *1*(3), 259–260.

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.

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.

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. (ms). Conditionalization. Unpublished manuscript.

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.

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.

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.

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*.

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.

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.

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, 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.

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.

Hedden, B. (2015a). Does MITE Make Right? In R. Shafer-Landau (Ed.) *Oxford Studies in Metaethics*, vol. 11. Oxford University Press.

Hedden, B. (2015b). *Reasons without Persons: Rationality, Identity, and Time*. Oxford, UK: 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.

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. C. (1983). *The Logic of Decision*. Chicago and London: University of Chicago Press, 2nd 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. (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.

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.

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.

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. (ms). *Probabilistic Knowledge*. Oxford University Press.

Nagel, T. (1978). *The Possibility of Altruism*. Princeton University Press.

Okasha, S. (2016). On the Interpretation of Decision Theory. *Economics and Philosophy*, *32*, 409–433.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Paul, L. A. (2014a). *Transformative Experience*. Oxford: Oxford University Press.

Paul, L. A. (2015). 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. (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. (taa). Aggregating incoherent agents who disagree. *Synthese*.

Pettigrew, R. (tab). 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. (tac). 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.

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.

Ramsey, F. P. (1931). Truth and Probability. *The Foundations of Mathematics and Other Logical Essays*, (pp. 156–198).

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.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons.

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.

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.

Talbott, W. J. (1991). Two Principles of Bayesian Epistemology. *Philosophical Studies*, *62*(2), 135–150.

Thoma, J., & Weisberg, J. (2017). Risk writ large. *Philosophical Studies*, *174*, 2369–2384.

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.

von Neumann, J., & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press, 2nd 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.

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.